Volume 38 Number 3 September 2014 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Special Issue:

Frontiers in Network Systems and Applications

Guest Editors:

Andrzej Chojnacki
Maciej Grzenda
Andrzej Kowalski
Bohdan Macukow

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Editors's Introduction to the Special Issue on "Frontiers in Network Systems and Applications"

The growing availability of network infrastructure is paralleled by the ever growing number of network applications. These applications rely on network services to answer business needs and enable the interaction of users and devices. Not surprisingly, the development of applications promotes further research in the field of network systems. In particular, the growing complexity of network infrastructure gives rise to the need for even more complex research and development environments. This need is observed both in the case of complex fixed network layouts and mobile environments. To answer these needs the efforts of industry leaders and academia can be combined. By combining the ability to build and deploy complex network layouts with more theoretically grounded research, further progress in the field can be attained. This illustrates the benefits arising from the collaboration between industry leaders and academia.

In 2012, to promote such cooperation of industry experts and researchers with academic research groups, the first Frontiers in Network Applications and Network Systems symposium was organised in Wroclaw, Poland. The symposium provided a forum for the exchange of ideas between network operators, designers and researchers. The symposium was organised as a part of the Federated Conference on Computer Science and Information Systems (FedCSIS). This created a unique opportunity to discuss frontiers in network system development in view of recent developments in other areas of computer science. The solutions from many of these areas, such as the solutions incorporated from the domain of artificial intelligence, or database systems, are also vital components of modern network systems.

The first edition of the conference was followed by even further cooperation with other network-related events of the FedCSIS multiconference. As a result, in 2013, an International Conference on Innovative Network Systems and Applications (iNetSApp) was organised for the first time. It included a variety of network-related topics with the emphasis on network systems, applications and services (SoFast-WS track of the conference) and wireless sensor networks (WSN track of the conference).

This special issue includes selected extended versions of papers, most of which were presented during SoFast-WS conference i.e. the Frontiers in Network Applications, Network Systems and Web Services conference organised in 2013 in Krakow, Poland. The SoFast-WS conference is co-organised by the Research and Development Centre of Orange Polska – a part of a global chain of R&D Orange Labs centres of Orange telecom group, the Faculty of Cybernetics of the Military University of Technology, the Faculty of Mathematics and Information Science of the Warsaw University of Technology, and Zayed University. Hence, the idea of promoting the cooperation between business and academia is directly reflected in the composition of both the organising team and the program committee of the conference.

The selection of papers contained in this special issue reflects various research activities in the field of network systems. The first work, Future proof access networks for B2B applications, authored by P. Parol and M. Pawłowski, discusses the development of the Gigabit-capable Passive Optical Network (GPON). Moreover, the authors propose a way the GPON network can provide a basis for Software-Defined Networking (SDN). A solution based on OpenFlow is proposed in this context. At the same time, the work provides a clear illustration of the complexity of network systems combining recent hardware developments with sophisticated novel protocols.

In the next work, The architecture of Distributed Database System in the VANET Environment, J. Janech, E. Kršák, and Š. Toth discuss the role of a database system in a vehicular network. The unique requirements that have to be met by a database system serving the needs of moving objects while taking into account location aspects are discussed. This clearly illustrates the interdisciplinary research needed to develop modern network systems, in this case involving database-related research.

Another perspective on network systems is offered by the work Prototype Implementation of a Scalable Real-Time Dynamic Carpooling and Ride-Sharing Application. In this work, D. Dimitrijević, V. Dimitrieski and N. Nedić propose a way in which progress in the development of a ride-sharing application can be attained. Hence, a user perspective on the system and the requirements of the users of the network application are fundamental for this work.

Finally, in Tiny Low-Power WSN Node for the Vehicle Detection, the development of network systems is again considered from the hardware point of view. Even though the work, authored by M. Chovanec, M. Hodon, and L. Cechovic, refers to wireless networks, it documents the development of a novel hardware device. The device is a special low-power sensor node embedding a magnetometer as the main sensing tool for vehicle presence monitoring.

The works contained in this special issue illustrate various directions of research conducted in the field of network systems, involving both novel hardware and software developments.

The editors would like to thank the SoFast-WS Program Committee members for their contribution to both the conference and the preparation of this special issue.

Andrzej Chojnacki, Maciej Grzenda,
Andrzej Kowalski, Bohdan Macukow

Editors of the special issue

Future Proof Access Networks for B2B Applications

Paweł Parol
Orange Labs Poland, Obrzeżna 7, 02-691Warsaw, Poland
Warsaw University of Technology, Faculty of Electronics and Information Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland
E-mail: pawel.parol@orange.com

Michał Pawłowski
Orange Labs Poland, Obrzeżna 7, 02-691Warsaw, Poland
E-mail: michal.pawlowski1@orange.com

Keywords: SDN, GPON, Optical LAN, B2B

Received: December 10, 2013

The paper offers an innovative approach for building future proof access network dedicated to B2B (Business To Business) applications. The conceptual model of considered network is based on three main assumptions. Firstly, we present a network design based on passive optical LAN architecture utilizing proven GPON (Gigabit-capable Passive Optical Network) technology. Secondly, the new business model is proposed. Finally, the major advantage of the solution is an introduction of SDN (Software-Defined Networking) paradigm to GPON area. Thanks to such approach network configuration can be easily adapted to business customers' demands and needs that can change dynamically over the time. The proposed solution provides a high level of service flexibility and supports sophisticated methods allowing users' traffic forwarding in efficient way. The paper extends a description of the OpenFlowPLUS protocol proposed in [18] . Additionally it provides an exemplary logical scheme of traffic forwarding relevant for GPON devices employing the OpenFlowPLUS solution.

Povzetek: Prispevek uvaja nova omrežja B2B z zajamčenim dostopom.

1 Introduction

In recent years one can observe skyrocketing of Internet traffic (its growth is exponential) as end users are consuming more and more Internet services (e.g. video or cloud based solutions). Growth is visible for all types of access, especially mobile but also fixed (fixed still account for vast majority of traffic). The situation is often highlighted by telecommunications providers but also organizations like enterprises, universities, governmental entities have to deal with high growth.. At the end many institutions have to adapt and bolster their traditional IT and network infrastructure in order to handle that phenomenon and to provide needed services and solutions More and more aspects of economy and our life is dependent on Internet thus assurance of quality and reliability as well methods to deal with its growth is crucial.

In the following chapter the overview of legacy campus networks, typically used by institutions, is given. Also LAN (and WAN access) solutions provided to business customers are described.

1.1 Office networks overview

Nowadays access to Internet is prevalent among companies. Moreover many enterprises have own intranet used for number of different purposes. In order to provide connectivity to end devices likes PC, laptops or tablets in-building network infrastructure is needed. It can be composed of a single modem/router but also tens of devices and substantial amount of transmission medium (optical fibers, twisted pair cables etc.). In case of big organizations all those components form a campus network – computer network interconnecting LANs (Local Area Networks) within a limited geographical area. The infrastructure is usually owned by campus owner / tenant e.g. enterprise, university, hospital.

Early LANs were large, flat networks with peer to peer Layer 2 (L2) communication based on Ethernet standard [4] . It was a simple approach but with network continuous growth ultimately led to disruptions (e.g. due to broadcast storms). Over the time Layer 3 (L3) has been introduced dividing campus network into smaller segments (allowing avoiding such problems). Additionally numbers of different solutions like VLANs (Virtual LANs [5]), RSTP (Rapid Spanning Tree Protocol [6]) or IP subnets have been developed making campus networks easier to maintain and manage, but with the cost of additional complexity

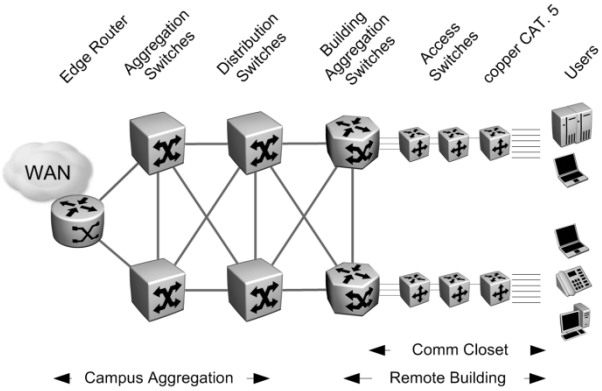


Figure 1: Campus network hierarchy.

Also the topology evolved towards more hierarchical and structured design. 4-tier network architecture (see Figure 1) has become common ([8] , [10]). In that approach access layer provides connectivity to end devices with copper twisted pairs (usually UTP CAT. 5 – Unshielded Twisted Pair Category 5, nowadays also CAT. 6 cables are gaining popularity). Fast Ethernet or Gigabit Ethernet (100BASE-TX or 1000BASE-T [4]) are commonly used. Access layer contains multiple L2 switches (usually managed and configurable). They are located nearby users, e.g. in communication closets on each floor of the building. At the next level additional switches aggregate traffic for each building (concentrating multiple access layer switches and providing access to higher levels equipment) Interconnection between access layer switches and building aggregation switches can be provided with copper cables, or with fibers. Building aggregation switches are connected to campus aggregation and distribution switches which then connect to router (or routers being a gateway to external networks). As a transmission medium for interconnecting building aggregation with campus aggregation segments fiber optic cables are often used due to higher bandwidth requirements (i.e. 10 GbE interfaces) and longer distance.

Tiered network design gives flexibility in terms of supporting numerous functions and plethora of end devices (for example growth of client population can be accommodated by adding access layer switches, but that approach is costly). The logical division for different layers does not need to be done with physical tiers; access and aggregation can be provided on the same equipment. It can be especially useful in case of smaller campus simplifying management of reduced number of devices [9] . However such approach requires equipment supporting, in many cases, complex functions as well as some expertise for design and configuration (meaning additional costs).

Important to note is fact that legacy Ethernet-based campus networks have significant limitations. Maximum length of copper Ethernet cables is restricted to 100 meters. In fact 4-tier topology with switches on each floor is an answer to that drawback. Ethernet LAN requires a cable connection to every single user port. This means significant number of access layer switches and wires (copper cables) and at the end it results in high costs. High-frequency signals (used for Fast and Gigabit Ethernet) require more sophisticated copper cable constructions which are physically larger than for lower frequencies (necessary to avoid signal disturbances). In consequence the space required for racks, communication closets is significant. Crucial amount of heat is produced (in many cases requiring additional air conditioning), power consumption is high. Management of high number of active devices is not easy. For those reasons legacy Ethernet LAN is not always the best answer for campus network requirements. That is why an important issue is to find a more effective approach for

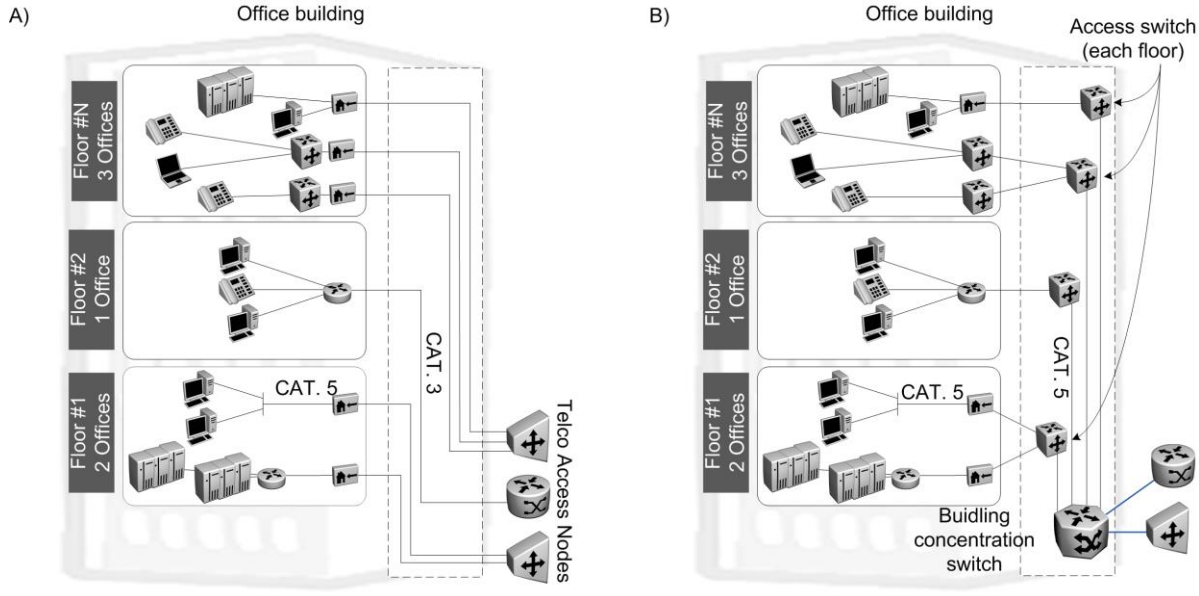


Figure 2: Legacy infrastructure in office buildings.
Scenario A: Telecommunication operators' cables CAT. 3 up to the office.
Scenario B: In-building infrastructure based on Active Ethernet LAN (copper cables CAT. 5)

office networks infrastructure. It especially relevant nowadays when all companies are seeking savings also in IT costs but requirements are getting higher and higher due to the importance of network and IT for business.

1.2 Scenarios for B2B services

B2B (Business to Business) telecommunications services' landscape is diverse. It includes services like Internet access, POTS (Plain Old Telephony Service), VoIP (Voice over IP), dedicated links, VPN (Virtual Private Network), etc. One can distinguish large (Enterprise), medium (SME – Small and Medium Enterprises) and small (SOHO – Small Office Home Office) market segments. However service overlapping (the same services) is possible, but often there are special offers for different segments.

Services' requirements largely depend on type of customers. Big entity owning campus network (and considerable number of network equipment) has other needs than company with small branches scattered around the country (and with lack of its own interconnection) and than small company located in single office building.

For entity with campus network usually telecommunications operator provides its services to location where campus edge router is placed, further propagation is the responsibility of the entity itself (compare Figure 1). In the second case (several branches) it is important to provide secure interconnection among branches.

For office building, in which many companies are located, there are two most common infrastructure scenarios (see Figure 2). First one is based on existing copper CAT. 3 cables which reach customers' desk / office and can be reused by telcos. Modem or router is the termination point of the services (Figure 2 Scenario A).

In the second scenario office building has infrastructure based on active Ethernet LAN with copper cables CAT. 5 (Figure 2 Scenario B). Telecommunications operators need to provide its interconnecting cables up to building's technology room. Separation of services / between different operators can be provided on logical level e.g. by means of VLANs.

For both scenarios the only responsibility of telco is to somehow access business customers. Herein, one could think of a new role for operators targeting office buildings environment: what added values are possible to be identified if telcos take the responsibility of building and administrating the entire in-building office network?

2 Optical LAN

Optical LAN is a new approach for office networks infrastructure and an answer to limitations of legacy Ethernet LANs. All-fiber LAN interconnecting existing Ethernet end devices allows reducing costs and making the network more reliable and future proof.

Proposed solution is based on GPON [1] . It is standardized, well known and widely adopted telecommunications access technology, used by many operators worldwide with millions of end customers. GPON uses point-to-multipoint topology and employs fiber optics as a transmission medium. As a real passive solution – no active equipment is used in-between GPON Access Node: OLT (Optical Line Termination) and line termination at customer side: ONT (Optical Network Termination). The campus network based on Optical LAN number of active equipment is significantly reduced comparing to traditional LAN scenarios. From OLT GPON port a single strand of fiber goes out to a passive optical splitter(s) which splits the signal onto fibers terminating at up to 64 (or even 128) ONTs (see Figure 3). All the fibers, splitters connected to one GPON port on OLT form a GPON tree. ONT device terminates GPON transmission and provides 10/100/1000-BaseT Ethernet connectivity to desktop equipment such as PC computers, laptops, voice over IP phones, and video phones using regular copper patchcords (or by 802.11 WiFi). ONT can be located on customer's desk (ONT per desk) or in office closet (ONT per office). Those two options are called respectively: Fiber-to-the-Desktop (FTTD) or a Fiber-to-the-Communications (FTTC) room. High flexibility of Optical LAN solution allows reusing existing copper infrastructure in buildings (for example GPON access is terminated on ONT located in the floor communication closet, from where existing copper cables are used up to customer's desk, see Figure 3 – Floor #2).

Thanks to fiber optics-based transmission Optical LAN is a long reach access solution – maximum reach is equal to 20 km in a standard mode. It is a tremendous improvement comparing to traditional copper Ethernet (100 m.). It allows placing OLT in distant locations, giving high flexibility in network design (in case of campus network OLT no longer need to be installed in the same building in which customers reside).

GPON technology assures 2.488 Gbps of downstream bandwidth and 1.244 Gbps of upstream bandwidth. Bandwidth is shared among customers connected to the same GPON tree. Advanced GPON QoS mechanisms assure appropriate bandwidth distribution among many users and different applications. In the future even higher bandwidth will be available with next generation of GPON standard. It will require exchange of end equipment (ONT and OLT) but with preservation of fiber optics.

Optical LAN solutions are present in portfolio of several vendors (e.g. Motorola [11] , Tellabs [12] , Zhone [13]). According to vendors estimations introducing of Optical LAN will reduce power consumption by up to 65%, space requirements by up to 90%, capital costs related to network elements by up to 74% [13] . Optical LAN is seen as a new paradigm in campus networking allowing optimization of investments and at the same time improving overall efficiency of the networks. Advances in Optical LAN and size of potential market leaded the main players to start standardisation efforts in order to get even higher adaption and better interoperability. As so one can expect that Optical LAN in the future become important contender for Ethernet.

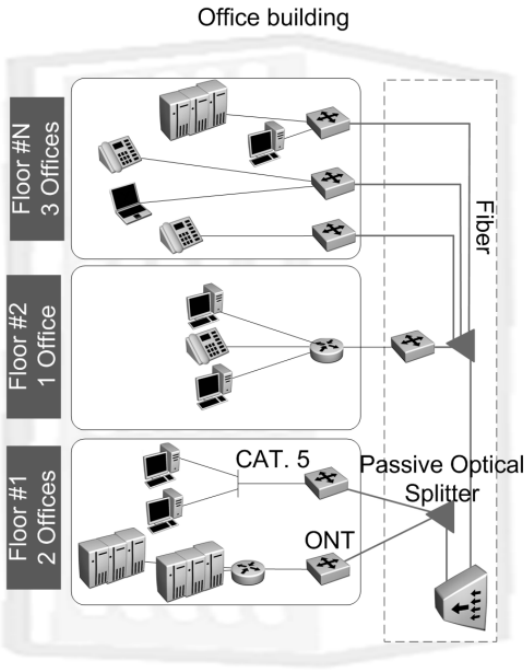


Figure 3: Office building infrastructure based on Optical LAN.

3 Future proof access networks for B2B

In this chapter we formulate three postulates (see Figure 4), which are, from our perspective, crucial for deploying future proof access networks for business applications:

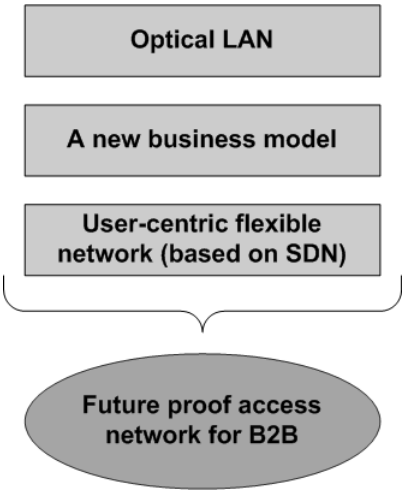


Figure 4: Future proof access network for B2B – conceptual base.

3.1 Applying Optical LAN concept

Currently Optical LAN vendors target big entities with large campus networks. In typical deployment Optical LAN is used by only one organization – the owner and the administrator of the campus network. Office buildings with many tenants, each of them having its own LAN network (at least up to some point) are not yet addressed.

In this paper we propose a solution to that deficiency. It is based on concept known from telecommunications world where many customers are connected to the same Access Node (different users served on the same equipment). In our proposition enterprises no longer need to operate any active network equipment or to build networks itself. LAN becomes a service, provided in similar fashion as e.g. Internet access. LAN service provider is responsible for service creation, administration and adjustment according to needs of customers (enterprises using LAN). That also means that network infrastructure is built for offices by LAN service provider. In fact such network is similar to GPON access networks used by telecommunications operators to provide services to its customers. For B2B scenario different customers are also served by the same GPON OLT unit.

3.2 A new business model

Telecommunications operators are well positioned to play the role of Optical LAN service providers. Usually they have necessary experience with GPON technology, operational resources and existing access network. Telcos are able to deploy optical fiber LAN in office buildings and to provide flexibility in management, service creation and administration.

Such approach has many advantages in terms of optimal usage of network resources. Single OLT can be used for several buildings, even if they are located in distant areas (due to long reach offered by GPON technology which capabilities in terms of maximum physical reach are not fully used in current optical LAN implementations). Also interconnection of distant branches becomes easier (in specific cases they can be served by the same OLT). Additionally a new type of services can be introduced called Office LAN services: e.g. on-demand LAN connections between companies located in the same building, access to in-building monitoring system, etc.

This novel approach also creates a new business model for telecommunications companies who become Optical LAN operator (builder and administrator). The main assumptions for such model are as follows:

- a telco company signs a contract with a building owner for building a complete office network based on Optical LAN solution (it covers all passive components like horizontal and vertical ducts, optical fibers, splitters, in-building Optical Distribution Frame, etc.); additionally copper CAT.5 infrastructure for office rooms (ducts, copper cables, sockets, connectors) can be built by telco company it does not exist in the considered building
- a telco company is responsible for administration and maintenance of the network
- business companies that rent space for their offices in the building sign contracts with a telco

company for providing them with telecommunications services
- business customers are given GPON ONTs which are installed in their offices, ONTs are connected to OLT (which is typically located in distant Central Office owned by telco company)
- telco company offer is assumed to cover wide range of services which can be optimized for particular B2B customers
- tenants that do not decide to be provided with services by telco company are also connected to Optical LAN-based in-building network (via ONTs which they are also given), they are free to be served by other telcos or service providers that establish a connection to considered OLT from their own networks – in such cases services' related traffic is transported in dedicated logical "channels" (e.g. in the sense of VLANs) over the considered optical access network

This opportunity to find new B2B market seems to be a good argument in convincing telco players to work on such solutions.

3.3 User-centric flexible network (based on SDN)

Another assumption for the presented approach is that it is based on user-oriented access network design. Service portfolio dedicated to business customers is typically more complex than the one for residential users. For business applications customized services need to be taken into account. Moreover, customer demands can change dynamically over short periods of time. That is why a challenge for networks deployed in business environments is to provide a high level of service flexibility and to forward user traffic in efficient way. To meet those requirements we present in this paper an access network architecture based on SDN (Software-Defined Networking) paradigm which assumes data plane and control plane abstractions separation ([16]). Thanks to such approach network devices become programmable units. In practice it means that network configuration can be easily adapted to the fast-changing needs.

4 SDN-based GPON solution for business applications

In order to introduce SDN paradigm to GPONs area one can propose different methods to accomplish that. One of the possible ways would be to develop a brand new protocol allowing GPON devices to become programmable units. Such approach is supposed to be an appropriate one for designing an optimal logical architecture of OLT and ONT in the scope of data processing and forwarding. However, development of generic SDN-based protocol for GPON would require a lot of standardization efforts and probably it would take a few years to obtain a solution being ready for deployment.

In this paper we present another approach. We propose a solution based on *OpenFlow* ([16]) which is the most widely deployed SDN-based protocol. *OpenFlow Switch* architecture consists of at least three parts ([15]) – see Figure 5:

- *Flow Table(s)* – a structure within switch implementation covering actions associated with each *flow entry*; the *Flow Tables* define the ways of how the traffic flows have to be processed by the switch
- *Controller* – an external unit running a remote control process that manages the switch via the *OpenFlow* protocol; the *Controller* can add, remove and update *flow entries* from the *Flow Table(s)*
- *OpenFlow Channel* – a channel which enables a communication (i.e. sending packets and commands) between the *Controller* and the switch

For a more detailed description of *OpenFlow*-specific logical components and functions please refer to [15] .

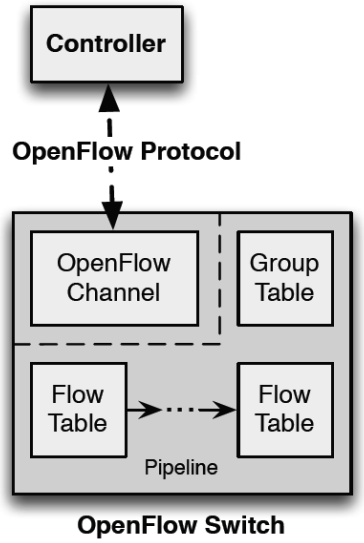


Figure 5: *OpenFlow Switch* logical scheme (source: [15])

OpenFlow was originally designed for L2/L3 switches (or routers) equipped with native Ethernet-based physical interfaces. That is why it is important to notice that it is useless to implement pure *OpenFlow* in GPON OLTs and ONTs. The reason for that is simple: although GPON effectively carries Ethernet frames, in practice it operates at Layer 1 (according the OSI model) with its own dedicated framing and GEM (GPON Encapsulation Method) protocol used for encapsulation higher-layer Protocol Data Units (e.g. Ethernet frames) into GTC (GPON Transmission Convergence) layer. The above mentioned specification of *OpenFlow* protocol does not support such kind of non-Ethernet-based physical interfaces. That is why some additional GPON-related functions have to be introduced to *OpenFlow*.

4.1 GPON-specific traffic instances

A single logical connection within the GPON system is called GEM Port and it is identified by GEM Port-ID. A GEM Port can be considered as a channel within GTC layer and is capable to transport one or more traffic flows. In the upstream direction GPON system also utilizes T-CONTs (Transmission Containers) corresponding to allocated timeslots within TDMA multiplexing existing in GPON. Each T-CONT represents a group of logical connections (GEM Ports) that appear as a single entity for the purpose of upstream bandwidth assignment on the PON (see Figure 6 – GPON-specific traffic entities identifiers are pointed in brackets).

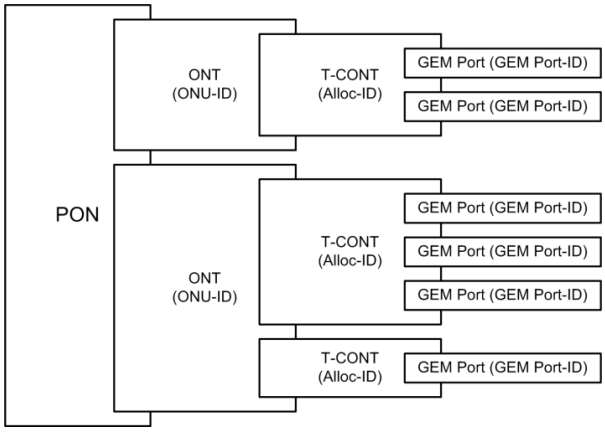


Figure 6: Upstream multiplexing in GPON system.

Each T-CONT can be seen as an instance of upstream queue with a certain bandwidth profile (a set of bandwidth parameters). The bandwidth assignment model applied in GPON system effectively introduces a strict priority hierarchy of the assigned bandwidth components ([2]):

- fixed bandwidth: with highest priority
- assured bandwidth
- non-assured bandwidth
- best-effort bandwidth: with lowest priority

Five T-CONT types which are defined by ([2]) are presented in Table 1.

Table 1: T-CONT types are defined by ([2]).

T-CONT	Traffic descriptor component			
	R_F	R_A	R_M	χ_{AB}
type 1	> 0	= 0	= R_F	= none
type 2	= 0	> 0	= R_A	= none
type 3	= 0	> 0	> R_A	= NA
type 4	= 0	= 0	> 0	= BE
type 5	> 0	> 0	≥ R_F + R_A	any χ_{AB}

Each T-CONT instance is associated with a bandwidth profile. Bandwidth profile is described using the traffic descriptor, which has the following components ([2]):

- fixed bandwidth R_F (bandwidth that is reserved exclusively for a given T-CONT and no other T-CONTs can use it; this bandwidth is statically allocated to a T-CONT)
- assured bandwidth R_A (bandwidth that is available for a given T-CONT on demand; this bandwidth is guaranteed)
- maximum bandwidth R_M (maximum amount of bandwidth, that can be allocated to a given T-CONT on demand; this bandwidth is not guaranteed)
- additional bandwidth eligibility χ_{AB} (type of additional bandwidth that a given T-CONT is eligible to get, can have the following values: none - no additional bandwidth, NA - non-assured bandwidth, BE - best-effort bandwidth)

Depending on the traffic type (latency-sensitive traffic, data transmission, etc.) the most appropriate T-CONT type should be selected to carry considered traffic flows. For instance, T-CONT type 1 characterized by fixed bandwidth component (R_F) only is dedicated to carry fixed-rate traffic which is sensitive to jitter or delay (e.g. ToIP traffic). Such kind of bandwidth is the only one which is allocated statically. In practice it means that it is reserved and always fully available for ONT for which mentioned T-CONT instance was created. Other T-CONT types (from 2 to 5) have to respect the volumes of bandwidth which they are assigned by DBA (Dynamic Bandwidth Allocation) algorithm. This mechanism takes a control over the bandwidth assignment within the entire PON tree. T-CONT type 2 is characterized by the assured bandwidth component (R_A) only and can be utilized for on-off type traffic with well-defined rate bound having no strict delay or jitter requirements. T-CONT type 3 is characterized by assured bandwidth component as well. Additionally it is supposed to be included in non-assured bandwidth sharing. It is suitable for carrying variable-rate traffic with requirement of average rate guarantee. T-CONT type 4 is assumed to participate only in best-effort bandwidth sharing. It is dedicated to carry variable-rate bursty traffic without strict delay requirements. T-CONT type 5 can be considered as a consolidation of other T-CONT types. It can be utilized for wide range of traffic flows [2] . It is worth noting that above described traffic-type-related applications for T-CONTs should be considered only as general recommendations, expressed in [2] . However, it is possible to use a T-CONT of particular type for different sort of traffic flows. Still, it is important to take into account that the T-CONT type specifies capabilities and limitations of the T-CONT instance and thus not all kind of traffic will be appropriate for it.

Upstream user traffic (Ethernet frames) is encapsulated into GEM Ports and then into T-CONTs. Each GPON ONT uses its own set of T-CONTs and GEM ports, a unique one within a GPON tree which ONT belongs to. A single GEM Port can be encapsulated into only one T-CONT, however a single T-CONT may encapsulate multiple GEM ports. In downstream direction only GEM Ports are used to carry traffic flows since no TDMA multiplexing exists there and thus the notion of T-CONT is not relevant for GPON downstream

transmission. For a more detailed explanation please refer to [2] .

As described above one of the key aspects of GPON-based network applications is to ensure effective traffic forwarding on the GTC layer. Traffic incoming from end users (upstream) and "from the network" (downstream) should be carried over passive optical network making use of the capabilities offered by this layer. In practice it means that a significant number of GTC-related traffic instances (GEM Ports and T-CONTs) should be utilized globally within the entire PON tree, especially if a high level of traffic diversity is assumed. By applying such approach it is possible to gain the following advantages:

- separation of traffic flows corresponding to different applications (by using multiple GEM Ports)
- improving QoS performance in upstream (by using multiple T-CONTs)

Thus, such approach improves security and QoS performance in the PON. In order to make it possible it is important to define appropriate rules (consistent and unambiguous ones) allowing to map traffic flows incoming from users to appropriate GEM Ports. In most of commercial implementations mapping rules "built-in" GPON ONTs are mono-criterion i.e. mapping is based on only one of the following criteria like: VLAN ID (Virtual LAN identifier), p-bit ([5]) or UNI (user port number on ONT). For some cases also double-criterion combinations of aforementioned parameters are available (e.g. VLAN ID + UNI) for the mapping purpose. Since GPON was originally designed for B2C (Business to Customer) market segment for which only Triple-Play (Internet, ToIP and IPTV) services are considered such approach was sufficient. For business applications where not only service portfolio is more complex but also customized services are taken into account, much more sophisticated methods (i.e. mapping rules) are required in order to ensure effective traffic forwarding through the system ([14]). In most scenarios currently deployed GPON ONT with limited set of hardcoded mapping and forwarding functions would not be able to address such needs. In such cases software upgrade is needed but it leads to higher operational costs - especially if business customer demands changes dynamically and it is possible that new set of functions is required. For such a scenario multiple software upgrades have to be taken into account.

4.2 SDN-based protocol for GPON

The solution for the issue is SDN-based protocol for GPON allowing OLT and ONT to become programmable units. In this paper we propose *OpenFlow*-based solution. As mentioned before the current specification of *OpenFlow* protocol does not support GPON natively. That is why our vision is to introduce GPON-related functions to the specification in order to develop a protocol extension which we called *OpenFlowPLUS*.

The main assumption for the *OpenFlowPLUS* is that it inherits all the functionality, architecture and capabilities of original *OpenFlow*. The essential improvement is an introduction of GPON-related functions to the protocol in terms of traffic forwarding in order to make the solution relevant also for GPON technology.

According to *OpenFlow Switch* architectural assumptions each device (OLT, ONT) within considered GPON tree contains *Flow Table(s)* and communicates over a *OpenFlowPLUS Channel* with remote *Controller* via *OpenFlowPLUS* protocol (see Figure 7).

For that purpose OLT and ONTs are supposed to have IP address configured. Since *Controller* and OLT are assumed to be connected to IP/Ethernet network they can establish L3 connection. ONTs are accessible by *Controller* only via OLT. One could take advantage of that and for the purpose of *OpenFlowPLUS* messages exchange between ONTs and *Controller* make use of GPON-specific mechanisms defined by [3] . In such a scenario protocol messages are transported through the PON via a dedicated OMCI (ONT Management and Control Interface) channel towards OLT and then they are sent directly to the *Controller* using OLT's *OpenFlowPLUS Channel*. Obviously, employing OMCI by *OpenFlowPLUS* for some new applications does not mean that the protocol takes the control over the entire GPON system. All functions which are out of the scope of traffic mapping and forwarding (e.g. ONT discovery and provisioning-related functions, Dynamic Bandwidth Allocation mechanism, optical layer supervision, alarms and performance monitoring etc.) are assumed to be realized in traditional way, i.e. in line with recommendations defined in [2] and [3] . The physical layer and access network topology remains the same. Each ONT (being a termination of optical network) is installed in customer premise. ONT aggregates traffic incoming from different end devices (PCs, laptops, ToIP phones, servers, etc.) and encapsulates Ethernet frames originated from office LAN into GEM frames. Mapping to GEM Ports and T-CONTs is executed in *FlowTable* based on sophisticated rules provided by *OpenFlowPLUS* protocol. The upstream traffic is sent to OLT using GPON uplink interface. OLT takes a control over the entire PON tree and it is responsible for making the whole system functions working properly. In the scope of GEM-based forwarding OLT assigns unique set of GEM Port and T-CONTs (corresponding to timeslots of TDMA) to each ONT which is connected to the PON. The traffic received from ONTs is forwarded towards other network segments (e.g. towards aggregation network routers) based on the flow entries existing in OLT *FlowTable*. The downstream transmission is supposed to be realized in similar fashion, i.e. making use of *FlowTables* implemented in OLT and ONTs. The important difference as compared to upstream is that in downstream no T-CONTs are used since TDMA does not exists there. That is why traffic incoming "from the network" starting from OLT is encapsulated only in GEM-Ports as the only GTC layer-related traffic instances utilized for the purpose of downstream transmission.

Taking into account above mentioned assumptions an optimal approach seems to be adding *OpenFlowPLUS*

controller as a functional module to the standard EMS (Element Management System) managing the system. Such solution would allow to manage the PON using single management entity. However it is important to stress that integration of GPON technology with *OpenFlowPLUS* protocol requires some efforts in the areas which are much wider than system management aspects. The implementations of *OpenFlowPLUS*-based OLT and ONT will differ from the traditional ones since *OpenFlowPLUS* introduces a different way of traffic forwarding. That is why a logical architecture of OLT and ONT is supposed to be modified in the context of invoking traffic forwarding functions. In other words *OpenFlowPLUS* logical components implemented in GPON devices have to be able to "communicate" with GPON-layer control plane.

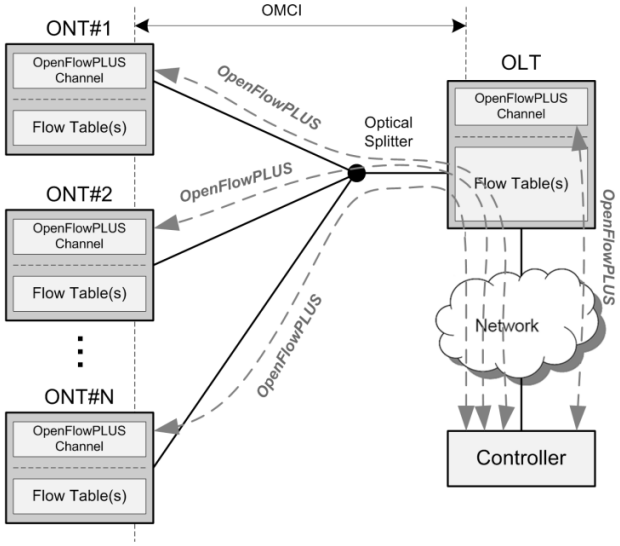


Figure 7: *OpenFlowPLUS*–based GPON solution overview.

As mentioned before the idea of *OpenFlowPLUS* is to provide GPON-related functions to the protocol in terms of traffic mapping and forwarding. Similarly to original *OpenFlow*, *OpenFlowPLUS* is assumed to use *Flow Table(s)* which perform packet lookups, modification and forwarding. Each *Flow Table* contains multiple *flow entries* (see Figure 8). Each *flow entry* contains ([15]):

- *match fields* – to match against packets; *match fields* include packet header fields (e.g. VLAN ID, MPLS label, IP destination address, TCP source port, etc.) and ingress port (optionally also metadata that pass information between tables can be used)
- *priority* – determines an order in which flow entry is matched against packets
- *counters* – which can be maintained for each port, table, flow, etc; counters are updated when packets are matched
- *instructions* – set of operations which are executed when a packet matches a *flow entry*
- *timeouts* – specify amount of time before flow is expired by the switch
- *cookie* – a field reserved for controller in order to modify or delete flows or to filer statistics; it is not used during packet processing
- *flags* – they impact the way flow entries are managed

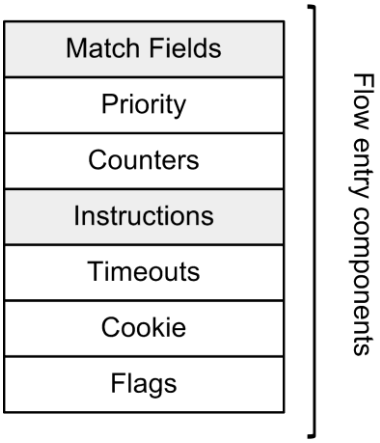


Figure 8: *OpenFlowPLUS* flow entry components.

Instructions define the ways of how single *action* is processed. *Actions* represent operations of packet modification or forwarding to the specified port. *Actions* are grouped by different action types, for instance *pop* action type (e.g. *pop VLAN header* action), *set* action type (e.g. *set MPLS traffic class* action), etc. *Instructions* executed during *OpenFlowPLUS* pipeline processing can either add appropriate *actions* to the current *action set* (a set of *actions* that are accumulated when the packet is processed by the tables and that are executed after exiting the processing pipeline by the packet), or force some *actions* to be applied immediately. *OpenFlowPLUS* defines new *actions* which are relevant to GPON technology. The considered functions are presented in Table 2.

OpenFlowPLUS introduces a brand new action type called *gpon* related to GPON-specific mapping methods. The considered action type provides two *actions*: *Map to GEM Port* action which represents an operation of mapping Ethernet frames to particular GEM Port instance and *Map to T-CONT* action which represents an operation of mapping GEM Ports to particular T-CONT instance. Additionally the new functionality for original *OpenFlow output* action is supposed to be supported: when a packet is destined to be forwarded to the GPON port, GTC framing is performed for the packet before exiting the interface. The aforementioned protocol improvements are the main GPON-related forwarding and mapping functions provided by *OpenFlowPLUS*.

Table 2: Main GPON-related forwarding functions provided by *OpenFlowPLUS.*

GPON unit	Action type /Action	Remarks
ONT, OLT	*gpon*: *Map to GEM Port*	introduction of a new *action* to the original *OpenFlow action set*
		function: mapping Ethernet frames to particular GEM Port instance
ONT	*gpon; Map to T-CONT*	introduction of a new *action* to the original *OpenFlow action set*
		function: mapping GEM Ports to particular T-CONT instance
ONT, OLT	*output*	action modification when executed for GPON interfaces
		new function: GTC framing before forwarding the packet on the GPON port

4.3 Use case

In this section we present a possible application for SDN-based GPON concept which is proposed in the paper. The following assumptions are made for the considered use case:

- the solution is dedicated to business customers who reside in office buildings
- operator acts not only as a service provider but is responsible also for administration of in-building network
- in-building network is based on Optical LAN solution
- different service types can be offered (the considered traffic flows are listed in Table 3):
    - Internet access,
    - ToIP (telephony over IP),
    - Metro Ethernet (corporate connections),
    - cloud computing-based services,
    - office LAN services
- access network architecture is based on *OpenFlowPLUS* GPON solution (see Figure 9)

Each office in the building is connected to the optical network via its dedicated ONT installed in customer premise. Copper cables terminated with RJ-45 sockets are deployed in office rooms. Each user device (PC, IP phone, application server, etc.) is connected to one of multiple Ethernet LAN ports which ONT is equipped with (see Figure 10). ONT aggregates the entire traffic incoming from user terminals (this traffic is in general assumed to contain no VLAN tags) and provides the functionality of L3 gateway. Public IPv4 addresses are assigned to ONT (IP@1.1) and to some selected user devices: web-based application server – IP@1.2 and intelligent installation system controller – IP@1.3. For other devices private addressing is used (IP@priv). ONT is supposed to act as an internal DHCP server for that purpose. Any kind of additional CPE (Customer Premise Equipment) acting as a BGW (Business Gateway) is not required in the considered network.

Table 3: Services and traffic flows overview.

Flow ID	Traffic flow/service	Remarks
F#1	Internet Access: HTTP, FTP, etc.	standard Internet services
F#2	Internet Access: web-based application hosting	connections from Internet are established using HTTPS (SSL + HTTP) protocol (TCP port 443)
F#3	Internet Access: remote access to intelligent installation system controller	connections from Internet to intelligent installation system controller physically located in the office are based on KNXnet/IP protocol (port 3671); IP address of the controller: IP@1.3
F#4	Metro Ethernet: connections to remote company branch	remote company branch is supposed to use IP@2.X address pool;
F#5	ToIP	IP phones used in the office are assumed to mark IP ToS field with DSCP "EF" value; IP address of ToIP platform: IP@4.1
F#6	Office LAN: on-demand connections to different companies located in the same building	connections allowed for a designated sub-pools of addresses from IP@1.X and IP@priv (office) and IP@3.X (different company office)
F#7	Office LAN: access to in-building monitoring systems	in-building monitoring system server is assumed to be connected to a dedicated ONT (installed in the building) with IP address: IP@5.1
F#8	Cloud computing: remote storage, backups	IP address of the server offering cloud computing services: IP@6.1

Thanks to applying *OpenFlowPLUS*, sophisticated mapping rules are supported in order to ensure effective traffic forwarding through GPON. As an example we show how *match fields* with corresponding *instructions* can be defined for *flow entries* within ONT *Flow Table*. To avoid complexity description of forwarding is limited to traffic flows transmitted in upstream direction (see Table 4 and Figure 10). The idea of traffic forwarding is as follows:

1. Traffic flows (packets) incoming from different end devices enter ONT *Flow Table*
2. ONT performs a table lookup. Packet match fields (which typically include various packets header fields) are extracted from the packet and then it is matched against the ONT *Flow Table*. Each *flow entry* in the table is uniquely identified by *match fields* and *priority* taken together. During matching operation only *flow*

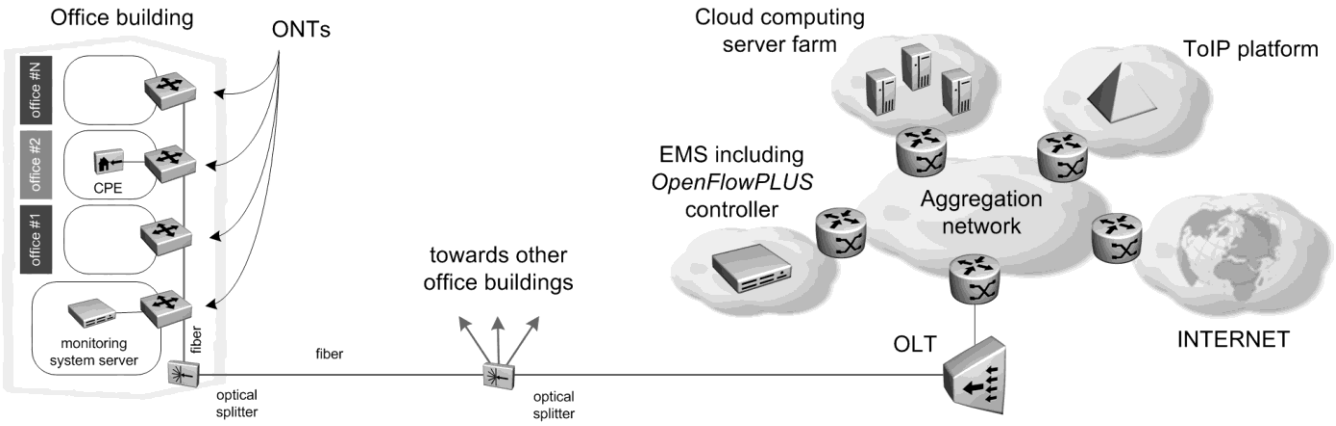


Figure 9: *OpenFlowPLUS*-based architecture for business applications overview.

entry with the highest priority (from those which match the packet) can be selected.

3. For the selected *flow entry instructions* are executed. In particular packet modification actions are applied in the first step, e.g. *Push VLAN header*, *Set VLAN ID*, *Set IPv4 ToS bits*, etc.
4. Next GPON-specific actions are executed, i.e. *Map to GEM Port* and *Map to T-CONT* that correspond to the operation of mapping user traffic flows (packets) to GPON-related traffic entities
5. Finally *output* action is applied which terminates *OpenFlowPLUS* processing pipeline and invokes further packet processing based on the mechanisms defined for GPON technology (GTC framing, allocation to dedicated timeslots, etc.)

Based on the user traffic to GPON-specific instances mapping methods (using limited set of parameters like VLAN ID, pbit, UNI) which are currently supported in typical commercial implementations it would be very difficult or even impossible to follow the traffic forwarding model presented in considered use case. For instance, applying traditional approach it would be impossible to map traffic flows F#1 and F#3 to different GEM Ports if they originated from the same end-device. Moreover it can be expected that the traditional approach:

- would require additional CPE (some kind of L3 Business Gateway - BGW) if standard L2 aware

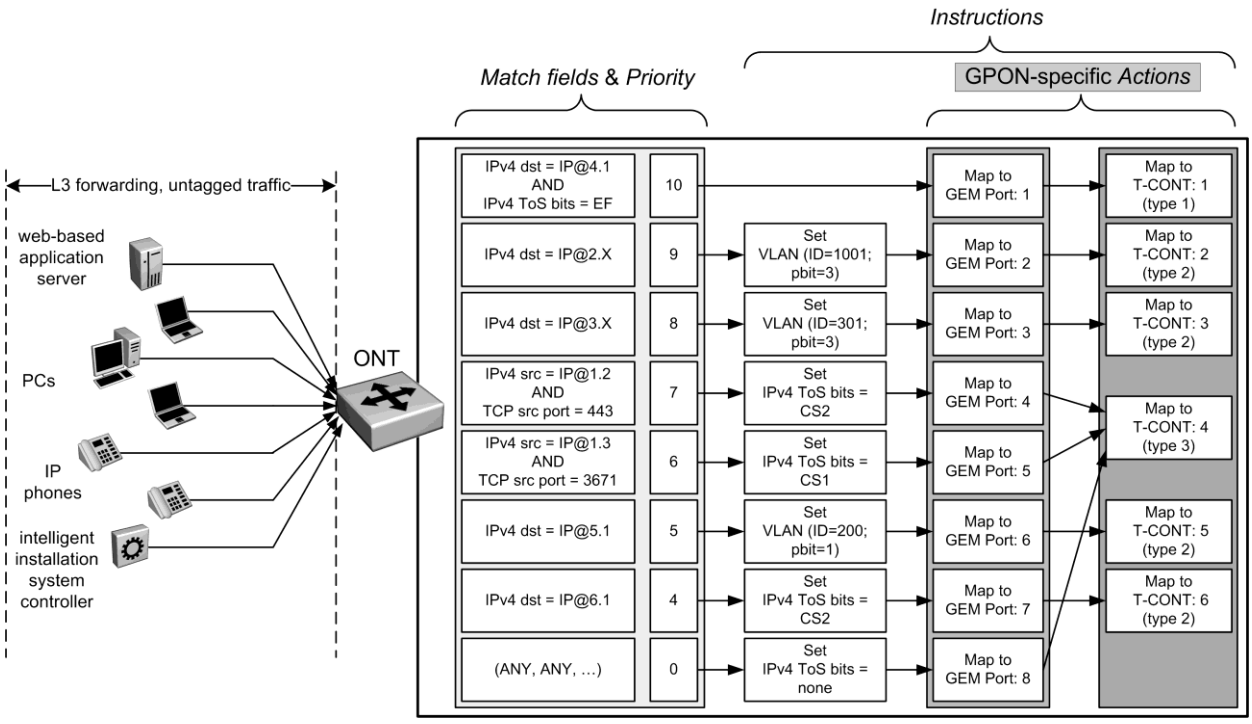


Figure 10: *OpenFlowPLUS*-based ONT forwarding: a logical scheme.

GPON ONT is used
- would introduce VLANs or even additional L2 switches (in some cases) to customer network
- would attach most of end devices to particular physical ports on L2 switches, BGW, ONT
- would partition customer network into "service domains" (e.g. in the sense of VLANs), it may lead to the situation that not all services are available for single end device

All above mentioned aspects make the traditional approach more expensive (more devices), more complex (more devices and more VLANs) and less flexible (fixed connections in customer network, limited set of available services per end device) On the other hand the proposed *OpenFlowPLUS*-based GPON solution is very flexible and convenient from customer perspective. Moreover, since it is assumed to be a programmable system its configuration can be easily adapted to support immediately new services and business needs once they appear.

It is important to stress that considered scenario is only example case for traffic forwarding realized in the sense of *OpenFlowPLUS* protocol. Here, for the sake of simplicity, only one *Flow Table* is used as well as simplified instructions scheme is proposed. However, for real deployment more advanced mechanisms can be applied like pipeline processing based on multi *Flow Table* entities. Also much more complex and sophisticated actions can be defined for traffic flows. The example of forwarding considered in this section is limited only to upstream forwarding for ONT (towards OLT). Obviously, the complete forwarding model for the entire system is supposed to cover:

- upstream forwarding for ONT (towards OLT)
- downstream forwarding for ONT (towards customer network)
- downstream forwarding for OLT (towards ONTs)
- upstream forwarding for OLT (towards aggregation network)

Presented access network model supports also openness for TPOs (Third Party Operators) what is typically required by country-specific regulations. Each customer served by TPO connects a dedicated CPE (which he/she was given by the TPO) to the *OpenFlowPLUS*-based ONT. For such application ONT provides only a basic functionality, i.e. it is configured to work as a traditional L2 bridge (equipped with GPON uplink) that passes assigned VLAN(s) through the system up to the first TPO's switch or router (see "office #2" in Figure 9) interconnected to primary operator's network. It is clear that in such case most of advanced forwarding functions offered by *OpenFlowPLUS* would not be available for such customer. However, it is important to note that it is not a matter of any technology limitations. It is rather a decision of primary operator who would offer "exclusive" features only for its own customers. Nevertheless, such customer always can change the telco company which he/she is served by. In particular the customer can leave the TPO and join primary operator's customers base. In this way he/she would be able to make use of the full range of features offered by *OpenFlowPLUS*-based solution.

match fields of flow entries & priority	**Matched flow**	***Instructions***
IPv4 dst = IP@4.1 AND IPv4 ToS bits = EF Priority: 10	F#5	*Apply-Actions* { *Map to GEM Port*: 1 *Map to T-CONT*: 1 (type 1) output}
IPv4 dst = IP@2.X Priority: 9	F#4	*Apply-Actions* { *Push VLAN header* *Set VLAN ID*: 1001 *Set VLAN priority*: 3 *Map to GEM Port*: 2 *Map to T-CONT*: 2 (type 2) output }
IPv4 dst = IP@3.X Priority: 8	F#6	*Apply-Actions* { *Push VLAN header* *Set VLAN ID*: 301 *Set VLAN priority*: 3 *Map to GEM Port*: 3 *Map to T-CONT*: 3 (type 2) output }
IPv4 src = IP@1.2 AND TCP src port = 443 Priority: 7	F#2	*Apply-Actions* { *Set IPv4 ToS bits* = CS2 *Map to GEM Port*: 4 *Map to T-CONT*: 4 (type 3) output }
IPv4 src = IP@1.3 AND TCP src port = 3671 Priority: 6	F#3	*Apply-Actions* { *Set IPv4 ToS bits* = CS1 *Map to GEM Port*: 5 *Map to T-CONT*: 4 (type 3) output }
IPv4 dst = IP@5.1 Priority: 5	F#7	*Apply-Actions* { *Push VLAN header* *Set VLAN ID*: 200 *Set VLAN priority*: 1 *Map to GEM Port*: 6 *Map to T-CONT*: 5 (type 2) output }
IPv4 dst = IP@6.1 Priority: 4	F#8	*Apply-Actions* { *Set IPv4 ToS bits* = CS2 *Map to GEM Port*: 7 *Map to T-CONT*: 6 (type 2) output }
(ANY, ANY, …) Priority: 0	F#1	*Apply-Actions* { *Set IPv4 ToS bits* = none *Map to GEM Port*: 8 *Map to T-CONT*: 4 (type 3) output }

Table 4: *Match fields* and *instructions* for flows incoming to ONT.

5 Conclusion

In the paper we presented a novel approach to deploy optical access networks addressed to B2B market segment where we defined a new role for telcos. The major advantage of our solution is its flexibility thanks to introduction of SDN paradigm to GPON-based networking. We believe our work can be considered as a

conceptual framework for further analysis and solution development.

References

[1] ITU-T Gigabit-capable Passive Optical Networks (GPON) : General characteristic, ITU-T G.984.1, 2008.

[2] ITU-T Gigabit-capable Passive Optical Networks (GPON) : Transmission convergence layer specification, ITU-T G.984.3, 2008.

[3] ITU-T Gigabit-capable Passive Optical Networks (GPON) : ONT management and control interface specification, ITU-T G.984.4, 2008.

[4] IEEE Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specification, IEEE Standard 802.3-2008, 2008.

[5] IEEE Standard for Local and metropolitan area networks – Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks, IEEE Standard 802.1Q-2011, 2011.

[6] IEEE Standard for Local and metropolitan area networks – Media Access Control (MAC) Bridges, IEEE Standard 802.1D-2004, 2004.

[7] FTTH Council Europe, FTTH Handbook, Edition 5, 2012.

[8] Cisco Systems. (1999). White Paper, Gigabit Campus Network Design – Principles and Architecture [Online]. Available: http://www.cisco.com/warp/public/cc/so/neso/lnso/cpso/gcnd_wp.pdf

[9] Brocade. Designing a Robust and Cost-Effective Campus Network [Online]. Available: http://www.brocade.com/downloads/documents/design-guides/robust-cost-effective-lan.pdf

[10] S. T. Karris, Networks Design and Management. 2nd ed. Orchard Publications, 2009.

[11] Motorola. (2012). White Paper, Creating Simple, Secure, Scalable Enterprise Networks using Passive Optical LAN [Online]. Available: http://moto.arrisi.com/staticfiles/Video-Solutions/Solutions/Enterprise/Passive-Optical-LAN/_Documents/_Staticfiles/WP_POL_Creating Networks_365-095-20298-x.1.pdf

[12] Tellabs. (2011). How Enterprises Are Solving Evolving Network Challenges with Optical LAN [Online]. Available: http://www.tellabs.com/solutions/opticallan/tlab_solve-net-challenges-with-optical-lan_an.pdf

[13] Zhone. FiberLAN Optical LAN Solution [Online]. Available: www.zhone.com/solutions/docs/zhone_fiberlan_solution.pdf

[14] P.Parol and M.Pawlowski, "How to build a flexible and cost-effective high-speed access network based on FTTB+LAN architecture," in Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on , vol., no., pp.655,662, 9-12 Sept. 2012

[15] The OpenFlow Switch Specification. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.3.3.pdf

[16] N.McKeown, T.Anderson, H.Balakrishnan, G.Parulkar, L.Peterson, J.Rexford, S.Shenker, and J.Turner, "OpenFlow: enabling innovation in campus networks". SIGCOMM Comput. Commun. Rev. 38, 2, pp. 69-74, March 2008

[17] Open Networking Foundation. (2012). White Paper, Software-Defined Networking: The New Norm for Networks [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf

[18] P.Parol and M.Pawlowski, "Towards networks of the future: SDN paradigm introduction to PON networking for business applications" in Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.829,836, 8-11 Sept. 2013

The Architecture of Distributed Database System in the VANET Environment

Ján Janech, Emil Kršák and Štefan Toth
Department of Software Technologies,
Faculty of Management Science and Informatics,
University of Žilina, Univerzitná 1, 010 26 Žilina, Slovakia
E-mail: {jan.janech, emil.krsak, stefan.toth}@fri.uniza.sk

Keywords: VANET, architecture, DDBS, communication protocols, security

Received: December 20, 2013

This paper describes principles of the data communication in the distributed database system AD-DB developed by the authors. The database system is designed to function properly in such a complex and dynamic network as the VANET is. That way, vehicles connected to the VANET could distribute traffic related data to the others. The paper concludes by proposing a solution for security problems by introducing cross-certificate to our system.

Povzetek: Predstavljena je izvirna arhitektura porazdeljenih podatkovnih sistemov v okolju VANET.

1 Introduction

VANET (Vehicular Ad-hoc NETwork) is a field of important research nowadays [25]. Many researchers are trying to develop new principles to make it possible to distribute information through this network. Applications for VANET could be divided into two categories: *safety applications* and *comfort applications*. Safety applications are more important ones. They are focusing on distributing information about traffic accidents, obstacles and other safety hazards to as many vehicles as possible [13, 14].

VANET is defined to be a special case of MANET (Mobile Ad-hoc NETwork) where network nodes are represented by vehicles in a road traffic. But problems with distributing data in VANET are completely different from the MANET ones. MANET nodes as computers with limited power source and limited computing resources have to communicate in small time frames to preserve as much power as possible. All research of the MANET communication is about minimizing communication and computing time and about conserving node power.

On the other hand, almost all of VANET nodes (vehicles in road traffic, road infrastructure) have good power source. So research in this area is focusing on the best way to distribute information for all nodes that are interested in it.

2 State of the art

2.1 Classic architecture of distributed database system

Architecture of DDBS (Distributed Database System) from data organization point of view is shown in figure 1. It is simple layered model with four layers. Each one of them represents some view on data itself.

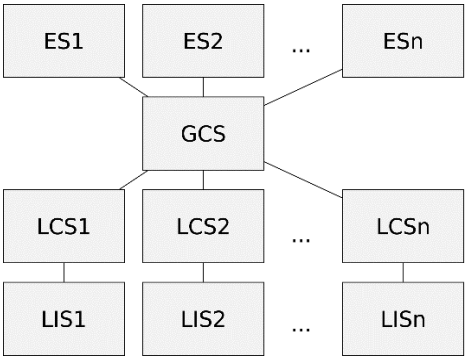


Figure 1: DDBS reference architecture [1].

There are four layers of distributed database system, each modeling one kind of view on distributed database [1]:

1. *LIS (Local Internal Schema)* represents physical representation of data stored at one node. It is analogy of internal schema from centralized databases.
2. *LCS (Local Conceptual Schema)* describes logical organization of data at one node. It is used to handle data fragmentation and replication.
3. *GCS (Global Conceptual Schema)* represents logical organization of data in whole distributed database system. This layer is abstracting from the fact that the database system is distributed.
4. *ES (External Schema)* represents user view into distributed database. Each external schema

defines which parts of database user is interested in.

The fact that the user is using only global conceptual schema through views defined in external schema, assures that the user can manipulate the data regardless of its position in the distributed database system. Therefore it is necessary to have a mapping from every local to the global conceptual schema. This mapping, named GD/D (Global Directory/Dictionary), is defined as part of the distributed database system.

The main role of GD/D is to provide access to mapping between local conceptual schemas and the global conceptual schema. So it has to be accessible from every node sending queries to the system. There are several ways to ensure it [4, 5]:

1. *Centralized directory* – whole GD/D is stored centrally at one node. The advantage of this solution is that it makes GD/D manipulation simpler. However, one central node represents single point of failure for the whole distributed system and can be a bottleneck as well.
2. *Fully redundant directory* – replication of the whole GD/D is stored on every node. That way it can be quickly accessed whenever needed. But its modifications are more complicated due to its multiple occurrences in the system.
3. *Local directory* – every node stores only its own part of the GD/D, so its management is very simple. On the other hand, global query requires communication with other nodes to make possible to create the query plan.
4. *Multiple catalog* – in the clustered distributed database system it is possible to assign whole GD/D replication to one node in each cluster. It is combination of first two ways.
5. *Combination of 1. and 3.* – every node has its own GD/D replication and there is one global replication as well. Each of this possibilities has its pros and cons. But they have all something in common: the system needs to recognize all of its parts.

Whether the GD/D is stored at one node or somewhat distributed through the system, there needs to be some way how to access it as a whole. This is not possible in VANET as there is no way to ensure communication between all of the nodes. In this situation, GD/D cannot be used to locate requested data.

As of the present time there is no solution designed specifically for VANET known to the authors. But there are few solutions for MANET, so we will describe them in next sections of the article.

2.2 TriM protocol

The TriM protocol is the one of first attempts for solving the problem of data distribution in MANET environment the generic way. It was designed as part of a PhD thesis at the University of Oklahoma [6]. The main focus of the protocol is to minimize power consumption and to utilize all three modes of communication [7]:

- *Data Push* represents data distribution using broadcast messages.
- *Data Pull* represents on demand data distribution.
- *Peer-to-peer communication* for querying data.

The main disadvantage of the TriM protocol is its requirement to have same data on all nodes. This requirement makes it practically unusable in the VANET environment.

2.3 HDD3M protocol

HDD3M protocol tries to solve TriM protocol problems. As in the original protocol, HDD3M aims to use all three modes of communication and to conserve as much power as possible. The main difference from the TriM protocol is a possibility to manage database fragments and to modify the distributed database via transactions.

HDD3M divides nodes into 3 categories:

- *Requesting node (RN)* is sending queries to distributed database system.
- *Database node (DBN)* is containing database fragments.
- *Database directory (DD)* stores GD/D for distributed database.

This protocol must solve problems with distribution GD/D. There is no guarantee that all of database directory nodes receive the GD/D update request. Some of the nodes could be inaccessible through MANET or shut down due to lack of energy. When the network is fragmented, keeping the data accurate and actual might be impossible.

The biggest problem for the deployment of distributed databases in the VANET environment is the necessity of the knowledge of all the nodes beimg available in the system. This problem persists in this solution as well because the GD/D is still used.

3 Principles of proposed solution

So the only way to ascertain the use of the distributed database system in the VANET environment is to remove the GD/D from the system and replace it with a different principle. As it has been said already, the GD/D describes the mapping between the local and global conceptual schemes. Without the mapping the system does not know where the data are located and how to query them.

Using the GD/D in the VANET environment is impossible because it requires knowledge of the whole system (*global directory*). In VANET every node knows only its immediate surroundings. So querying a distributed database is fairly limited in such environment. The only nodes which can be addressed to using queries are those in the immediate surroundings in the network. So the system naturally creates virtual clusters of nodes that can communicate with each other. The clusters might overlap, so each of the nodes of the cluster can communicate with another set of nodes.

The only possibility to introduce principles of distributed database systems into VANET environment

lies in allowing to query data only from clusters containing the query node. That way we can replace GD/D with another principle – CD/D (Cluster Directory/Dictionary). But there is still question, how to store CD/D and how to distribute it throughout the database system. The possibilities are same as they were for storing GD/D. They were described in the subsection 1. in section 1 of this paper.

Best possibility for VANET seems to be storing their own part of CD/D at each of network nodes. Other possibilities would be complicated to implement due to highly dynamic nature of VANET.

This is the way distributed database management system AD-DB (AD-hoc DataBase) is working [24]. AD-DB was created as the result of a PhD thesis at University of Žilina [2] by one of authors.

4 Query processing in AD-DB

As we already said, it is impossible to keep CD/D as a whole and distribute it throughout the VANET. Instead of that, AD-DB is using broadcast messages for data communication and lets each data node to decide whether it has requested data or not by looking to its own part of CD/D.

AD-DB supports two methods of communication each based on slightly different principle:

- *Pull method* is an application of pull mode of data communication into AD-DB database. It allows each node to query data from cluster.
- *Push method* is an application of push mode of data communication into AD-DB database. It allows to share own data to other nodes without any prior query.

4.1 The Pull method

The Pull method represents the standard method of query processing in classic distributed database systems. One of the nodes sends query to the system and waits for the results.

The method could be used in such a situation where a client does not have to update the data periodically but needs to query it once instead. One time search for nearby cinemas could be taken as an example of such a situation.

It is also possible to use the pull method as a means for data replication but it is much more ineffective than using the push method [15].

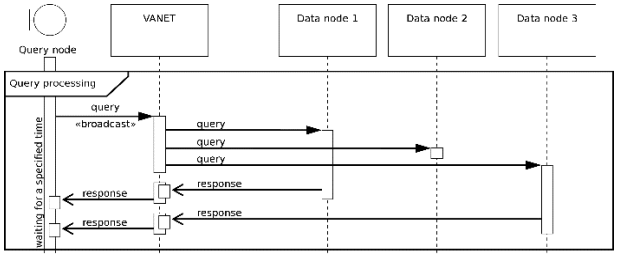


Figure 2: Query processing using the pull method [2, 3].

The query principle is shown on Fig. 2. Communication is done in the following steps:

1. *Global query optimization*. It is important to optimize a query to minimize the size of resulting data.
2. *Sending a query*. The Query node sends optimized query using broadcast message. That way all of the data nodes in cluster receive the query. The query node waits for the specified time.
3. *Query fragmentation*. Every data node which receives the query fragments searches for subqueries that the node is able to execute.
4. *Local subquery optimization*. A data node optimizes each of the found subqueries and prepares it for execution.
5. *Subquery execution*. The data node executes each of subqueries.
6. *Sending the result*. The data node sends back the resulting data together with the identification of executed subquery using the unicast message.
7. *Results evaluation*. After the specified time runs out, the query node evaluates all results received from the data nodes and merges them to one complete result.

4.2 The push method

Given that the organization of the network structure is changing rapidly in VANET, it is clear that sometimes there is a need for querying the same data repeatedly. A possibility of using the push method of data communication in AD-DB can be handy in such situation.

This is also the reason why the push method is more effective to be used in data replication algorithms than the pull method [15].

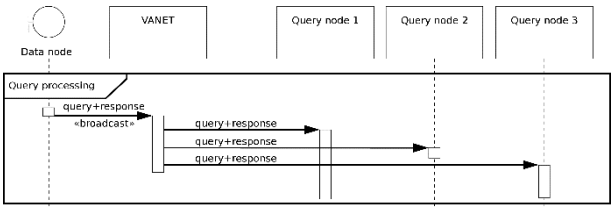


Figure 3: Schematic illustration for push method [3].

The schematic principle of push method is shown in Fig. 3. The Communication is done in the following steps:

1. *Local query optimization*. The data node optimizes the query and prepares it for execution.
2. *Query execution*. The data node executes the optimized query.
3. *Sending the data*. The data node sends resulting data packed with the query through VANET as broadcast message.
4. *Results evaluation*. When the query node receives the data, it analyzes the attached query to determine whether it needs the data or not. If

it needs the data, it forwards the data to the user application to process it.

5 High level communication protocol for AD-DB

Schematic representation of the communication protocol used in AD-DB is shown in Fig. 4. The data node can process the processQuery message. This is sent by a query node in the form of a broadcast message in the pull method of communication.

The query message has following structure [2]:

- *Schema uuid* is a unique identifier of the current database schema. It is important to include this for the data node to be able to determine whether it should process the query or not.
- *Serialized query* represents the query itself. The best way to transfer the query is in a form of serialized abstract syntax tree, as it is easy to process by the data node.

There is no need to transfer the session identifier of any kind, because the query and uuid could be used as a unique identifier of the request.

The response message structure is as follows [2]:

- *Schema uuid* as part of response unique identifier.
- *Serialized query* as part of response unique identifier. It is possible to use the query as part of unique identifier because the query processed by the database system is expected to be simple and short. If this assumption was not true, it would still possible to use value computed from the query by some hash function instead.
- *Query part* is the identifier of the processed subquery.
- *Data* as a collection of the resulting objects.

Using the schema uuid and query pair as a unique identifier of a request has one advantage over using surrogate identification number. This way the response message format can by the same for the pull and the push methods of communication.

Important part of a response message is the query part identification. It represents a unique identifier of query part processed by a data node as a subquery. This identifier is needed by a query node to be able to merge all responses from all responding data nodes.

There are two possibilities how to use the same system of numbering for all query parts by both the query and the data node:

- Inserting the identifier directly into the serialized query. The process of identifier inserting is done directly by the query node after the global optimization. Example of a query with identifiers (syntax of the query language used by AD-DB is published in multiple publications by authors [2, 3, 8]):

$$^{(1)} \bowtie (^{(2)}projects//(\lambda x|x \lhd name{=}\text{KANGO}),^{(3)}employees)$$

where (1) identifies the whole query as one part, (2) identifies collection of all projects with the name KANGO, and (3) identifies collection of all employees.

- Automatic numbering of all operations by theirs priority. The priority of an operation can not change as it is defined by the query language, so the numbering will be same on both query and data node. This system is preferred and it is used by AD-DB as it does transfer slightly smaller quantity of data between the query and data node.

The push method is using the processResponse message. It is sent by the data node in the form of broadcast message.

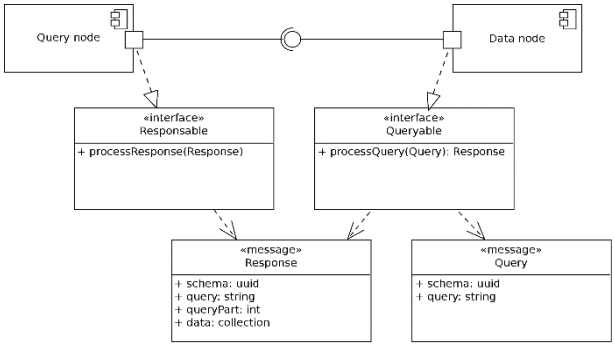


Figure 4: Schematic representation of communication protocol used by AD-DB [2].

6 The OSACP protocol

OSACP (Object Structure Aware Communication Protocol) is an application protocol designed specifically to transfer structured data through VANET. It is designed as part of PhD thesis at University of Žilina [9]. OSACP is using UDP transport protocol on top of IPv6 network protocol. Its design allows it to transfer any structured data through VANET and reconstruct it on the other side even if part of data was not transferred correctly [10, 3].

Missing parts of the structure are replaced by special object UNKNOWN to indicate incomplete message. It is up to the user of the distributed database (person, or another application) to decide whether it can process the message or not.

7 Security of DDBS

Since VANET is a very dynamically changing system, we must pay special attention to the communication security. Connections between network elements are constantly changing according to the current position of the elements and the impact of their communication devices. Looking at VANET as a distributed system without the possibility of at least partial centralization presents a high risk of abuse (threats model, authentication, privacy, secure identification of the position, etc.). It is necessary for network elements to create an appropriate security architecture that will protect the participants from various types of attacks.

Safety aspects that are required in networks can be divided into 2 areas:

- Availability

- Authentic information and privacy

On that basis, we can define commonly known security threats and divide them into the following categories.

7.1 Threats to availability

Denial of service – overload the node resources so it is not able to perform other important and necessary tasks. It occurs very often as distributed denial of service. Similar situation can be made by jamming the channel.

Black hole attack – an attack by misbehavior of the node. It can forcibly redirect network traffic to a non-existent node by falsifying the routing information, causing the data to be lost.

Malware – malware attacks, such as viruses in VANETs, have the potential to cause serious disruption to its normal operation. Malware attacks are more likely to be carried out by a malicious node. The attacks may be introduced into the network when VANET units in cars and roadside station receive software updates or special plugins.

GPS spoofing – using GPS simulators to generate fraudulent signals so nodes believe they are in different location or in different time.

Broadcast tampering – an attacker can inject false traffic safety message into the network. Broadcasting this message can cause accidents or manipulating the flow of traffic to clear chosen route.

Spamming – spam messages on VANETs elevate the risk of increased transmission latency. The lack of centralized administration causes serious problems in VANET.

7.2 Threats to authentication

Masquerading – an attacker presents itself as legitimate node in the vehicular network by using false information or by using message fabrication, message alteration, or message replay. For example, an attacker acts as an emergency vehicle to mislead other vehicles to slow down and yield [16].

Replay attack – this attack happens when an attacker replay the transmission of earlier information to take advantage of the situation of the message at time of sending [17].

Sybil attack – in this attack type, a node sends multiple messages to other nodes and each message contains a different fabricated source identity in such a way that the originator is not known [18, 19]. The basic goal of the attack is to provide an illusion to other nodes about a traffic jam, and force them to take an alternate route

Message Tampering – this type of attack is changing messages while being transferred from a source to theirs destination. Everyone within the same zone in VANET can listen to all messages sent by other users. Thus, malicious users can modify the contents of a message before it is received by real destination [20].

ID Disclosure – is about disclosing the identity information. Using this method, attacker can track the current location of his target [20].

7.3 Threats classification

From this point of view we should define which attack described in sections 7.1 and 7.2 is mapped to which layer of our distributed database system. All of them should be expected only on two lowest layers. One can expect these type of attack form group of *Threats to availability* to the LIS layer:

- Denial of service
- Black hole attack
- Broadcast tampering
- Spamming

From group of *Threats to authentication* come into consideration:

- Masquerading
- Reply attack

On the *second level LCS* (Local Conceptual Schema) it can be form group of *Threats to availability*:

- Malware
- GPS spoofing

And from group of *Threats to* authentication:

- Sybil attack
- Message Tampering
- ID Disclosure

Many of these attacks can be avoided by using PKI (Public Key Infrastructure). In conventional database systems PKI and cryptography ensure the highest layers. The VANET situation is reversed. We need to ensure the lowest layers.

Classic PKI is composed of CA (Certification Authority) tree with one root CA. Each CA has a network of RA (Registration Authority) to verify the applicant's identity certificate. Security is based on cryptography. The parent node in the hierarchy "Node certificate - CA certificate – root CA certificate" guarantees the public key and identifies subordinate by form of a certificate that is signed electronically document. In the case of private key compromising the certificate relevant node or CA needs to be revoked.

When using PKI in VANET, we must take into account the specificities that make it impossible to use certain features and protocols used in PKI.

7.4 Local storage certificates of root CAs

Many CA roots are expected in VANET and therefore fluctuation of CA is very likely. In environment, with many root CAs, it is very complicated to store every root certificate in local storage in every node. On the other hand it is possible to use cross-signed certificates so a node in its local storage must have only one or few root certificates. When node receives signed message, it receives also chain of certificates of CA hierarchy and if there is a root CA that believes it (in its local storage), message is valid.

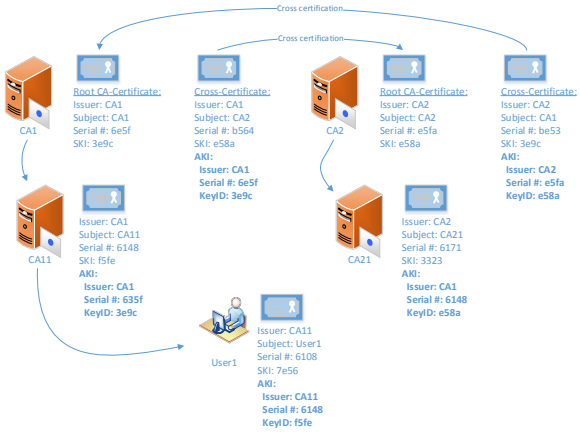


Figure 5: N tier cross certificate [21].

7.5 Revocation of certificates

In standard PKI there is a mechanism to revoke certificate by issuing CRL (Certification Revocation List) by CA. It contains a list of every revoked certificate. It can be the absolute list or the differential list of revoked certificates from last published CRL. There is also OCSP protocol (Online Certificate Revocation Protocol), which is used for online checking of validity the certificate. In VANET OCSP has very low performance [22, 23] and is not very useful.

For our distributed database system we recommend to use the cross-certificate and the absolute CRL. Using the absolute CRL we risk very large CRL. It depends on the length of certificate validity. The short period of validity of the certificate reduces probability of compromising the private key, but increases overhead and re-applying for the issuance of the certificate. For that process it can help implementation of automatic renewal of certificates by CA. Choosing a suitable length of validity of the certificate depends on the particular use of DDBS.

8 Conclusion

There is no known distributed database systems that would be possible to operate in the VANET environment. There are some attempts to do so for MANET, but they are unusable for VANET.

The paper presented the communication system of the distributed database system AD-DB. The database system is designed to be used in the VANET environment and so its basic principles had to be altered for such usage.

In the nearest future we would like to focus on enhancing the query optimization algorithms, but there are many other areas which would be interesting to explore. For example, many of the data in VANET are of highly temporal character, e.g. current weather, traffic flow speed, traffic obstacles, etc. It would be interesting to have a possibility to query current state of those temporal data.

We have some accomplishments in this area even now. We have designed system to query visual objects recognized by vehicle cameras through VANET [11, 12], so the next logical step would be to integrate this system into distributed database system AD-DB.

Acknowledgment

This contribution/publication is the result of the project implementation:

Centre of excellence for systems and services of intelligent transport II., ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

"Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

References

[1] M. T. Ozsu, P. Valduriez (2011). *Principles of Distributed Database Systems*. 3rd ed.

[2] J. Janech (2010). *Riadenie procesov pri distribúcii databáz* (Data distribution process control). PhD disertation, Dept. of Software Technologies, University of Žilina, Žilina, Slovakia.

[3] J. Janech, T. Baca, A. Lieskovsky, E. Krsak, K. Matiasko (2013). Distributed Database Systems And Data Replication Algorithms For Intelligent Transport Systems. *Communications: Scientific Letters of the University of Žilina*. Vol. 15, No. 2.

[4] P. Sokolovský, J. Pokorný, J. Peterka (1992). *Distribuované databázové systémy* (Distributed Database Systems). 6th ed.

[5] C. J. Date (2003). *An Introduction to Database Systems*. 8th ed.

[6] L. D. Fife (2005). *TriM: Tri-Modal data communication in mobile ad-hoc network database systems*. Ph.D. disertation, University of Oklahoma.

[7] L. D. Fife, L. Gruenwald (2003). Research issues for data communication in mobile ad-hoc network database systems. *SIGMOD Rec.*, Vol. 32, No 2.

[8] J. Janech, A. Lieskovský, E. Kršák (2012). Comparation of Strategies for Data Replication in VANET Environment. *26th International Conference on Advanced Information Networking and Applications Workshops* (WAINA).

[9] T. Bača (2012). *Optimalizácia prenosu správ v ad hoc sieťach* (Optimization of Message Distribution in Ad-hoc Networks). PhD dissertation, Dept. of Software Technologies, University of Žilina, Žilina, Slovakia.

[10] T. Bača (2012). Optimisation of message distribution in Ad-hoc networks. *Information Sciences and Technologies : bulletin of the ACM Slovakia*, Vol. 4, No. 4.

[11] Š. Toth, J. Janech, E. Kršák (2013). Query Based Image Processing in the VANET. *5th International Conference on Computational Intelligence, Communication Systems and Networks* (CICSyN2013).

[12] Š. Toth (2013). *Spracovanie obrazu s využitím dopytov v prostredí VANET* (Query Based Image Processing in the VANET). PhD dissertation, Dept. of Software Technologies, University of Žilina, Žilina, Slovakia.

[13] E. Kršák, P. Hrkút, P. Vestenický (2012). Technical infrastructure for monitoring the transportation of oversized and dangerous goods. *Federated Conference on Computer Science and Information Systems* (FedCSIS 2012).

[14] S. Badura, A. Lieskovsky (2010). Intelligent traffic system: Cooperation of MANET and image processing. *1st International Conference on Integrated Intelligent Computing* (ICIIC 2010).

[15] T. Bača (2011). Data replication in distributed database systems in VANET environment. *Proceedings of 2011 IEEE 2nd international conference on software engineering and service science*, Beijing, China.

[16] A. K. K. Aboobaker (2010). *Performance analysis of authentication protocols in vehicular ad hoc networks (VANET)*. Technical Report, Dept. of Mathematics, University of London.

[17] B. Parno and A. Perrig (2005). Challenges in securing vehicular networks. *The Fourth ACM Workshop on Hot Topics in Networks* (HotNets-IV).

[18] F. Sabahi (2011). Vehicular Ad - hoc Networks Security Analysis. *International Conf. on Computer Engineering and Applications* (ICCEA).

[19] S. Zeadally, R. Hunt, Y. S. Chen, A. Irwin, and A. Hassan (2010). Vehicular Ad Hoc Networks (VANETS): Status, Results, and Challenges. *Telecommunication Systems.* Vol. 50, Issue 4, pp 217-241.

[20] F. Sabahi (2012). Impact of Threats on Vehicular Adhoc Network Security. *International Journal of Computer Theory and Engineering*, Vol. 4, No. 5.

[21] C. L. Mankowski (2012). Answer on post Can a certificate have multiple chains and multiple self-signed roots? http://security.stackexchange.com/questions/15562/can-a-certificate-have-multiple-chains-and-multiple-self-signed-roots

[22] J. Serna-Olvera, V. Casola, M. Rak, J. Luna, M. Medina, and N. Mazzocca (2010). Performance Analysis of an OCSP-Based Authentication Protocol for VANET, *Int. J. Autonomous and Adaptive Communications Systems*. Vol. 3, No. 2, pp 19-45.

[23] K. Papapanagiotou, G. F. Marias, P. Georgiadis (2007). A Certificate Validation Protocol for VANETs. *IEEE Globecom Workshops.* Washington. pp: 1-9.

[24] J. Janech, Š. Toth (2013). Communication in Distributed Database System in the VANET Environment. *Federated Conference on Computer Science and Information Systems* (FedCSIS 2013). pp. 795–799.

[25] K. Mäkká, J. Dicová (2013). Possibility of using the software product AIMSUN in the process of modelling transport. *Proceedings in ITS 2013 - Intelligent Transportation Systems 2013, Virtual Conference*.

Prototype Implementation of a Scalable Real-Time Dynamic Carpooling and Ride-Sharing Application

Dejan Dimitrijević, Vladimir Dimitrieski, and Nemanja Nedić
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000, Novi Sad, Serbia
E-mail: {dimitrijevic, dimitrieski, nemanja.nedic}@uns.ac.rs

Keywords: real-time, carpooling, ride-sharing

Received: September 29, 2014

Setting out to build a real-time carpool and ride-sharing solution, which would be able to attain a global user base and is initially designed as highly-scalable, this paper describes some of the selected designed concepts, distribution and cloud computing strategies needed to do so. Our selections were based on experiences of others with same or similar-purpose solutions which were developed in the previous decade. Some of these solutions were either outdated, mostly by leaving its users with a subpar user experiences as their user base grew, or outgrown by having limited client reach, leaving them available only to a small portion of mobile client devices or desktop browsers and users. This paper presents an implementation of a ridesharing application prototype that follows all of aforementioned strategies. The goal of this prototype is to show that it is possible to make very scalable and ubiquitous ridesharing application, which is able to successfully reach and serve a global user base.

Povzetek: Članek predstavlja nov prototipa aplikacije za delitev stroškov prevoza z avtomobilom.

1 Introduction

This paper is a follow-up to a paper presented at the FedCSIS'13 conference [1] which extends further upon it with some experimental data and even more prototype implementation details. The aforementioned paper explored some currently available positions concerning an implementation of any present and future carpooling and ride-sharing applications. These applications should be real-time, dynamic (more about meaning of dynamic in following sections) and scalable enough to reach a worldwide audience. The explored positions were chosen so that any such application could and should work without either omitting support for any future mobile platform or leaving its clients with deteriorating quality of service as it client base grows. Because of the prior of the two just previously noted requirements, we identified some novel client web technologies which are or will be available on all modern mobile platforms. Therefore, we felt confident that any real-time dynamic carpool and ride-sharing application built upon these technologies could be ubiquitous enough. By ubiquitous we mean making it available across both various mobile and desktop platforms, current and future ones. Also, because of the latter requirement, which was also explored and outlined in our previous paper, some effort was also put in identifying server side technologies capable of producing a low-cost development and maintenance real-time dynamic carpooling and ridesharing mobile and web application. Considering existing solution experiences, primary idea was to build a scalable solution that could start out small but allow easy growth later. This is to accommodate the fact that the ride-sharing industry has only recently started becoming globally interesting. However, carpooling formally appeared in the US in the mid-1970s, after the 1973 oil crisis [2]. At that time the rising costs of using a personal vehicle for transportation of only one passenger made it prudent to drive more than one person, usually co-workers commuting to and from same workplace, splitting transportation costs. However, the reduction of oil and gas costs in the 1980s and the breakdown of a typical 9AM to 5PM workday in the 1990s led to a spiral down trend in carpooling popularity. Federal government in the US tried to counter such a trend by incentivising carpooling drivers, growing the number of no-toll carpool lanes across many highways. Those lanes were also allowing for relief from ever growing traffic jams and gridlock, as the number of vehicles on the roads increases. In 2000 it exceeded 740 million globally [3] and projections are there will be over 2 billion motorized vehicles by 2030 [4]. Large number of vehicles creates many well-documented problems for urban areas, such as increased traffic, increased pollution, parking congestion, and the need for expensive infrastructure maintenance. To reduce all of those and personal transportation costs also, we set out to create a low-cost, ubiquitous and global audience capable real-time carpool and ride-sharing solution.

1.1 Problem

The expenses, both environmental and fiscal, of single occupancy vehicles can be reduced by utilizing more of the empty seats in personal transportation vehicles. Carpooling and ride-sharing targets those empty seats: taking additional vehicles off the road thus reducing traffic and pollution, whilst providing opportunities for social interaction. However, historically carpool scheduling often limited users to consistent schedules and fixed rider groups–carpooling to the same place at

the same time with a set person or a group of people. To make that problem worse, the leading problem concerns, given in a 2009 survey about why people don't carpool, were difficulty to organize carpools and the inconvenience of organization far in advance [5].

Due to aforementioned reasons, this paper proposes some main guidelines and cloud distribution strategies that we fell will bring best value for any future global carpool and ride-sharing solution.

The rest of the paper is organized as follows: Section 2 provides an overview of related work. In Section 3 we present overall design concepts and our objectives. Section 4 elaborates on our proof-of-concept prototype system implementation choices, with subsections focusing on several specifics. Section 5 concludes the paper and discusses our future work.

2 Related work

In this section we present existing ridesharing and carpool solutions and research. We have divided this section into two main parts: Subsection 2.1 reviews carpool and ride-sharing related solutions currently available and Subsection 2.2 surveys some of the literature and papers on the subject.

2.1 Current carpool and ride-sharing solutions

Entering carpool and ridesharing search terms in some of the largest mobile app store and internet search engines returns a great deal of mobile apps and internet websites offering either classic or dynamic carpooling and ride-sharing. By the notion of classic carpool mobile app or website, we denote applications within which users just schedule and advertise their plans for a trip well in advance. This is accomplished effectively through an, in essence, a searchable electronic bulletin board, seeking other users travelling in the same direction at the same time. Although some of those apps and websites, such as carpooling.com [6] and its mobile client apps have a large user bases, the static routing problems they help solve makes their usage fairly limited.

The inconvenience of having to search though large carpools or even smaller but fixed choice driver groups, hoping to amongst them find a pre-scheduled and advertised trip adequately consistent with one's schedule, makes classic carpooling apps or websites non-practical for relatively short and near-immediate on-the-go carpool and ride sharing trip plans. It is for that reason that even a solution presented on [6] and its large network of European subsidiary websites, added advanced time constrained search features. These features are used to "find a lift", which to a certain extent alleviate some of the inconveniences for their on-the-go passenger users. However, the added hourly time-constrained advanced searching still inconveniences their vehicle driving users to be mindful of their advertised pre-trip given schedules, even though that may not always be objectively possible, due to unforeseen events such as: road accidents, gridlocks, etc.

Thus, a new form of dynamic carpool and ride-sharing mobile apps and websites is emerging. They are indicated by their use of real-time passenger requests along with real-time vehicle driving users' location data, foregoing the need for well in advance pre-scheduled and advertised trips. Amongst some of the most known and pioneering mobile apps and websites offering dynamic carpooling and ride-sharing are Lyft [7] and SideCar [8]. Both mobile apps are available for iOS as well as for Android, but neither app currently has a web browser user interface. This may be intentional since both are fully natively written, and by our observations both use TCP sockets to communicate with their respective backend services. Therefore both would need changes to make them web browser-friendly. Unfortunately, those code changes could include some rather tedious transformations. This is due to the fact that native TCP socket traffic has not been well suited for consumption in web browsers relying mostly on HTTP, until recently. Another popular application and website is Waze [9], which isn't intended for carpool and ride-sharing. Waze is used for gridlock traffic reporting and avoidance, but accommodates for pickup requests also, and it seems to have taken another approach in comparison to the two aforementioned applications. Although Waze mobile apps are also fully natively written, their website presents a "live map" user interface which maps events reported by their users. Such events include pickup requests and replies of passenger and vehicle driving Waze users who are otherwise linked either via a popular social network, or through email and SMS. Fortunately, access to those real-time events has now been extended from its native app users to even non-Waze users through private URLs which lead to a live map connected to the Waze backend via a GeoRSS [10] feed through HTTPS (Figure 1).

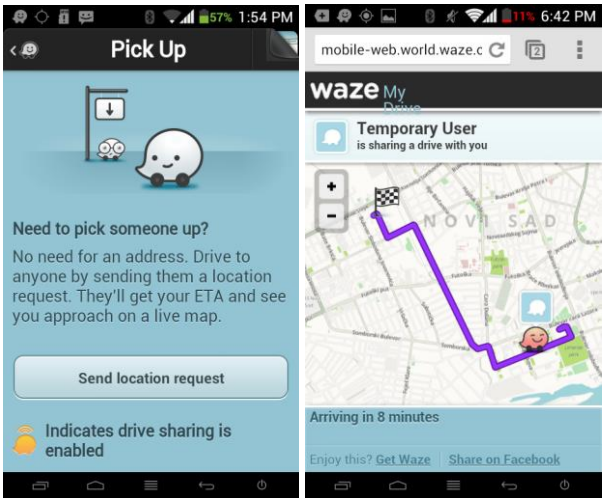


Figure 1: Waze "pick up" and mobile "live map" UI.

The original GeoRSS XML format is however transformed to JavaScript Object Notation (JSON) presumably for easier (JavaScript) client-side consumption and traffic overhead reduction. However it is still limited by a set update interval time. To our knowledge, there are no other globally popular websites

and mobile apps that currently allow for carpool and ride-sharing uses, excluding commercial taxi dispatchers.

2.2 Current carpool and ride-sharing papers

Because static carpool still represents the majority of existing solutions, almost all of the available papers and literature on carpool and ridesharing mainly deal with the static ridesharing issues. In the static carpooling users must pre-schedule their trips, neglecting the dynamic aspect. Despite much of the progress experimented on dynamic carpooling and ride-sharing concepts, it still remains in the early stages regarding publicly available works and literature. In order to make up for that shortfall, some of the papers which mention carpooling and ride-sharing, and even consider the dynamic aspect, like [11], also considered other issues at the same time. Some papers are especially involved in the concepts of traceability, communication and security services. Their authors feel that none of the current solutions evoked these concepts, identifying the security issues as one of the main reasons hindering their success [12]. All cited papers provided us with a lot of beneficial ideas and food for thought transferred onto this paper. They also influenced this paper's findings and conclusions, and out of that still quite disorganized literature, we have identified some yet non-tackled issues, laid out in the following sections. Mainly, we take issue with web browser user interfaces and standardized web technologies which seem to be the unifying way forward, putting ubiquity in the grasp of every hybrid web and mobile application.

3 Design considerations

In the previous section we presented some of the carpooling and ride-sharing solutions, ideas and issues tackled recently. In this section we are building up on those solutions and ideas, proposing some of our own design concepts for a global dynamic real-time carpooling and ride-sharing solution. In Subsection 3.1 we describe some of design concepts that are suitable for a real-time dynamic carpooling and ride-sharing solution. Subsection 3.2 further describes our ideas, allowing for the proposed real-time solution to tackle the problem of being able to serve up to a global user base, adding cloud and distribution design concepts. Finally, in Subsection 3.3 we deal with the ubiquity problem, considering the client user interface technology we feel will be future-proof and available on almost all new mobile and desktop platforms.

3.1 Real-time dynamic solution design concepts

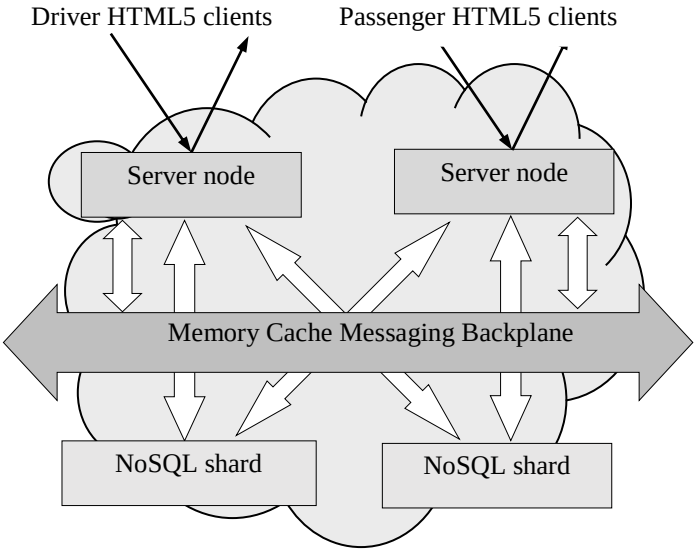
As it was noted in Section 2, real-time dynamic carpooling and ride-sharing solutions are becoming more common. However it takes more designing effort to achieve real-time dynamic capabilities than for mere static carpooling and ride-sharing. The reason for the recent increase is obviously because real-time dynamic solutions are more convenient, and thus more likely to be used in greater numbers by end users. Additionally, some technologies previously used for seemingly real-time communication on the web, have only recently matured and have been standardized.

In the beginning of a so called Web 2.0, at the time when real-time updating websites were only just starting to appear, most of those websites used Asynchronous JavaScript and XML (AJAX) [13]. AJAX is a group of interrelated web development techniques used on the client-side to create asynchronous, seemingly real-time web applications. Most of those techniques relied upon regular HTTP, a simple request-response and stateless protocol. Having to achieve what was usually two-way communication took some effort for websites and web applications. Developers were using various workarounds, techniques involving the use of the browser XmlHttpRequest object or some other web browser plugins.

The first workarounds developed into techniques known as: frequent polling, long-polling and the so called forever-frames. Although all of those techniques were, and still are, very much usable for seemingly real-time web page updates without requiring full page refreshes, they had drawbacks. Their primary drawback was the amount of server-side and network resources they consume. The server is either forced to respond to a large number of frequent requests, or it opens up a number of long running responses, which additionally occupy its hardware resources. On the other hand, using workarounds such as various browser plugins, although they used less network and server-side resources, turned out to be non-practical, because of the lack of plugin support on current mobile devices. For such reasons, new techniques were developed, and recently standardized by the World Wide Web Consortium (W3C). As part of the HTML5 specification Server-Sent DOM Events (SSE) were standardized in 2011 [14], but have not yet been implemented by all desktop browsers, namely, Internet Explorer. However, Web Sockets API [15] was drafted back in 2009 and it is currently supported by all major web browsers. Web Sockets provide a full-duplex communication channel over a single TCP connection, thus allowing for a lower network latency time due to less traffic overhead compared to HTTP. Compared to SSE and other polling techniques, Web Sockets provide the best option for building real-time communication on the web. Due to aforementioned reasons, this protocol is an integral part of our proposed design.

3.2 Distribution and cloud design concepts

Having chosen Web Sockets (WS) as a preferred means of communication, although helping solve latency issues, left another issue unsolved. WS based communication, as all others, supports clients' connections to a single server node. Even though a number of users may be greater when using WS, it still depends on available server hardware resources. Since vertical scaling of server hardware resources can be expensive and still ultimately limiting, the best solution to the near-infinite scaling



problem is horizontal distribution, across multiple server nodes. Ideally, any global real-time solution would be best served in one's own server farm, but given hardware and its maintenance costs, renting cloud resources is a more realistic option. However, for horizontal scaling, one needs to be able to scale data also. Since traditional, i.e. relational data scaling is much harder [16], we have turned to non-relational data (NoSQL). NoSQL databases allow easier scaling and offer better performance in data writes. Additionally, they offer a possibility of scaling reads onto multiple database nodes, combining so called sharding and some parallelism approaches. The notion of sharding denotes horizontal distribution of data across multiple servers. Utilizing aforementioned advantages in a document oriented NoSQL data store that supports geospatial data indexing would make it a perfect fit for our proposed solution and storing our users' location based data. Also, a key-value memory caching NoSQL

Figure 2: Basic architecture design.

data store could be used as a messaging backplane for communication between our individual server nodes, but that use and setup is trivial in the case of our chosen real-time library.

3.3 User interface architecture design concepts

To make a website or mobile app truly ubiquitous, one needs to support as many different desktop and mobile platforms as possible, ideally all of them. Although client applications can be natively written for each platform, there are some unifying user interface technologies for almost all current desktop and smartphone mobile platforms.

So, to achieve ubiquity, we propose the usage of combined HTML5/CSS3 canvas map and form elements for user interface (UI) rendering. Building the UI around streamed real-time data flows of state changes created by passenger and vehicle driving user events is also why the paradigm of reactive programming seems to be the perfect choice. Reactive programming should not be confused with responsive web design, which is also utilized, for same UI reuse across various device screen resolution sizes. Any changes in state registered by user client applications are asynchronously processed when transferred preferably by a Web Socket (as it offers lowest latency compared to other real-time web transport mechanisms) to cloud server nodes and back again to either desktop or mobile clients which are schematically displayed in Figure 2.

4 Prototype implementation

This section describes in more detail implementation choices we made to build the prototype of our distributable cloud-based dynamic real-time carpooling and ride-sharing solution. Subsection 4.1 describes our prototype's real-time communication transport library. Subsection 4.2 deals with our use of geospatial indexed data and gives a comparison of the previously used relational database solution, along with future used NoSQL database and a fast memory caching messaging backplane solution. In Subsection 4.3 we present UI implementation details.

4.1 Real-time communication

During our previous research, we have helped develop the first online taxi dispatching solution in Serbia, named TaxiProxy [17]. This solution was fully realized in .NET and is cloud-hosted on Microsoft Azure. Our participation on this project influenced a lot of our primary technology choices for the prototype of a real-time dynamic carpooling and ride-sharing solution described in this paper.

As noted in the design section, the need to have a real-time dynamic carpooling and ride-sharing solution is imperative, since those solutions are what most users currently wish to use. To make our prototype solution real-time capable, the choice to implement it using a library capable of WS protocol communication in .NET came down to a library named SignalR [18]. SignalR is an open-source library for ASP.NET adding real-time web functionality to .NET applications. It also allows for server-side code to push content to the connected clients

as it happens, in real-time. SignalR server is capable of supporting clients written in several programing languages including .NET and JavaScript. The server-side code can push content to those connected clients by a number of transport techniques, most suitable being bi-directional WS, if available. For a server-side the WS transport requirements are either a self-hosted ASP.NET 4.0+ application or one hosted within Internet Information Services (IIS) 8. Server-side hosted on earlier versions of IIS falls back to other means of message transports. For clients, WS requirement issue is a bit lengthier to describe, so it is listed in better detail in the client UI subsection.

SignalR library allows for two types of implementation approaches. First, less abstract, uses persistent connection, a low level class option, which offers basic real-time mechanisms. It just provides mechanisms to notify the server-side backend of client connection and disconnection, receiving and sending asynchronous individual and bulk messages to connected clients. However, second option, named SignalR hubs, provides an easier to use development interface. It allows abstracting away the need to serialize and deserialize request and response messages manually. The power and ease of use of the second option made it a prudent choice for the speedy development of our prototype application. Even though SignalR hubs allow for the collective segregation of the real-time connected clients into groups, such as a driver and passenger group, the connected clients' membership in those groups is unfortunately not automatically persisted. Also, a dependency resolving GlobalHost server-side object, which is just an implementation of the service locator pattern, can also be easily used to communicate real-time messages from one hub onto another. It is for that reason that we decided to split drivers and passenger clients onto two distinct SignalR hubs: a Passenger and a Driver SignalR hub. That choice allows easy implementation of membership persistence later on, which is important when scaling the application onto multiple server nodes. Persisted group membership allows for a single server node instance to fail without losing all of the membership data of our perspective clients connected to the failing node. As each server node replicates the original driver and passenger SignalR hub architecture, any passenger or driver clients connected to a failing node server should be unaware of any server-side problems, this due to the fact that the load balancer would then just instruct them to another fully working server node and its SignalR hubs.

4.2 Database implementation and performance considerations

For this prototype implementation we have identified two main entities:

- Driver entity which comprises main identifier, position and availability. For each driver, SignalR updates drivers' geospatial coordinates, i.e. position, based on a signal it receives from drivers' devices. When a driver picks up enough passengers, to utilize desired number of empty seats, it becomes unavailable for other passengers to join.
- Passenger entity comprises main identifier and position. Passengers' position is also updated through SignalR and data received from passengers' device.

We have chosen two databases with different data models for this prototype implementation. As we have already implemented our previous project TaxiProxy on Windows Azure platform, we have chosen Windows Azure SQL Database (SQL Azure) as relational database to test in this research. Additionally, we have decided to use one NoSQL database as we plan to make our solution scalable and deploy it in a distributed environment. There is plethora of available NoSQL databases, but we have opted for MongoDB [19] due to our previous experience with this database. Since IIS 8 was our prototype's hosting platform of choice, it should be noted that the server node could then only have been hosted within the Windows 8 / Server 2012 OS platforms. Fortunately, Windows Server 2012 was made available to end-users on the Windows Azure cloud platform as a virtual machine operating system choice since late 2012, and it is deployable onto an Extra Small low-cost machine instance. Extra Small Windows Azure instance, which entails a shared core processor with only 768MB RAM, may not be the first choice from a performance standpoint. However, it is sufficient for proof-of-concept deployments and for limiting cloud-hosting costs.

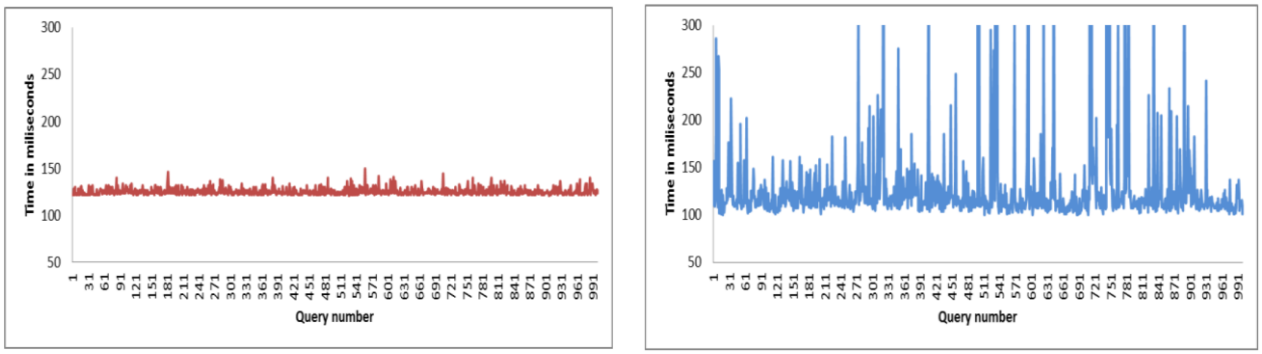
SQL Azure was our first choice to store data as we have used Windows Azure cloud to implement other parts of this prototype. In order to support faster and easier distance calculations between drivers' and passengers' positions, we have used SQLGeography data type for columns containing geospatial data. This data type offers a number of methods for creating and manipulating geospatial data. In order to speed up query execution over geospatial data eve further, we have created indexes on these SQLGeography columns.

One of main motivations to choose MongoDB database came from the fact that MongoLab [20] service offers up to 500MB of free storage for a single-machine instance of MongoDB. This is a major advantage as we want to keep our ridesharing system's costs at minimum to allow cost-free usage of our application for end users. The only drawback is that it is deployed on a single machine which doesn't fully utilize MongoDB's functions as it is supposed to run on a distributed environment. MongoDB is a document oriented NoSQL database and thus we have created driver and passenger documents to store appropriate data. For each driver and passenger data is stored in JSON. Driver's and passenger's positions are stored as longitude and latitude pairs. Similarly to SQL Azure, we have created geospatial indexes to speed up execution of queries.

In order to see which database suits our needs better, we have decided to test these databases based on two criteria:

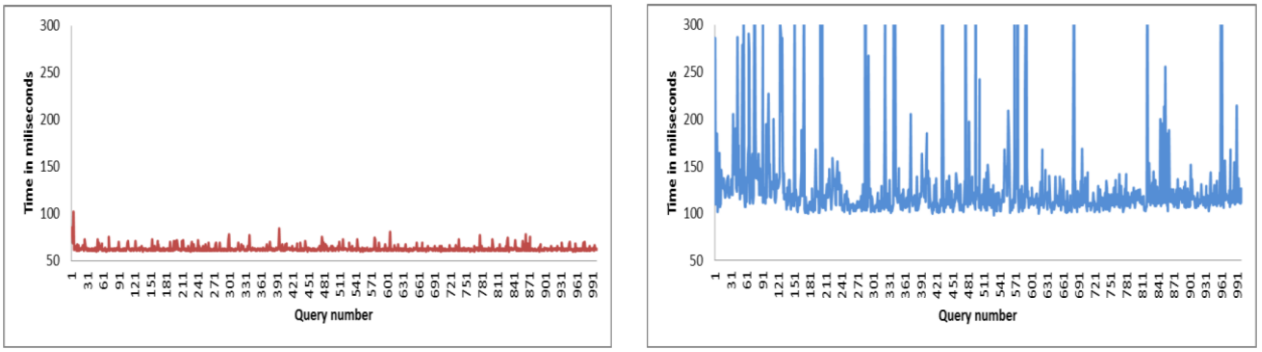
1. *Time needed to execute geospatial query*. This is tested with the most used query in our system that finds five nearest drivers for a passenger.
2. *Amount of storage space needed to store all passengers and drivers*. As the cost of renting cloud based storage is proportional to size of data, we want to keep cost and thus storage size as low as possible.

To prepare a test environment for testing these two database instances, we have set up as similar instances as it is possible with two different data models. We have populated both databases with the same amount of data: 10000 drivers and 20000 passengers. Locations are randomly assigned to all entities. All locations are distributed uniformly within the city area of Novi Sad, Serbia. Both servers, Mongo DB and SQL Azure, are located in EU-West region (geographically closest to us) and are deployed on a single machine. Single machine deployment was chosen as we want to keep service costs as low as possible. Both databases were accessed from a computer with 8GB of DDR3 RAM, 2.67GHz quad core CPU and with 10Mbs/1.5Mbs bandwidth. We have executed 1000 geospatial queries, querying for five nearest drivers to a random location in Novi Sad. Each query was executed using an index on geospatial data. We have taken into consideration only the queries that needed between 50ms and 300ms to execute.



a) MongoDB *b) Windows Azure SQL Database*

Figure 3: Query execution time including time needed to open connection to a database



a) MongoDB *b) Windows Azure SQL Database*

Figure 4: Query execution time without time needed to open connection to a database

In Figures 3 and 4, query execution times are depicted for aforementioned databases. In Figure 3 we have presented query execution times that include time needed to establish and close connections to appropriate database. Both databases performed similarly if we take average times into consideration. MongoDB needed 124.22ms to establish connection, execute query and iterate over results. For the same process, SQL Azure needed 124.83ms. The only difference in these tests is the fact that all query execution times on MongoDB had a small dispersion unlike SQL Server query times. Of all queries on SQL Azure, 33 queries took over 300ms to complete, while all queries on MongoDB were completed in the required time frame between 50ms and 300ms.

In Figure 4 we have presented query execution times without taking into consideration time needed to establish and close connections databases. MongoDB was twice as fast as SQL Azure in these tests. MongoDB needed 62.95ms to execute query and iterate over results. For the same process, SQL Azure needed 122.43ms. Similarly to the test that included time needed to establish and close a connection, MongoDB again had a small dispersion unlike SQL Azure queries. While querying SQL Azure, 20 test cases took over 300ms to complete, while all queries on MongoDB were completed in the projected time frame.

In addition to previously described query time considerations, other main factor in choosing database for final product is the size its data are occupying. Both Windows Azure and MongoLab have limitations on the amount of data their cheapest options can store. As it is already stated, we have populated both databases with the same amount of data: 10000 drivers and 20000 passengers. This data, stored in SQL Azure occupies 1598 KB. At the same time, data stored in MongoDB occupies 3727KB which is twice as many as SQL Azure. This is a direct consequence of the overhead that storing data in JSON format has.

As this amount of data is our prediction for a Serbia-based application, storage size issues are not much of a problem as there is more than 500MB left in both databases to keep them in the same cost range. Even if this application is intended for worldwide use, there is still much space for user base growth. Of greater significance is the speed of query execution. If a connection pool is created to keep connections open, MongoDB is the obvious choice as it is two times faster than SQL Azure. Even in the case of opening one connection per query, MongoDB has more stable execution times. Furthermore, as this is single-machine instance, we expect even greater improvements by deploying MongoDB instance in distributed environment thus allowing Map-Reduce algorithm to fully show its potential. Therefore, we will use MongoDB for a further development of our prototype.

In addition to databases that store driver and passenger data, we have used a NoSQL database to support SignalR server nodes. SignalR server-side code may be deployed alongside a NoSQL key-value memory cache data store named Redis [21]. With the minimum amount of 250MB of RAM allocated to Redis, a fully functioning server node could be produced. Such a node is capable of serving initially large enough number of simultaneous users on its own. Since a new server node can be cloned, and any cloned node's Redis object instance can then be easily subscribed to an existing Redis instance node, we can easily increase the number of new server nodes to meet our future scaling needs. Scale out is easily achieved in part due to SignalR's in-built scaling mechanisms, which uses Redis pub/sub features for a messaging backplane. Each SignalR server node could then be notified of any new WS or other real-time connection channels opened on any SignalR server node through its Redis instance. The load balancer of connected computing cloud instances, which is built into Windows Azure, takes care of diverting traffic to a most appropriate SignalR server node. Such node is chosen based on its current traffic, and it is able to process any incoming new or reoccurring real-time request. But since each node has by then been notified, by its Redis instance each connection should be replicable by any other SignalR node. Therefore each node is capable of replying to any previously opened real-time connection request on another SignalR node.

Table 1: Web browser support for Web Sockets.

Web browser	**Supported since version**	**Supported**
Internet Explorer	10.0 (fully)	Yes
Firefox	4.0 (partially) 6.0 (fully)	Yes
Chrome	4.0 (partially) 14.0 (fully)	Yes
Safari	5.0 (partially)	Yes
	6.0 (fully)	
Opera	11.0 (partially)	Yes
	12.1 (fully)	
iOS Safari	11.0 (partially)	Yes
	12.1 (fully)	
Opera Mini	-	No
Android Browser	-	No
BlackBerry Browser	7.0 (fully)	Yes
Opera Mobile	11.0 (partially)	Yes
	12.1 (fully)	
Crome for Android	25.0 (fully)	Yes
Firefox for Android	19.0 (fully)	Yes
Firefox OS Boot2Gecko	1.0.0-prerelease (fully)	Yes
Tizen OS	2.0.0a-emulator (fully)	Yes

4.3 User interface

Finally, for ubiquity reasons, our choice of prototype client's UI rendering was LeafletJS [22]. LeafletJS is an open-source library that provides HTML5 Canvas [23] mapping. Encouraged by the results of our previous project, TaxiProxy, we felt confident that HTML5/PhoneGap was a right choice. PhoneGap [24] allows a developer to develop a fully functional HTML5/CSS UI and then generate native mobile applications. Therefore, this client could be used on both desktop and mobile devices and have an ubiquitous UI. In both desktop and mobile web client, sharing the same JavaScript logic codebase offers a unified access to a geolocation [25] feature of the device clients are running on. That is a necessary feature for a dynamic carpooling and ride-sharing applications, but also although the look and feel across smaller resolutions changes accordingly, it is not drastically changed. The learning curve for using clients across multiple platforms is thereby reduced by utilizing a responsive CSS3 web design incorporated in Twitter's Bootstrap [26] library.

As we have previously mentioned in Subsection 4.1, real-time request and response messages used by clients to communicate with the SignalR server-side backend are in JSON format. LeafletJS utilizes this format for encoding a variety of geographic data structures named GeoJSON [27]. JSON and its derivatives tend to be lightweight, compared to XML, in an attempt to reduce the latency caused by the need to parse out data from server requests and responses. Additionally, reducing any network latency is also achieved by an attempt to support Web Socket transport, as it is data just sent at the TCP instead of HTTP level but still accessible by the browser.

However, support for WS as a mean of communication transport, depends primarily on a platform web browser's capabilities which is for current desktop and mobile web browsers given in Table 1.

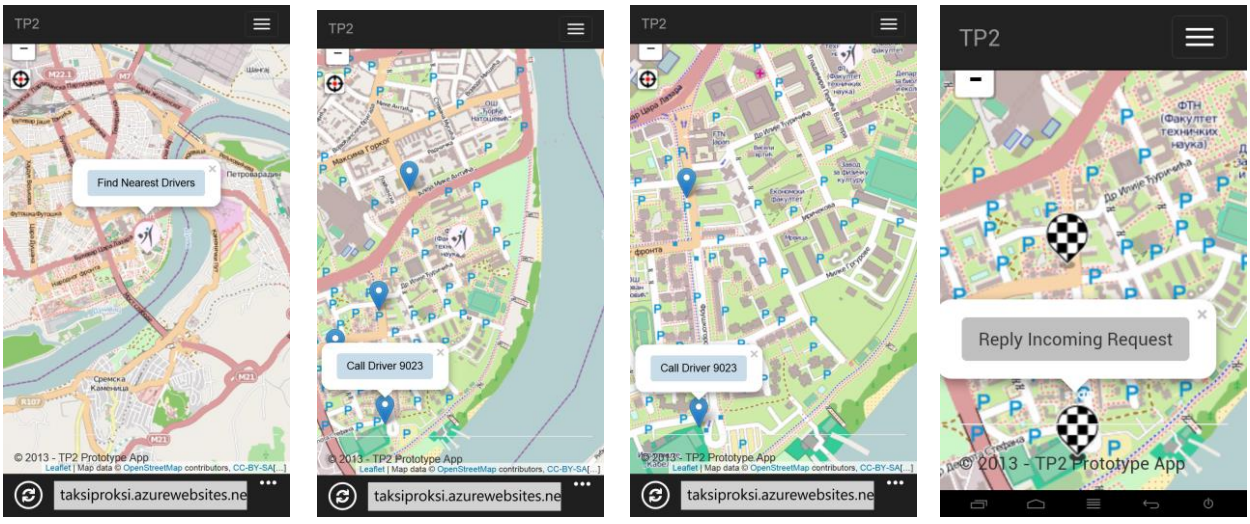


Figure 5: Prototype app screenshot on WP8 and Android platforms, respectfully showing passenger and driver UI

5 Future work and conclusion

Having described our prototype application, of which the early-development stage UI is depicted in Figure 5, our future work and plans envision for it to be deployed and further tested in the real-world. Since the prototype clients were based on previous work done for a commercial online taxi dispatcher, it will initially be tested and deployed as part of that solution to a limited number of taxi drivers. Early adopters of the online taxi dispatching service will then get the benefit of being able to track a few assigned taxis in real-time. The drivers of those taxis will be either issued mobile devices with pre-installed HTML5/PhoneGap web clients or those client apps will be installed on their own devices. Such real-world tests will hopefully lead to identifying problems not yet foreseen. Once a stable solution is reached, the prototype application could and will become a standalone service, open for public use and not just for taxi dispatching and the cost of its operational maintenance could then also be better estimated. If deemed low enough to be offset by ad support according to [28], its use could be free for end users unlike currently popular services like Lyft and SideCar [7, 8].

To reach that point however, some other issues, such as security and privacy, will also need to be tackled. In [29] the solution for the security and privacy issue was implied by use of a 3rd party location based service (LBS). This LBS used claims based authentication protocol OAuth to authenticate and subsequently authorize which exact set of users would be allowed access to the authorizing user's location. Unfortunately, from February 2013 the 3rd party LBS was shut down, and an alternative solution should either be found or developed a standalone service.

Trying to avoid the repeat of having to find alternatives to a 3rd party components not being operational any more, the focus in this paper was on starting to build up our own LBS features respective of privacy using NoSQL. Additionally, our goal in the future would be to also rely on information which can be provided from popular social networks for user authentication. To aid us in that endeavour, instrumental part of the puzzle could be Windows Azure built-in Access Control Service (ACS), allowing for users to

single sign-on to the proposed carpool and ride-sharing service just as if they were signing into the selected social networks. If those users comply, their location data could then only be made accessible to a subset of their social network friends, a widely acceptable solution from a current privacy standpoint.

This paper tried to underscore the need for developing dynamic real-time carpool and ride-sharing solutions, instead of already outdated static ones. Our approach comprises novel web technologies and approaches. Since a prototype has been successfully developed following the outlined design concepts, distribution and cloud strategies, it is obviously possible to build other such solutions using the same approaches. Especially interesting is the possibility to develop a web platform application that runs across multiple devices and their web browsers, be they mobile or desktop. Using an open-source Bootstrap library and Apache Cordova [30] mobile developer platform, which was derived from PhoneGap, is our main topic of interest. We feel this approach could be the unifying tool for any future service supposedly usable across multiple operating systems, current and future. Combining those with some other frameworks which use the HTML5 UI elements such as the canvas element thus adding the ability to render graphical data such as street level maps for carpool should, by our position, be the leading way forward.

References

[1] Dmiitrijevic, D., Nedic, N., & Dimitrieski, V. (2013, September). Real-time carpooling and ride-sharing: Position paper on design concepts, distribution and cloud computing strategies. In Computer Science and Information Systems (FedCSIS) 2013 Federated Conference on (pp. 781-786). IEEE.

[2] Ozanne, L., & Mollenkopf, D. (1999). "Understanding consumer intentions to carpool: a test of alternative models." In Proceedings of the 1999 annual meeting of the Australian & New Zealand Marketing Academy. smib.vuw.ac.nz (Vol. 8081).

[3] Fraichard, T. (2005). "Cybercar: l'alternative à la voiture particulière." Navigation (Paris), 53(1), 53-74.

[4] Dargay, J., & Hanly, M. (2007). "Volatility of car ownership, commuting mode and time in the UK." Transportation Research Part A: Policy and Practice, 41(10), 934-948.

[5] Massaro, Dominic W., et al. (2009) "CARPOOLNOW: Just-in-time carpooling without elaborate preplanning." the 5th International Conference on Web Information Systems and Technologies. Lisbon, Portugal. 2009.

[6] The largest car sharing network for cheap, green travel in Europe. Web – www.carpooling.com

[7] Lyft. Web – www.lyft.me

[8] SideCar. Web – www.side.cr

[9] Outsmarting traffic, together. Web – www.waze.com

[10] GeoRSS. Web – georss.org

[11] Sghaier, M., Zgaya, H., Hammadi, S., & Tahon, C. (2011). A Distributed Optimized Approach based on the Multi Agent Concept for the Implementation of a Real Time Carpooling Service with an Optimization Aspect on Siblings. International Journal of Engineering (IJE), 5(2), 217.

[12] Sghaier, M., Zgaya, H., Hammadi, S., & Tahon, C. (2010, September). A distributed dijkstra's algorithm for the implementation of a Real Time Carpooling Service with an optimized aspect on siblings. In Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on (pp. 795-800). IEEE.

[13] Garrett, J. J. (2005). Ajax: A new approach to web applications.

[14] Hickson, I. Server-Sent Events, W3C Working Draft 20 October 2011.

[15] Hickson, I. (2010). The Web Sockets API, W3C Working Draft 29 October 2009.

[16] Cattell, R. (2011). Scalable SQL and NoSQL data stores. ACM SIGMOD Record, 39(4), 12-27.

[17] Najbrži put do slobodnog vozila. Web – taxiproxy.com

[18] ASP.NET SignalR: Incredibly simple real-time web for .NET. Web – www.signalr.net

[19] MongoDB. Web – www.mongodb.org

[20] MongoLab. Web MongoDB. Web – www.mongolab.org

[21] Redis. Web – www.redis.io

[22] LeafletJS. Web – www.leafletjs.com

[23] HTML5 Canvas. Web – www.w3.org/TR/2009/WD-html5-20090825/the-canvas-element.html

[24] PhoneGap. Web – www.phonegap.com

[25] Popescu, A. (2010). Geolocation api specification. World Wide Web Consortium, Candidate Recommendation CR-geolocation-API-20100907.

[26] Bootstrap. Web – www.getbootstrap.com

[27] GeoJSON. Web – www.geojson.org

[28] Goldstein, D. G., McAfee, R. P., & Suri, S. (2013, May). The cost of annoying ads. In Proceedings of the 22nd international conference on World Wide Web (pp. 459-470). International World Wide Web Conferences Steering Committee.

[29] Dimitrijević, D., & Luković, I., & Dimitrieski, V., & Vasiljević, I. (2013) "Orchestrating Yahoo! FireEagle location based service for carpooling" 3rd International Conference on Information Society Technology and Management, Kopaonik, Serbia, 2013.

[30] Apache Codrova. Web – cordova.apache.org

Tiny Low-Power WSN Node for the Vehicle Detection

Michal Chovanec, Michal Hodon and Lukas Cechovic
University of Zilina
Univerzitna 8215/1, 010 26 Zilina, Slovakia
E-mail: michal.chovanec@fri.uniza.sk, michal.hodon@fri.uniza.sk, lukas.cechovic@fri.uniza.sk

Keywords: wsn, node, low-power, msp-430, TDMA, magnetometer, open-source, gcc

Received: November 20, 2013

In this paper, a tiny low-power wireless sensor network (WSN) node based on msp430 microcontroller is introduced. The node should be serving the needs of a special WSN used for traffic monitoring, or more precisely for the indication of vehicle's presence. The single vehicle detection is based on the utilization of magnetometer sensor, where the change of Earth's magnetic field, caused by the movement of passing vehicle, is measured. Communication subsystem of the node is based on the low-power sub-GHz transceiver cc1101 implemented together with the special, energy-efficient communication protocol in 868 MHz ISM band. In line with low-power architecture, environment friendly power consumption of the node was reached. It is equal on average to 300uA, when transmitting period is 1s and magnetometer data collection rate is equal to 20Hz.

Povzetek: Opisan je novi senzor oz. senzorsko omrežje za zaznavanje prometa.

1 Introduction

Traffic monitoring becomes very important with the growing number of registered motor vehicles. Based on [1], there are more than 250 million vehicles on the European roads and this number tends to increase. The higher this number is, the higher is the probability of accidents on the roads. Therefore, in modern society, both traffic monitoring systems that monitor different road networks and intersections, as well as other intelligent transportation systems (ITS) are implemented. These systems substitute the role of humans.

Conventional methods utilized for the vehicle's detection are based on the use of video cameras or magnetic loops in roadway [2]. In [2], where these methods are described in detail, it is also pointed that disadvantages of these methods are not inconsiderable. Beside the fact that the camera systems are very expensive and the magnetic loops require roadway cut, both methods require external power supply because of their high power consumption.

In [3], other disadvantages of such systems are depicted, from the ineffectiveness of magnetic-loops or camera-based solutions, up to uncertainties brought along with the new technologies, such as magnetometers or microradars. Moreover, if such systems are utilized for traffic data collecting on local roads, the costs connected with their installation are quite significant though dependant on the type of installation, either portable which offers short-duration counts (up to 48 hrs) or permanent which could work up to one year [4]. In this case, the cheapest solutions are the portable ones, though their acquisition costs are still high. According to [5], if the sensors are utilized for the traffic volume, vehicle's length and average speed classifications, video- and radar-based solutions provide the lowest average percentage errors. However, as it was already mentioned in [2], these solutions are quite energy demanding and thus not suitable for long-time measurements. It could be assumed that the special low-power variations of these technologies could bring the desired performance.

Therefore, with respect to [6], other technologies come to the forefront allowing utilisation of such systems even at places without electricity. One of this low-power methods for correct vehicle detection is based on the utilization the Earth magnetic field disturbance [7]. As it was outlined in [7], a tiny three-axis magnetometer can be used for this purpose. With regards to [5], if only traffic volume is measured, magnetometric error-rate is really similar to the classification errors of other technologies. And since the magnetometric measures are really power-friendly, application of this technology for the purposes of the traffic volume monitoring, even at places without electricity, could be feasible. This assumption was successfully verified in [8] as well, where a kind of magnetometer-based vehicle classification method was introduced.

Therefore, upon the findings mentioned, as well as with considerations of [9], the prototype of a special low-power sensor node was developed embedding magnetometer as the main sensing tool for the vehicle's presence monitoring. The node is a part of a special wireless sensor network (WSN), which was partly described in [10]. The WSN is based on the multihop principle, using multiple places for sensing and broadcasting information about vehicle's presence to the server station. This approach was chosen because of the high power efficiency as well as low computational requirements of the sensor nodes. In order to minimize the power consumption and maximize operat-

ing range of the network during communication, sub-GHz 868MHz ISM band was used as the communication frequency of the WSN protocol [11].

2 Node hardware

WSN network consists of several nodes communicating on 868MHz free ISM band. As a communication unit, a low-power sub-1 GHz RF transceiver TI CC1101 unit was used. Power consumption of the receiver and the transmitter in operating mode is ca. 16 - 20mA, however when all clocks are stopped and the module is in the sleep mode, the power consumption is 0.7uA.

As a main control unit, we used ultra-low power 16-bit microcontroller (MCU) TI MSP430F2232 (16MHz, 8kB Flash, 512B RAM) with power consumption about 270uA@1 MHz in active mode, about 30uA when just ACLK is active (1024Hz), and 0.1uA in LPM4 mode where all clocks are stopped [12].

For the purposes of magnetic field detection, three-axis digital low power magnetometer (MAG) MAG3110 from Freescale was used. It is measuring the three components of the local magnetic field. All the mentioned modules comprise the main parts of the developed sensor node. A summary of the power consumption of all the node parts is provided in the table below.

Device	Active Mode [uA]	Low-power Mode [uA]
MSP430f2232	270	30
CC1101	16400-tx 18400-rx	0.7
MAG3110	137.5 @10Hz rate	2

Table 1: Node parts - power consumption

The node has been designed so that the board space and power consumption were as small as possible (Figure 1). Keeping the node compactness, the device can be powered from a single cell Li-Pol battery, which has @ 1000mAh capacity and also relatively small dimensions.

Considering the node design, MAG module was placed as far as possible from the transceiver so that the noise brought along with the magnetic field sensing could be reduced. MAG communicates with MCU via I2C bus @ 100kHz clock rate. It can trigger the data-ready interrupt, which can be used for the MCU waking up. For correct working MAG needs only three ceramic external capacitors.

For correct functioning, the transceiver needs an output LC filter as well as a balun circuit for unbalanced Whip antenna. Additional band stop filter may be added to build an optional filter to reduce emission at 699MHz. For precise frequency synthesis the integration of 27MHz crystal is necessary.

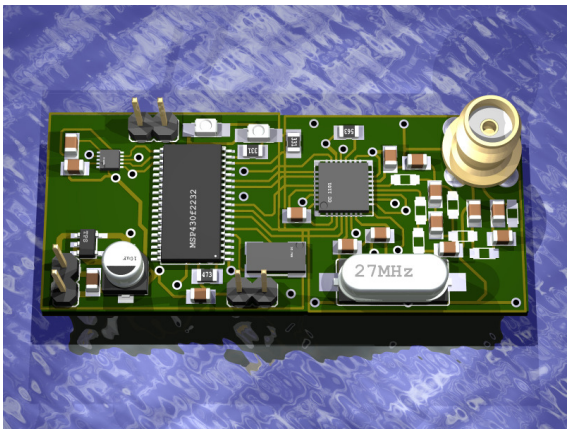


Figure 1: 3D model of the node

MCU was in this application clocked on 1MHz. For the debugging purposes, two led diodes and the UART interface were hooked up.

The hardware of the node is shown in the figure below.

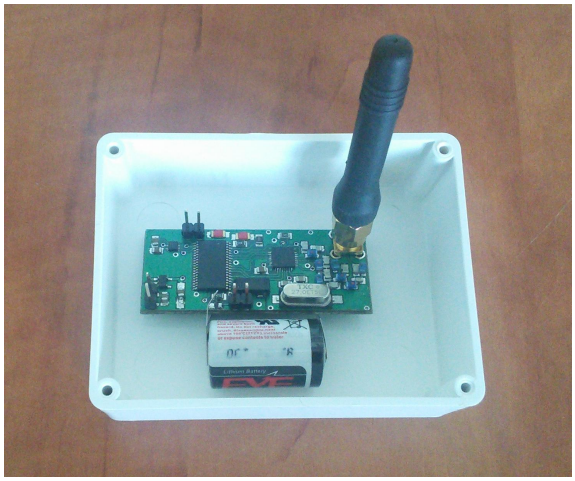


Figure 2: HW of the developed node - communication testing

3 Firmware

The control software of the node manages the magnetometer data reading as well as single packet transmitting/receiving tasks. When we write this as an infinite while(1) { } loop, the power consumption will be noticeably high, since everything will be done in power-on mode. By this approach we also can not ensure the real-time processing of asynchronous incoming packets. The solution is to write the software as an event-driven, when the crucial tasks of the program are determined by the events. Nowadays microcontrollers have refined an event-driven support within the interrupts. Therefore, the firmware of the node

was based on the implementation of the timer interrupts assuring execution of a certain event at a given frequency in the defined time intervals. This approach allows the nodes to fall into the sleep mode and hence save the energy considerably.

The firmware was then implemented as follows: after MCU's GPIO initialisation, *timer_0* was set to 1024Hz periodic interrupt. The frequency was divided from an external 32.768kHz crystal which was configured as ACLK clock. When the timer triggers, MCU reads magnetometer control registers and increments the software counter. When magnetometer control register sets the data ready, the data are read through I2C bus into the digital filter.

The software counter counts from 0 to 1023 with the 1s time interval of one loop. There are few important time moments within the loop:

1023 - MCU wakes up the radio to receive;

0 - MCU waits for an incoming packet;

1 - power down radio;

16 - MCU wakes up the radio and trasmits the packet;

other - in the other time slots MCU sleeps;

The end of packet reception or transmission is reported via GPIO interrupt. MCU then does not need to poll and synchronously read CC1101 registers. Thus, the sleep time of nodes can be maximized. Moreover, due to the preservation of enough space for the received packet processing in the time slot 0, the packet transmitting was defined to be done each time in the 16th time slot. Thanks to this, every node knows when to ask for a data.

3.1 The synchronization of the nodes

Since the system is functioning on an event-driven principle utilizing the triggering from the timer interrupt, TDMA method (Time Division Multiple Access) was used as the channel access method of WSN. As it was already mentioned, one program time loop takes 1 second and it is divided into 1024 time slots. Every time slot has then unambiguously defined specific task which should be done according to the implemented crystal frequency. In order to ensure the proper functioning of every node within the network, all nodes have to be synchronized so that they transmit/receive the data only when the neighbouring nodes are awaken. Therefore, in the network, there is always one master node, which is always synchronized with itself, and which is periodically sending packet to its antenna, so that other nodes could synchronize according to it. The address of the master node was set to 0x01.

All other "slave" network nodes synchronize according to the previous node synchronization packet with the address *node_adr-1*. This means that the master node synchronizes the most close node, which synchronizes the following one, etc. When the synchronization packet is received, after the packet receiving is done, software counter of each node sets the time slot to 1. Next, in the time slot 16, it transmits the node synchronization packet to the next nodes. This approach allows the implementation of the master/slave tree network architecture. In our application a "string" network architecture was however used. This architecture implies placement of all nodes in one logical subnet, where the current node synchronizes only next node. Within this structure, the data from master can be shifted to the last node in the string. Network depth is in this case limited by clock precision, and receiving time-window width.

3.2 Data filtering

Whilst the sensors are planned to be deployed in a rush, noisy environment of traffic roads, the magnetometer data output needs to be filtered. In the vehicle-detecting application, only fast magnetic field changes just within X and Y axis have to be monitored. To prevent slow magnetic field fluctuations, sensor moving, or parameters floating in time, the long term magnetic field process has to be recorded. This can be done by using a low pass filter with a very long response time, e.g. 0.5 .. 2 seconds.

For Z coordinate, a simple filter can be a written as

$$Y(z) = \frac{\alpha(1 - z^{-1})X(z)}{1 - \alpha z^{-1}} \tag{1}$$

where α is time constant in range (0, 1), X(z) is a filter input and Y(z) filter output. For field change, derivation (difference in discrete domain) can be used. After rewriting into time domain we have

$$y(t) = \alpha y(t) + \alpha(x(t) - x(t-1)) \tag{2}$$

We can see that slow changing signal will be zeroed, but fast signal change will produce high output amplitude for signal change duration, or low pass filter response time. The final output from N axis will be produced as:

$$m(t) = \sum_{i=0}^{N} |y(t)_i| \tag{3}$$

where y(t) is a vector result from all three filtered axis of magnetometer. Vehicle presence is in this case defined as the threshold m(t) with a constant value.

From the equations above it is obvious that the utilized filter is really simple, almost trivial. This approach was chosen because of the minimisation of the computational power of a node that is necessary for the signal preprocessing. Correctness of the filter was investigated in a set of measurements, where the straight movement of a common vehicle (in our case Skoda Octavia 1.9TDi) was monitored (Figure 3).

The sensor was placed with the X-axis set in parallel according to the vehicle's movement, so the highest signal-change should be observed within the Y axis of the sensor, which was placed orthogonally. The vehicle then performed three two-way rides forward and backward (six in

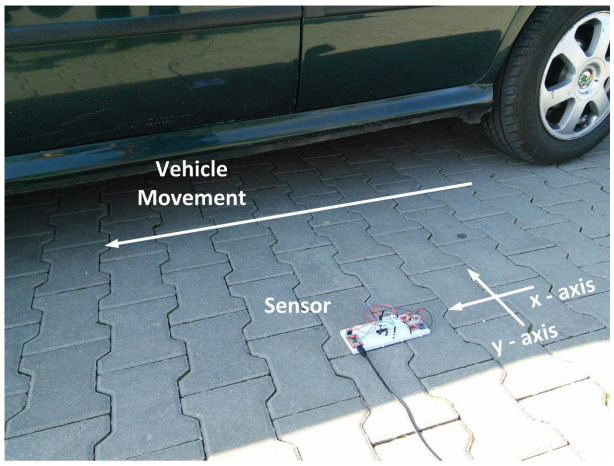


Figure 3: Evaluation of the filter functioning

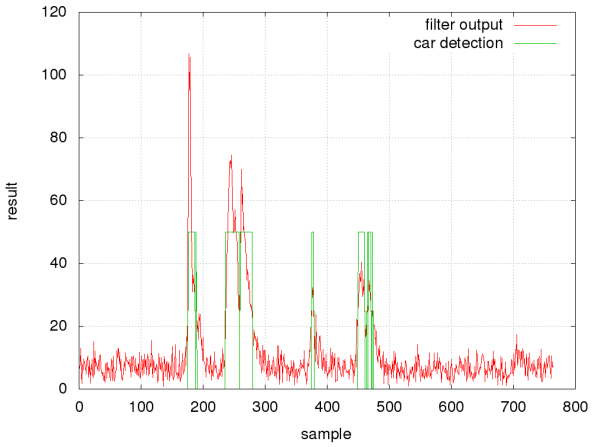


Figure 5: Filtered result

total), at constant speed ca. 20km/h along with the sensor. The drives were distanced stepwise 0.5m, 1m, 3.5m from the sensor.

In the figures below, the output filter results of magnetometer car detection with $\alpha = 0.9$ and 20Hz sampling frequency are depicted. The superposed DC value on all three axis - Earth magnetic field - can be observed too.

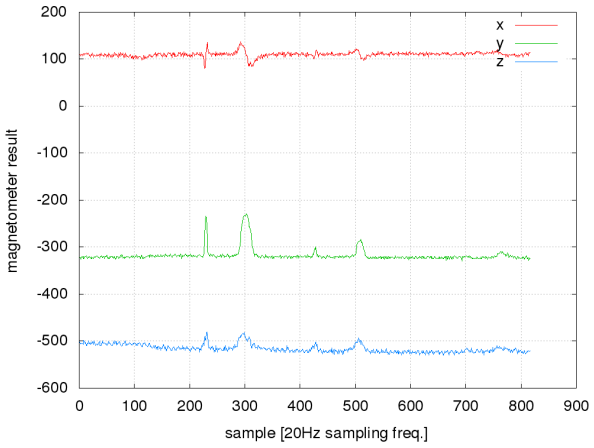


Figure 4: Magnetometer output

3.3 System architecture

In the final implementation, there are three node types [10]:

1 Sensor node,

2 Retransmitting node,

3 Signal node.

All three node types have the same hardware, they differ only by the firmware implemented within the main control unit.

The sensor node is used for the single vehicle detection, so its proper placement is the most critical task. As shown in test measurements @20km/h vehicle speed, magnetometer detection works well at the distance 1m. Therefore, it could be assumed that by its installation on the edge of the road, full coverage of one road traffic line could be assured.

Retransmitting nodes work just like repeaters since they only re-send the received message to the next node. As it was shown in [10], nodes are synchronized and wake up only for a short time. When a node receives a message from another neighbour, it just modifies the packet destination address by its incrementing by one and resends the packet further. Because the linear network structure is applied, each node abut on two neighbours.

The signal node is utilized for the signalling purposes to inform/warn other traffic participants about the presence of the vehicle. It can be implemented either as the LED road sign [10], or as the server/data-logger when sending the data into computer via UART interface. In this case, integration of UART to USB converter is mandatory.

4 Conclusion

The tiny node for a vehicle presence detection, described in this study, represents a good solution for the data collection at the places without mains power. The main advantages of this solution are: low power consumption, tiny dimensions as well as a simple communication protocol. Small modification of the node, reached by the integration of other different sensing units, such as accelerometer or sensors for measuring temperature, humidity, pressure, etc., within the node's layout, allows the mutuation of its functioning according to the direct needs of the WSN. Therefore, apart from the transportation domain, it could be applied also within different other application domains, such as rails vibrations monitoring, building temperature/humidity sensing, shopping center's traffic monitoring, etc.

However, the developed prototype has also some defi-

ciencies. As far as the node layout is concerned, the node needs two crystals and many LC components for proper functioning. Utilization of an external antenna is also not an ideal solution, especially from the final price of the product viewpoint. Integrated PCB or ceramic antenna will be in this case better. Also, even though the software is written so as to minimize the time delay, its hardware dependency is still too high. Therefore, in the next releases the firmware will be rewritten more modularly.

For the massive production, the best solution is to implement the node control unit within ASIC or FPGA integrated circuits, whilst the sensor signal processing can be implemented more effectively by using logic gates and parallel processing [13]. The ratio between computing power and power consumption will increase then. Utilization of alternative power sources, as these described in[14] or [15] is in this case worth considering.

5 Acknowledgement

This publication was supported within the project *Centre of excellence for systems and services of intelligent transport II.* ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

"Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

References

[1] ANFAC (Spanish Automobile Association) *Latest Report on motor vehicles in use in Europe*, 2010 Edition.

[2] Ashish Dhar (2008) *Traffic and Road Condition Monitoring System*, M.Tech.Stage 1 Report, Roll No: 07305041.

[3] Greg Davies *Scottsdale Transportation Commission Report - Bicycle Signal Detection*, Transportation Commission, July 18, 2013.

[4] Transportation Research Synthesis *Collecting and Managing Traffic Data on Local Roads*, TRS 1207, Minesota Department of Transportation, September 2012.

[5] Michael Marti, Renae Kuehl, Scott Petersen *Traffic Data Collection Improvements*, Research Project Final Report 2014RIC51B, Minesota Department of Transportation, February 2014.

[6] Lawrence A. Klein, Milton K. Mills, David R.P. Gibson *Traffic Detector Handbook*, Third Edition—Volume I, Publication No. FHWA-HRT-06-108, October 2006.

[7] Juraj Micek, Michal Hodon (2013) *Wireless Sensor Networks for Intelligent Transportation Systems*, IEEE CommSoft newsletter, No.2, 2013.

[8] Ondrej Karpis *Sensor for vehicles classification*, FedCSIS: proceedings of the Federated conference on computer science and information systems : September 9-12, 2012, Wrocław, Poland, IEEE, ISBN 978-83-60810-51-4, pp. 785-789.

[9] Dan Middleton, Hassan Charara, Ryan Longmire *Alternative Vehicle Detection Technologies for Traffic Signal Systems: Technical Rreport*, FHWA/TX-09/0-5845-1, Texas Transportation Institute, February 2009.

[10] Michal Hodon, Michal Chovanec, Martin Hyben *Intelligent traffic-safety mirror*,In: Studia informatica universalis, ISSN 1621-7545, Vol. 11, no. 1, 2013, pp. 87 -101.

[11] Mitko Tanevski, Alexis Boegli and Pierre-Andre Farine (2012) *Firmware for Ensuring Realtime Radio Regulations Compliance in WSN*, 7th IEEE International Workshop on Practical Issues in Building Sensor Network Applications - SenseApp 2012, Clearwater, Florida, page(s): 917 - 920, print ISBN: 978-1-4673-2130-3.

[12] Texas Instruments *MSP430 Ultra-Low-Power Microcontrollers*, http://www.ti.com /lit/sg/slab034w/slab034w.pdf.

[13] Martin Hyben, Michal Hodon *Low-cost command-recognition device*, Teleinformatics Review (in press).

[14] Michal Kochlan, Peter Sevcik *Supercapacitor power unit for an event-driven wireless sensor node*, FedCSIS: proceedings of the Federated conference on computer science and information systems : September 9-12, 2012, Wrocław, Poland, IEEE, ISBN 978-83-60810-51-4, pp. 791-796.

[15] Peter Svcik, Oldrich Kovar *Alternative energy sources for WSN node power supply*, ITS 2013 = Intelligent transportation systems 2013, August 26-30, 2013,- ISSN 1339-4118, Žilina, ISBN 978-80-554-0763-0, CD-ROM, pp. 146-149.

Genetic Algorithm with Fast Greedy Heuristic for Clustering and Location Problems

Lev A. Kazakovtsev and Alexander N. Antamoshkin
Siberian State Aerospace University named after Academician M. F. Reshetnev
prosp. Krasnoyarskiy rabochiy, 31, Krasnoyarsk 660014, Russian Federation
E-mail: levk@bk.ru

Keywords: continuous location problem, p-median problem, genetic algorithms, discrete optimization, clustering, k-means

Received: April 4, 2014

Authors propose new genetic algorithm for solving the planar p-median location problem and k-means clustering problem. The ideas of the algorithm are based on the genetic algorithm with greedy heuristic for the p-median problem on networks and information bottleneck (IB) clustering algorithms. The proposed algorithm uses the standard k-means procedure or any other similar algorithm for local search. The efficiency of the proposed algorithm in comparison with known algorithms was proved by experiments on large-scale location and clustering problems.

Povzetek: Razvit je nov algoritem za gručenje in lokalizacijo s hitro požrešno hevristiko.

1 Introduction

The aim of a continuous location problem [18] is to determine the location of one or more new facilities in a continuum of possible locations. Main parameters of such problems are the coordinates of the facilities and distances between them [54, 19, 22]. Examples of the location problems include the location of warehouses [22], computer and communication networks, base stations of wireless networks [44, 30], statistical estimation problems [41], signal and image processing and other engineering applications. In addition, many problems of cluster analysys [21, 34, 47] can be considered as location problems [37, 32] with squared Euclidean [21, 26], Euclidean [37, 48] or other distance functions [23].

The Weber problem [52, 54] is the problem of searching for such a point that a sum of weighted Euclidean distances from this point to the given points (existing facilities which are also called "demand points" or "data vectors" in case of a clustering problem) is minimal:

$$\arg\min_{X\in\mathbb{R}^2} F(X) = \sum_{i=1}^{N} w_i L(X, A_i). \quad (1)$$

Here, $L(\cdot)$ is a distance function (norm), Euclidean in case of Weber Problem.

For solving this problem (serching for its center), we can use an iterative Weiszfeld procedure [53] or its improved modifications [51, 17]. Analogous problems with Manhattan and Chebyshev distances are well investigated [55, 50, 11]. Convergence of this algorithm is proved for various distance metrics [40].

One of possible generalizations [22, 14] of the Weber problem is the p-median problem [22] where the aim is to find optimal locations of p new points (facilities):

$$\arg\min F(X_1, ..., X_p) = \sum_{i=1}^{N} w_i \min_{j\in\{\overline{1,p}\}} L(X_j, A_i). \quad (2)$$

Here, $\{A_i | i = \overline{1,N}\}$ is a set of the demand points (data vectors), $\{X_j | j = \overline{1,p}\}$ is a set of new placed facilities, w_i is a weight coefficient of the ith demand point, $L(\cdot)$ is a distance function defined on a continuous or discrete set [24]. In this paper, we consider continuous problems in an n-dimensional space. In the simplest case, $L(\cdot)$ is Euclidean norm. In this case, the Weiszfeld procedure is implemented up to p times at aech iteration of the the iterative alternating location-allocation (ALA) method [13].

If the distance function (metric norm) is squared Euclidean (l_2^2) then the solution of the single-facility problem (1) is the mean point (centroid) [22]:

$$x_j = \sum_{i=1}^{N} w_i a_i / \sum_{i=1}^{N} w_i. \quad (3)$$

Here, we assume that $X = (x_1, ..., x_d)$, $A_i = (a_{i,1}, ..., a_{i,d})$ $\forall i = \overline{1,N}$.

The simplest and probably most popular clustering [9, 49] model is k-means [33, 34] which can be formulated as a p-median problem (2) where $w_i = 1$ $\forall i = \overline{1,N}$ and $L(\cdot)$ is squared Euclidean norm l_2:

$$L(X, Y) = \sum_{i=1}^{d} (x_1 - y_1)^2$$

where $X = (x_1, ..., x_d) \in \mathbb{R}^d$, $Y = (y_1, ..., y_d) \in \mathbb{R}^d$. Searching for the centroid takes less computational resourses than searching iteratively for the center in case of

the Weber problem and the ALA method works faster in this case.

The p-median problem with Euclidean (l_2), squared Euclidean (l_2^2) or other l_p distances [16] is a problem of global optimization: the objective function is not concave nor convex [13]. The ALA method and analogous algorithms can find one local minimum of the objective function, their result depends on the initial solution. Moreover, such global optimization problems are proved to be NP-hard [20, 2, 35] for both continuous and discrete location [45, 46, 25] which makes usage of a brute force methods impossible for large datasets. Therefore, many heuristics [39] are used to improve the obtained results. One of widely used heuristics for initial seeding is k-means++ [8]. Most popular ALA procedures for the k-means problem are based on an algorithm proposed by Lloyd [33]. The algorithm known as standard k-means procedure [34] is fast local search procedure based on Lloyd's procedure. However, many authors proposed faster methods based on this standard procedure [56, 12, 1] for datasets and continuous supplemented data streams.

Many authors propose approaches based on data sampling [38]: solving reduced problem with randomly selected part of the initial dataset and using this result as the initial solution of the ALA algorithm on the whole dataset [15, 29, 43]. Analogous approach was proposed for discrete p-median problems [6].

Many authors propose various genetic algorithms (GA) for improving the results of the local search [28, 36, 31, 42]. Most of such algorithms use evolutionary approach for recombination of the initial solution of the ALA algorithm.

Hosage and Goodchild [27] proposed the first genetic algorithm for the p-median problem. Genetic algurithms are based on the idea of recombination of elements of a set ("population") of candidate solutions called "individuals" coded by special alphabet. In [10], authors propose a genetic algorithm providing rather precise results but its convergence is slow. Alp et al. [3] proposed a quick and precise genetic algorithm with special "greedy" heuristic for solving discrete p-median problems on networks which was improved by Antamoshkin and Kazakovtsev [4]. This algorithm can be used for generating the initial solutions for the ALA algorithm [42]. The idea of the "greedy heuristic" is as follows. After selecting two "parent" solutions, new infeasible solution (a candidate solution) is composed as the union of the facility sets of the "parent" solutions. From new solution, the facilities are eliminated until the solution becomes feasible. At each step, algorithm eliminates such facility that its elimination gives minimal addition to the objective function. If this algorithm is used for the continuous p-median problem, it generates the initial solution for the ALA algorithm [42] which must be implemented at each step to estimate the result of eliminating of each facility from the candidate solution.

In this paper, we present a new genetic algorithm with floating point alphabet based on the ideas of algorithm proposed by Alp et al. [3]. Original Alp's algorithm uses integer alphabet (number of vertices of the network) in "chromosomes" (interim solutions) of the GA. Its version for planar location problems [42] uses integer alphabet for coding numbers of data vectors used as initial solutions of the ALA algorithm. In our algorithm, we use floating point alphabet. Elements of "chromosomes" of our genetic algorithm are coordinates of centers or centroids of the interim solution which are altered by steps of the ALA algorthm and eliminated until a feasible solution is obtained. Such combination of the greedy heuristic and ALA procedure allows the algorithm to get more precise results.

In case of continuous locating problems, the greedy heuristic is a computationally intensive procedure. We propose new procedure which allows eliminating sets of the centers or centroids from the candidate solution which gives multiple reduce of the running time.

2 Known algorithms

The basic idea of the alternating location-allocation ALA is recalculating the centers or centroids of the clusters and reallocating of the data vectors to the closest center or centroid:

Algorithm 1. *ALA method [28].*

Require: Set $V = (A_1, ..., A_N)$ of N data vectors in d-dimensional space, $A_1 = (a_{1,1}, ..., a_{1,d}), ..., A_N = (a_{N,1}, ..., a_{N,d})$, initial solution: a set of centers or centroid of p clusters $X_1 = (x_{1,1}, ..., x_{1,d}), ..., X_p = (x_{p,1}, ..., x_{p,d})$.

1: For each data vector, find the closest centroid:

$$C_i = \arg \min_{j=\overline{1,p}} L(A_i, X_j) \; \forall i = \overline{1, N}.$$

2: For each cluster $\mathcal{C}_j^{clust} = \{i \in \{\overline{1, N}\} | C_i = j\}$, recalculate its center or centroid X_j. In the case of Euclidean (l_2) metric, Weiszfeld procedure or its advanced modification can be used. In the case of squared Euclidean (l_2^2) metric, equation (3) is used to obtain each of d coordinates.

3: If any center or centroid was altered at Step 2 then go to Step 1.

4: Otherwise, STOP. $X_1, ..., X_p$ are local minima.

To improve the performance of Algorithm 1, recalculation of the centers or centroids are performed for the altered clusters only. In the case of Euclidean metric l_2, this allows to avoid running Weiszfeld procedure for each of the clusters at each iteration.

In case of the squared Euclidean metric l_2^2, this algorithm is called Standard k-means procedure.

The ALA methods is a local search procedure, its result depends on proper selection of the initial solution. In the simplest case, p data vectors can be randomly selected as the initial centers or centroids. A popular procedure called k-means++ for initial seeding [8] guarantees $O(\log(p))$ accuracy by proper choosing initial centers. The idea of this method is based on probability change of choosing data

vectors as the initial centers depending on distances to the closest previously chosen vectors. Analogous method for discrete location problems was proposed in [4]. The k-means++ algorithm is as follows:

Algorithm 2. *k-means++ [8].*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters.

1: Initialize the probabilitiy distributions vector $P = (p_1, ..., p_N)$ with equal values (e.g. 1). Initialize the set of centroids $\chi = \emptyset$.

2: Chose one data vector X from the set of data vectors V at random using weighted probability distributions P: calculate $S = \sum_{i=1}^{N} p_i$, generate a random value $r \in [0; S)$ with the uniform distribution and use $i_{min} = \arg\min_{i \in \{\overline{1,N}\}: \sum_{j=1}^{i} p_i < r} i$. Set $\chi = \chi \cup \{A_{i_{min}}\}$.

3: For each $i \in \{\overline{1, N}\}$ set $p_i = \min_{X \in \chi} L(X, A_i)$. Here, $L(\cdot)$ is the distance metric.

4: If $|\chi| < p$ then go to Step 2.

5: Otherwise, STOP. χ is the initial set of centers or centroids.

The idea of sampling k-means [38, 15] is very simple:

Algorithm 3. *Sampling k-means.*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters, parameter $s \in (0; 1)$.

1: Choose randomly s data vectors from V and form new set V_s.

2: Initialize the set of initial centers or centroids χ. To improve the results, k-means++ procedure (Algorithm algkplus) can be performed.

3: Run Algorithm 1 with the initial set χ and set of data vectors V_s. After this procedure, we have the modified set χ.

4: Run Algorithm 1 with the initial set χ for the whole set of data vectors V.

5: STOP.

In [15], authors propose a method of choosing an optimal value of the parameter s.

Sampling k-means approach, k-means++ initial seeding and other techniques improve the results of the k-means procedure, however, they do not eliminate its most important flaw: all of them perform local search. The simplest approach used for global optimization is random multistart [5]. In this case, the local search procedure runs with various randomly generated initial data. For the p-median or k-means problem, this algorithm is as follows.

Algorithm 4. *Random multistart.*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters.

1: Set $F^** = +\infty$.

2: Initialize the sets of data vectors indexes $\chi : \chi \subset \{\overline{1, N}\}, |\chi_k| = p$. Uniform random generation or k-means++ procedure can be used.

3: Perform the ALA procedure with the initial solution χ and obtain a local minimum F^* of the objective function (2) and a set of corresponding centers or centroids χ^*. Instead of the "pure" ALA procedure, sampling k-means can be used in case of a large dataset.

4: If $F^{**} > F^*$ then set $F^{**} = F^*; \chi^** = \chi^*$.

5: Check the stop conditions. If they are not reached then go to Step 2.

6: Otherwise, STOP. The solution is χ^{**}.

The scheme of the genetic algorithm with greedy heuristic proposed by Alp et al. for continuous location problems is as follows [3, 42].

Algorithm 5. *GA with greedy heuristic.*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters, popultion size N_p.

1: Initialize N_p sets of data vectors indexes $\chi_1, ...\chi_{N_p} : \chi_i \subset \{\overline{1, N}\}, |\chi_k| = p \; \forall k = \overline{1, N_p}$. For each $k \in \{\overline{1, N_p}\}$, calculate the fitness function. In case of the continuous p-median problem, to obtain the fitness function value for χ_k, algorithm performs the ALA procedure with initial solution χ_k and calculate

$$\mathcal{F}_k = F(\chi_k^*) = \sum_{i=1}^{N} w_i \min_{X \in \chi_k^*} L(X, A_i). \qquad (4)$$

Here, χ_k^* is the result of running the ALA procedure with the initial solution χ_k.

2: If the stop conditions are reached then go to Step 7.

3: Choose randomly two "parent" sets χ_{k_1} and χ_{k_2}, $k_1, k_2 \in \{\overline{1}, N_p\}, k_1 \neq k_2$. Running special crossover procedure with greedy heuristic, obtain "child" set of indexes χ_c. Calculate the fitness function value $\mathcal{F}_c$ in accordance with (4).

4: If $\exists k \in \{\overline{1, N_p}\} : \chi_k = \chi_c$ then go to Step 2.

5: Choose index $k_{worst} = \arg\max_{k=\overline{1,N_p}} \mathcal{F}_k$. If $\mathcal{F}_{wotst} < \mathcal{F}_c$ then go to Step 2.

6: Replace $\chi_{k_{worst}}$ with χ_c, set $\mathcal{F}_{k_{worst}} = \mathcal{F}_c$ and go to Step 2.

7: STOP. The result is a set χ_k^*, $k^* = \arg\min_{k=\overline{1,N_p}} \mathcal{F}_k$.

In the above version of this algorithm, at Steps 5 and 6, the worst solution χ_{worst} is replaced by new solution. In our experiments, we used other procedure at Step 5 (simplest tournament selection): choose randomly two indexes k_1 and k_2, $k_1 \neq k_2$; set $k_{worst} = \arg\max_{k \in \{k_1, k_2\}} \mathcal{F}_k$. This version of Step 5 gives better results.

In both random multistart and genetic algorithms, various stop conditions can be used. We used maximum running time limit.

Unlike most genetic algorithms, this method does not use any mutation procedure. However, the crossover procedure uses a special heuristic:

Algorithm 6. *Greedy crossover heuristic.*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters, two "parent" sets of centers or centroids χ_{k_1} and χ_{k_2}.

1: Set $\chi_c = \chi_{k_1} \cup \chi_{k_2}$. Note that $p \le |\chi_c| \le 2p$, i.e., the candidate solution χ_c is not feasible.
2: If $|\chi_c| = p$ then STOP and return the solution χ_c.
3: Calculate $j^* = \arg\min_{j \in \chi_c} F(\chi_c \setminus \{j\})$.
4: Set $\chi_c = \chi_c \setminus \{j^*\}$.
5: Go to Step 2.

At each iteration, one index of the center or centroid is eliminated (Step 4). At Step 3, Algorithm 6 chooses the index of a center or centroid which can be eliminated with minimum change of the fitness function. To estimate the fitness function, ALA procedure must be performed. Thus, Step 3 of Algorithm 6 is computationally intensive. In case of Euclidean metric, iteratice Weiszfeld procedure must run at each iteration of the iterative ALA procedure performed $|\chi_c|$ times.

Therefore, Algorithm 6 is a computationally intensive procedure, slow for very large datasets in case of k-means problem and almost inapplicable in case of large-scale continuous p-median problems with Euclidean metric. Idea of this heuristics correlates to ideas of the information bottleneck (IB) clustering method [48]. In the IB algorithms, at the start, all data vectors form individual clusters. At each step, one cluster is eliminated, its members join other clusters. To choose such cluster, for each of them, algorithm calculates the "information loss". In the case of "geometric" clustering based on distance metrics, this loss can be estimated as the distance function increase. The computational load in case of the IB clustering allows to implement this method to small datasets only ($N < 1000$). This form of the method proposed by Alp et al. for continuous location problems is a compromise between the IB clustering simpler heuristics like traditional genetic algorithms or random multistart.

This algorithm can be used for solving a discrete p-median problem (2) with an additional condition:

$$X_j \in V \; \forall j \in \{\overline{1,p}\} \tag{5}$$

which can be used as an initial solution of the ALA method.

Algorithm 7. *Greedy crossover heuristic for initial seeding.*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters, two "parent" sets of centers or centroids χ_{k_1} and χ_{k_2}.
1: Set $\chi_c = \chi_{k_1} \cup \chi_{k_2}$.
2: If $|\chi_c| = p$ then STOP and return the solution χ_c.
3: Calculate $j^* = \arg\min_{k \in \chi_c} \sum_{i=1}^{N} (\min_{j \in (\chi_c \setminus \{k\})} w_i L(A_i, A_j))$.
4: Set $\chi_c = \chi_c \setminus \{j^*\}$.
5: Go to Step 2.

In this case, the ALA method always starts from a local minimum of the discrete problem (2) with an additional constraint (5). This version of the algorithm is much faster, it gives better results than the random multistart (Algorithm 4) for most popular test datasets (see Section 4). However, such results can be improved.

We propose two modifications. One of them decreases the computational intensiveness of the algorithm, the second one improves its accuracy. Their combination makes new algorithm faster and more precise in case of large datasets.

3 Our contribution

Let us consider Steps 3 and 4 of Algorithms 6 and 7. At each iteration, Step 3 selects one index of data vectors and eliminates it from the candidate solution. Let us assume that at some kth iteration, j^*th index is eliminated and at $(k+1)$th iteration, algorithm eliminates j^{**}th index. Our first modification is based on the supposition that if A_{j^*} is distant from $A_{j^{**}}$ (i. e. $L(A_{j^*}, A_{j^{**}}) > L_{min}$, L_{min} is some constant) then the fact of eliminating or keeping j^{**}th index "almost" does not depend on the fact of elimination or keeping of j^*th index at previous iteration.

If the facts of choosing of indexes of two distant data vectors at Step 3 in two successive iterations are independent then the decisions on their eliminating (or keeping) can be made simultaneously. We propose the following modification of Steps 3 and 4.

Algorithm 8. *Fast greedy heuristic crossover: modified steps of the greedy heuristic procedure (Algorithm 6).*

3: For each $j \in \chi_c$, calculate $\delta_j = F(\chi_c \setminus \{k\})$.
4.1: Sort δ_i and select a subset $\chi_{elim} = \{e_1, ..., e_{n_\delta}\} \subset \chi_c$ of n_δ indexes with minimal values of δ_i. Value $n_\delta \in \{\overline{1, |\chi_c| - p}\}$ must be calculated in each iteration. Maximum number of the extra data elements of set χ_c must be eliminated in the first iterations and only one element in the final iterations (final iterations coincide with Algorithm 6 or 7):

$$n_\delta = \max\{[(|\chi_c| - p) * \sigma_e], 1\}. \tag{6}$$

We ran Algorithm 8 with $\sigma_e = 0.2$. Smaller values ($\sigma_e < 0.0$) convert it into Algorithm 6 and make it work slower. Big values ($\sigma_e > 0.3$) change the order of eliminating the clusters and reduce the accuracy.
4.2: From χ_{elim}, remove close data vectors. For each $j \in \{\overline{2, |\chi_{elim}|}\}$, if $\exists k \in \{\overline{1, j-1}\} : L(A_{e_j}, A_{e_k}) < L_{min}$ then remove e_j from χ_{elim}.
4.3: Set $\chi_c = \chi_c \setminus \chi_{elim}$.

Algorithm 6 performs up to p iterations. For real large datasets, computational experiments demonstrate that p or $p-1$ iterations are performed in most cases (data vectors of the "parent" solutions at Step 3 of Algorithm 5 do not coincide). In each iteration, ALA algorithm runs $|\chi_c|$ times. Thus, ALA algorithm runs up to $2p + (2p-1) + ... + 1 = 2p^2 - p + 1$ times. Popular test datasets, BIRCH 1–3 are generated for testing algorithms on problems with 100 clusters. Thus, the ALA algorithm must run up to 19901 times.

Depending on parameter L_{min}, in each iteration of Algorithm 8 eliminates up to $\sigma_e \cdot p$ members from χ_c. If L_{min} is big and $\sigma_e = 0.2$, in the first iteration, ALA runs $2p$ times, in the second iteration $[1.8p]$ times, then $[1.64p]$, $[1.512p]$ times etc. In case of 100 clusters and big L_{min}, the ALA procedure runs $200 + 180 + 164 + 152 + 142 + 134 + 128 + 123 + 118 + 116 + 113 + 111 + 109 + 108 + 107 + 106 + 105 + 104 + 103 + 102 + 101 = 2626$ times only. Taking into account computational intenseness or the ALA procedure such as standard k-means algorithm which is estimated $O(N^{34}p^3 4\log^4(N)/\sigma^6)$ in case of independently perturbed data vectors by a normal distribution with variance σ^2 [7], reducing number of runs of the local search procedure is crucial in case of large-scale problems.

Step 3 of Algorithm 7 can be modified as follows.

Algorithm 9. *Fast greedy heuristic crossover for initial seeding: modified steps of the greedy heuristic procedure (Algorithm 7).*

3: For each $j \in \chi_c$, calculate $\delta_j = \sum_{i=1}^{N} (\min_{j \in (\chi_c \setminus \{k\})} w_i L(A_i, A_j))$.

4.1: Sort δ_i and select a subset $\chi_{elim} = \{e_1, ..., e_{n_\delta}\} \subset \chi_c$ of n_δ indexes with minimal values of δ_i.

4.2: For each $j \in \{\overline{2, |\chi_{elim}|}\}$, if $\exists k \in \{\overline{1, j-1}\} : L(A_{e_j}, A_{e_k}) < L_{min}$ then remove e_j from χ_{elim}.

4.3: Set $\chi_c = \chi_c \setminus \chi_{elim}$.

The aim of Step 4.2 of Algorithm 8 is to hold the order of elimination of the clusters provided by Algorithms 6 or 7. In Fig. 1, two cases of running Algorithm 8 are shown. Let us assume that $p = 4$ and distances between the centers of clusters 1 and 3, 3 and 4, 1 and 4, 6 and 7 are less than L_{min}. Let us assume that parameter σ_e allows eliminating up to 3 clusters simultaneously in the 1st iteration. After Step 3 of Algorithm 8 and sorting δ_i, we have a sequence of clusters $4,\ 3,\ 6, \ldots$. If Step 4.2 is included in Algorithm 8 then only one of clusters 1, 3 and 4 can be removed in the 1st iteration (Case A). Thus, only two clusters (4 and 7) are eliminated in the 1st iteration. If we remove Step 4.2 from our algorithm or assign big value to L_{min} then the simultaneous elimination of clusters 3 and 4 is allowed (Case B) which gives worse value of the squared distances sum. If the original Algorithm 7 runs, it eliminates cluster 4 first, then cluster 6. In its 3rd iteration, Algorithm 7 eliminates cluster 1 and we have the set of clusters shown in Fig. 1, Case A after two iterations which coincides with the result of Algorithm 8.

Algorithm 6 starts the ALA procedure many times, it is a precise but slow method. Having included Algorithm 8 into Algorithm 6, we reduce the number of starts of the ALA procedure, however, as explained above, at least 2626 starts of the local search algorithm in each iteration of the genetic algorithm in case of 100 clusters make using this method impossible for very a large dataset, especially for the Euclidean metric. Algorithm 7 optimizes the fitness function calculated for the initial seeding of the ALA procedure. This approach is fast, however, an optimal value of the fitness function for the initial seeding does not guarantee its optimal value for the final result of the ALA procedure.

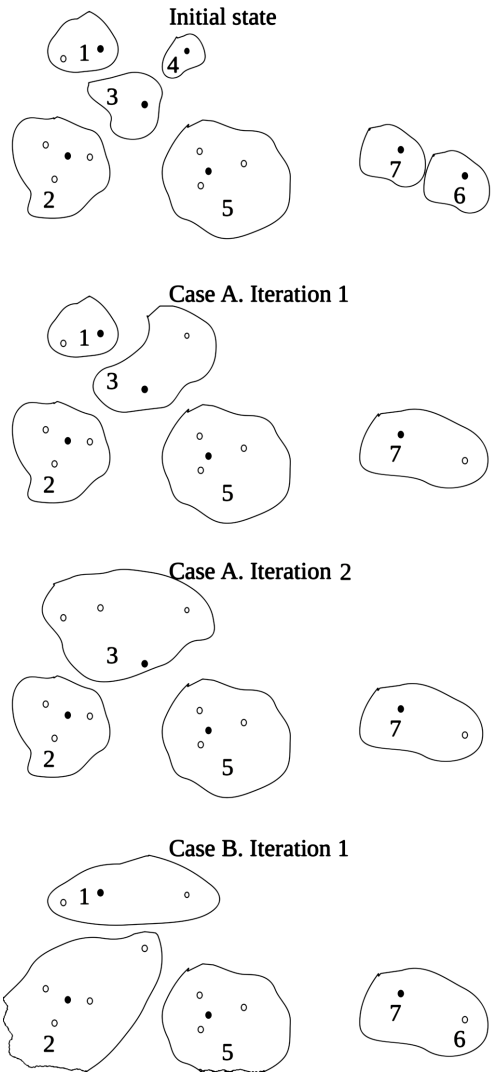


Figure 1: Succeeding and simultaneous elimination of clusters.

We propose a compromise version of two algorithms which implements one step of the ALA procedure after each elimination of the clusters. Since the result of the ALA procedure does not coincide with the data vectors A_i (in general), using integer numbers as the alphabet of the GA (i.e. for coding the solutions forming the population of the GA) is impossible and we use real vectors (coordinates of the interim solutions of the ALA procedure) for coding the solutions in the population of the GA. The whole algorithm is as follows.

Algorithm 10. *GA with greedy heuristic and floating point alphabet.*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters, population size N_p.

1: Initialize N_p sets of coordinates $\chi_1, ..., \chi_{N_p}$: $chi_i \subset \mathbb{R}^d$, $|\chi_k| = p \ \forall k = \overline{1, N_p}$ with solutions of the ALA algorithm initialized by the k-means++ procedure (Algorithm 2). Thus, each χ_i is a local minimum of (2). Store corresponding values of the function 2 to array $\mathcal{F}_1, ..., \mathcal{F}_{N_p}$.

2: If the stop conditions are reached then go to Step 7.

3: Choose randomly two "parent" sets χ_{k_1} and χ_{k_2}, $k_1, k_2 \in \{\overline{1}, N_p\}$, $k_1 \neq k_2$. Running Algorithm 11, obtain "child" set of coordinates χ_c which is a local minimum of (2). Store the value of (2) to $\mathcal{F}_c$.

4: If $\exists k \in \{\overline{1, N_p}\} : \chi_k = \chi_c$ then go to Step 2.

5: Choose index $k_{worst} = \arg\max_{k=\overline{1,N_p}} \mathcal{F}_k$. If $\mathcal{F}_{wotst} < \mathcal{F}_c$ then go to Step 2.

6: Choose randomly two indexes k_1 and k_2, $k_1 \neq k_2$; set $k_{worst} = \arg\max_{k \in \{k_1, k_2\}} \mathcal{F}_k$.

6: Replace $\chi_{k_{worst}}$ with χ_c, set $\mathcal{F}_{k_{worst}} = \mathcal{F}_c$ and go to Step 2.

7: STOP. The result is a set χ^*_k, $k^* = \arg\min_{k=\overline{1,N_p}} \mathcal{F}_k$.

The greedy heuristic procedure is modified as follows.

Algorithm 11. *Greedy crossover heuristic with floating point alphabet.*

Require: Set $V = (A_1, ..., A_N) \in \mathbb{R}^d$, number p of clusters, two "parent" sets of centers or centroids χ_{k_1} and χ_{k_2}, parameters σ_e and L_{min}.

1: Set $\chi_c = \chi_{k_1} \cup \chi_{k_2}$. Run the ALA procedure for $|\chi_c|$ clusters starting from χ_c. Store its result to χ_c.

2: If $|\chi_c| = p$ then run the ALA procedure with the initial solution χ_c, then STOP and return its result.

2.1: Calculate the distances from each data vector to the closest element of χ_c.

$$d_i = \min_{X \in \chi_c} L(X, A_i) \ \forall i = \overline{1, N}.$$

Assign each data vector to the corresponding cluster with its center in an element of χ_c.

$$C_i = \arg\min_{X \in \chi_c} L(X, A_i) \ \forall i = \overline{1, N}.$$

Calculate the distances from each data vector to the second closest element of χ_c.

$$D_i = \min_{Y \in (\chi_c \setminus \{C_i\})} L(Y, A_i).$$

3: For each $X \in \chi_c$, calculate $\delta_X = F(\chi_c \setminus \{X\}) = \sum_{i:C_i]X} (D_i - d_i)$.

4.1: Calculate n_δ in accordance with (6). Sort δ_X and select a subset $\chi_{elim} = \{X_1, ..., X_{n_\delta}\} \subset \chi_c$ of n_δ coordinates with minimal values of δ_X.

4.2: For each $j \in \{\overline{2, |\chi_{elim}|}\}$, if $\exists k \in \{\overline{1, j-1}\} : L(X_j, X_k) < L_{min}$ then remove X_j from χ_{elim}.

4.3: Set $\chi_c = \chi_c \setminus \chi_{elim}$.

4.4: Reassign data vectors to the closest centers or centroids.

$$C^*_i = \arg\min_{X \in \chi_c} L(X, A_i) \ \forall i = \overline{1, N}.$$

For each $X \in \chi_c$, if $\exists i \in \{\overline{1, N}\} : C_i = X$ and $C^* i \neq X$ then recalculate center or centroid X^* of the cluster $\mathcal{C}^{clust}_X = \{A_i | C^*_i = X, i = \overline{1, N}\}$. Set $\chi_c = (\chi_c \setminus \{X^*\}) \cup \{X\}$.

5: Go to Step 2.

An important parameter of Algorithms 8 and 11 is L_{min}. Performed series of experiments on various data, we propose the following method of its determining for each pair of centers or centroids X_j and X_k (see Step 4.2 of Algorithm 11):

$$L_{min} = \min_{X \in \chi_c} \{\max\{L(X, X_j), L(X, X_k)\}\}.$$

We ran this algorithm with large datasets and proved its efficiency experimentally.

4 Computational experiments

4.1 Datasets and computing facilities

For testing purposes, we used real data and generated datasets collected by Speech and Image Processing Unit of School of Computing of University of Eastern Finland[1] and UCI Machine Learning Repository[2]. Other authors use such problems for their experiments [56, 1, 43]. Number of data vectors N varies from 150 (classical Iris plant problem) to 581013 (Cover type dataset), number of dimensions d varies from 2 to 54, number of clusters from 3 to 1000. In addition, we used specially generated datasets for p-median problems (uniformly distributed data vecrots in $\mathbb{R}^2$, each coordinate in interval $[0; 10)$ with uniformly distributed weights in range $[0; 10)$).

Computational experiments were performed for problems with Euclidean (l_2) and squared Euclidean (l_2^2) distances (p-median and k-means problems, correspondingly).

For our experiments, we used a computer Depo X8Sti (6-core CPU Xeon X5650 2.67 GHz, 12Gb RAM), hyperthreading disabled and ifort compiler with full optimization and implicit parallelism (option -O3).

For algorithms comparison purposes, we ran each algorithm with each of datasets 30 times.

4.2 Algorithm parameters tuning

An important parameter of the genetic algorithm is number of individuals (candidate solutions) N_p in its population (population size). Running Algorithm 10 for the generated datasets ($d = 2$, $N = 1000$ and $N = 10000$, $p = 10$ and

[1] http://cs.joensuu.fi/sipu/datasets/

[2] https://archive.ics.uci.edu/ml/datasets.html

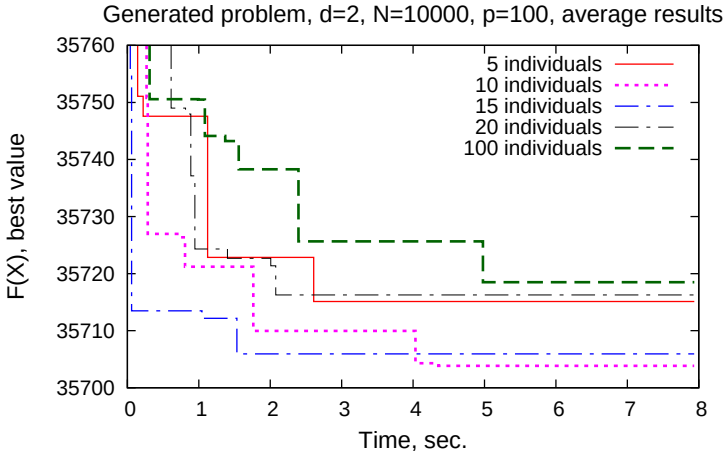


Figure 2: Results of Algorithm 10 with various population sizes.

$p = 100$) and real datasets (MissAmerica1 with $p = 100$) show that large populations ($N_p > 50$) slow down the convergence. Analogously, running with very small populations ($N_p < 10$) decrease the accuracy.

Results of running our algorithm for a generated problem with squared Euclidean metric are shown in Fig. 2. In this diagram, we fixed the best results achieved by the algorithms and the spent time after each improvement of the results in a special array. This diagram shows the average or worst results of 30 starts of the algorithms.

Experiments show that a population of 10–25 individuals is optimal for all tested problems for all greedy crossover heuristics considered in this paper (Algorithms 6, 7, 8 and 11). There is no obviois relation between d and N_p, p and N_p nor N and N_p. In all experiments below, we used $N_p = 15$.

4.3 Numerical results

For all datasets, we ran the genetic algorithm with greedy heuristic (Algorithm 5) with various crossover heuristics (Step 3 of Algorithm 5): its original version proposed by Alp et al. [3, 42] (Algorithm 6), its version for initial cluster centers seeding only (Algorithm 7), our modifications allowing elimination of many cluster centers in one step (Algorithm 8) and our new genetic algorithm with floating point alphabet (Algorithm 11).

Results for each of datasets are shown in Tables 1 and 2.

We used the sampling k-means procedure (Algorithm 3) for datasets with $N \geq 10000$ as the ALA procedure at Step 1 of Algorithms 5, 10 and 11. For smaller datasets, we ran all algorithms without sampling. However, running algorithms without sampling k-means procedure for large datasets equally delays the genetic algorithm with all considered greedy crossover heuristics.

Computation process with each of the algorithms was performed 30 times. Time limit shown in the first column was used as the stop condition. Value of this maximum running time was chosen so that adding equal additional time does not allow to obtain better results in case of using the original greedy crossover heuristic for initial seeding (Algorithm 7) in at least 27 attempts of 30. In addition, for the problems listed in Table 1, we fixed the average time needed for achieving the average result of the original genetic algorithm with greedy crossover heuristic (Algorithm 5 + Algorithm 6, see [3, 42]). For more complex problems listed in Table 2 where the original greedy crossover procedure is inapplicable due to huge computational complexity, we fixed the average time needed for achieving the average result of the original genetic algorithm with greedy crossover heuristic applied for optimizing the fitness function value of the initial dataset of the ALA procedure (Algorithm 5 + Algorithm 7).

Algorithm 5 with the original greedy crossover heuristic (Algorithm 6) proposed by Alp et al. [3, 42] shows excellent results for comparatively small datasets (see Table 1). For the least complex problems ("Iris" dataset), using the algorithm proposed in this article (Algorithm 10, Problems 1 and 3) reduces the accuracy of the solution in comparison with the original algorithm of Alp et al. [3, 42] (Algorithms 5 and 6). For larger datasets, new algorithm (Algorithm 10+Algorithm 11) is faster and more precise.

Moreover, using the original greedy crossover heuristic is impossible for large datasets (for all larger gatasets with $p > 30$, $N \geq 10000$) due to very intensive computation at each iteration. For such datasets, we used the algorithm of Alp et al. applied for optimizing the fitness function value of the initial dataset of the ALA procedure (Algorithms 5 and 7) for comparison purposes. In this case, for all solved large-scale test problems with both Euclidean (l_2, planar p-median problem) and squared Euclidean (l_2^2, k-means problem) metrics, our Algorithm 10 with floating point alphabet and modified greedy crossover heuristic (Algorithm 11) works faster and gives more precise results than Algorithm 5 with greedy heuristic implemented for the initial seeding only (Algorithm 7, [3, 42]).

Table 1: Results of running new Algorithm 11 and original genetic algorithm with greedy crossover heuristic.

Dataset, and its parameters, time limit	Distance	Method	Avgerage result	Average time needed for reaching result of the original method, sec.	Worst result	Avg. speedup (new vs. original method)
Iris ($n = 150, d = 4$, $p = 3$), 100 sec.	l_2^2	Original	$1.40026039044 \cdot 10^{14}$	0.0096	$1.40026039044 \cdot 10^{14}$	
		ALA mult.	$1.40026039044 \cdot 10^{14}$	0.0103	$1.40026039044 \cdot 10^{14}$	
		New	$1.400262 \cdot 10^{14}$	-	$1.4002858 \cdot 10^{14}$	-
Iris ($n = 150, d = 4$, $p = 10$), 100 sec.	l_2^2	Original	46916083985700	2.4	46916083985700	
		ALA mult.	46917584582154	-	46935815209300	
		New	46916083985700	2.5	46916083985700	-
				-		-
MissAmerica1 ($n = 6480, d = 16$, $p = 30$), 1500 sec.	l_2^2	Original	105571815.061	603	105663081.95	
		ALA mult.	105714622.427	-	106178506.965	
		New	105440299.602	13.8	105440299.601	43.69
				-		-
Europe ($n = 169309$, $d = 2, p = 10$), 1500 sec.	l_2^2	Original	1099348562.46	1050.8	1099355026.03	
		ALA mult.	1099345009.09	15.6	1099345033.08	
		New	1099345067.99	123.8	1099345210.55	8.48
				-		-

Note:"Original" algorithm is Algorithm 5 with original greedy crossover heuristic (Algorithm 6),
"ALA mult." algorithm is multiple start of the ALA procedure (Algorithm 4),
"New" algorithm is the genetic algorithm with floating point alphabet (Algorithm 10 with Algorithm 11 as the greedy crossover procedure).

To illustrate the dynamics of the solving process, we present the timing diagrams which show the average results of 30 runs of each algorithm for various datasets in Fig. 3 and 4. Diagrams show that new algorithm with floating point alphabet allows to increase the accuracy at early stages of the computation process in comparison with known methods which allows to use it for obtaining quick approximate solutions. In addition, results of the fast greedy heuristic (Algorithm 8) are shown in the diagrams. Using this heuristic without other modifications to the genetic algorithm can reduce the accuracy of the results.

5 Conclusion

New genetic algorithm based on ideas of the p-median genetic algorithm with greedy heuristic and two original procedures can be used for fast and precise solving the large-scale p-median and k-means problems. For the least complex problems, the results of our method are less precise than the original GA with greedy heuristic proposed by Alp et al. However, new algorithm provides more precise results in appropriate time for the large-scale problems.

References

[1] M. R. Ackermann et al. (2012) StreamKM: A Clustering Algorithm for Data Streams, *J. Exp. Algorithmics*, Vol 17, Article 2.4 (May 2012), DOI: 10.1145/2133803.2184450

[2] D. Aloise, A. Deshpande, P. Hansen, P. Popat (2009) NP-Hardness of Euclidean Sum-of-Squares Clustering, *Machine Learning*, Vol. 75, pp. 245–249, DOI: 10.1007/s10994-009-5103-0

[3] O. Alp, E. Erkut and Z. Drezner (2003) An Efficient Genetic Algorithm for the p-Median Problem, *Annals of Operations Research*, Vol.122, Issue 1–4, pp. 21–42, doi 10.1023/A:1026130003508

[4] A. Antamoshkin, L. Kazakovtsev (2013) Random Search Algorithm for the p-Median Problem, *Informatica (Ljubljana)*, Vol. 37, No. 3, pp. 267–278

[5] A. Antamoshkin, H. P. Schwefel, A. Toern, G. Yin, A. Zhilinskas (1993) *Systems Analysis, Design and Optimization. An Introduction*, Krasnoyarsk, Ofset

[6] P. Avella, M. Boccia, S. Salerno and I. Vasilyev (2012) An Aggregation Heuristic for Large Scale p-median Problem, *Computers & Operations Research*, 39 (7), pp. 1625–1632, doi 10.1016/j.cor.2011.09.016

Table 2: Results of running new Algorithm 11 and original genetic algorithm with greedy crossover heuristic used for initial seeding.

Dataset its parameters, time limit	Distance	Method	Avgerage result	Average time needed for reaching result of the original method, sec.	Worst result	Avg. speedup (new vs. original method for init. seeding)
Europe ($n = 169309$, $d = 2, p = 100$), 1200 sec.	l_2	Orig. init. ALA mult. New	400370576 400755480 400345100	397.6 - 193.6	400904292 401007437 400595350	- 2.05
Generic ($n = 10000$, $d = 2, p = 100$), 300 sec.	l_2	Orig. init. ALA mult. New	85318.44 85374.62 85167.01	47.65 - 0.895	85482.67 86114.83 85179.72	- 53.24
Europe ($n = 169309$, $d = 2, p = 100$), 1200 sec.	l_2^2	Orig. init. ALA mult. New	$2.2767 \cdot 10^{12}$ $2.3039 \cdot 10^{12}$ $2.2752 \cdot 10^{12}$	557.6 - 143.1	$2.2933 \cdot 10^{12}$ $2.3111 \cdot 10^{12}$ $2.2862 \cdot 10^{12}$	- 3.89
BIRCH1 ($n = 100000$, $d = 2, n = 100$), 120 sec.	l_2^2	Orig. init. ALA mult. New	$9.277283 \cdot 10^{13}$ $9.746921 \cdot 10^{13}$ $9.277282 \cdot 10^{13}$	46.08 - 31.56	$9.277287 \cdot 10^{13}$ $9.276386 \cdot 10^{13}$ $9.274231 \cdot 10^{13}$	- 1.46
BIRCH3 ($n = 100000$, $d = 2, p = 1000$), 2400 sec.	l_2^2	Orig. init. ALA Mult. New	$4.08652 \cdot 10^{12}$ $4.16053 \cdot 10^{12}$ $4.02040 \cdot 10^{12}$	802.8 - 8.81	$4.105231 \cdot 10^{12}$ $4.162683 \cdot 10^{12}$ $4.022190 \cdot 10^{12}$	- 91.12
MissAmerica2 ($n = 6480$, $d = 16, p = 100$), 300 sec.	l_2^2	Orig.init. ALA Mult. New	80264286 81869326 79837119	85.8 - 0.752	81039148 82316364 80147971	- 114.1
CoverType ($n = 581013$, $d = 54, p = 100$), 2400 sec.	l_2^2	Orig.init. ALA Mult. New	3122934.7 3163271.9 3069213.6	905.4 - 53.1	3146107.4 3182076.8 3072299.5	- 17.05

Note: "Orig. init." algorithm is Algorithm 5 with original greedy crossover heuristic (Algorithm 7),
"ALA mult." algorithm is multiple start of the ALA procedure (Algorithm 4),
"New" algorithm is the genetic algorithm with floating point alphabet (Algorithm 10 with Algorithm 11 as the greedy crossover procedure).

[7] D. Arthur, B. Manthey and H. Roglin (2009) k-Means Has Polynomial Smoothed Complexity, *in Proceedings of the 2009 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS '09)*, IEEE Computer Society, Washington, DC, USA, pp. 405–414 DOI: 10.1109/FOCS.2009.14

[8] D. Arthur and S. Vassilvitskii (2007) k-Means++: The Advantages of Careful Seeding, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete algorithms*, ser. SODA '07, pp. 1027–1035

[9] K. Bailey (1994) Numerical Taxonomy and Cluster Analysis, in: Typologies and Taxonomies, Sage Pubns, DOI: 10.4135/9781412986397

[10] B. Bozkaya, J. Zhang and E. Erkut (2002) A Genetic Algorithm for the p-Median Problem, *in Z. Drezner and H. Hamacher (eds.), Facility Location: Applications and Theory*, Springer

[11] A. V. Cabot, R. L. Francis and M. A. Stary (1970) A Network Flow Solution to a Rectilinear Distance Facility Location roblem, *American Institute of Industrial Engineers Transactions*, 2, pp. 132–141

[12] L. O'Callaghan, A. Meyerson, R. Motwani, N. Mishra and S. Guha (2002) Streaming-Data Algorithms for High-Quality Clustering, *Data Engineering, 2002. Proceedings. 18th In-*

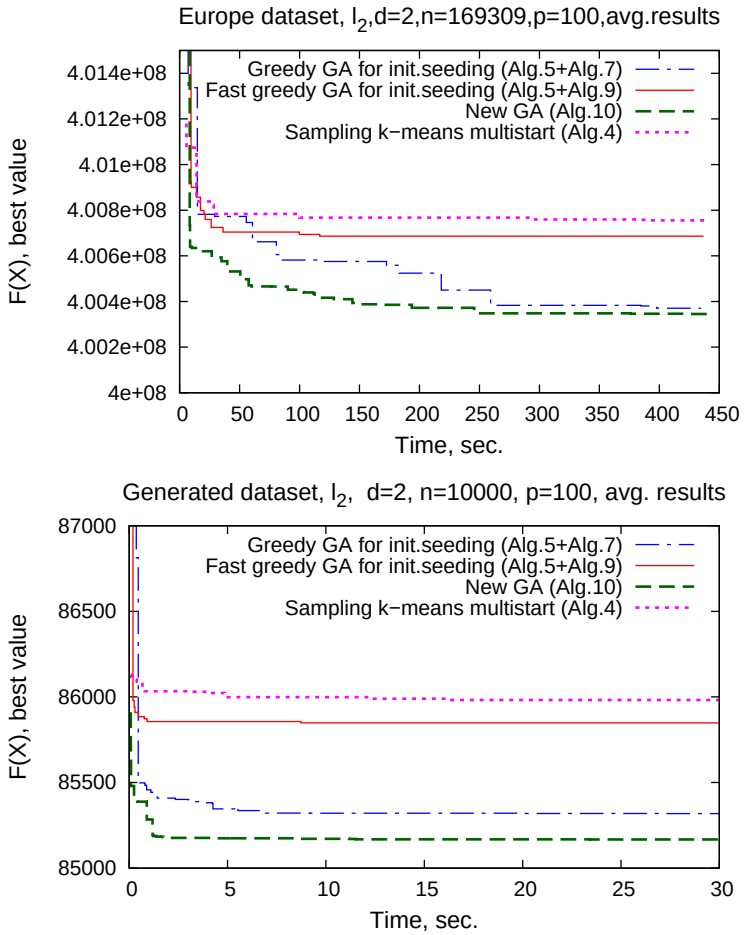


Figure 3: Average results of running new GA in comparison with other algorithms for $p-$median problems.

ternational Conference on, pp. 685–694 DOI: 10.1109/ICDE.2002.994785

[13] L. Cooper (1963) Location-allocation problem, *Oper Res*, vol. 11, pp. 331–343

[14] L. Cooper (1968) An Extension of the Generalized Weber Problem, *Journal of Regional Science*, Vol. 8, Issue 2, pp.181–197

[15] A. Czumaj and C. Sohler (2004) Sublinear-Time Approximation for Clustering Via Random Sampling, *in Automata, Languages and Programming, Lecture Notes in Computer Science*, Vol. 3142, Springer Berlin Heidelberg, pp. 396–407, DOI: 10.1007/978-3-540-27836-8_35

[16] M. M. Deza, E. Deza (2013) Metrics on Normed Structures, *Encyclopedia of Distances,* Springer Berlin Heidelberg, pp.89–99, DOI: 10.1007/978-3-642-30958-8_5

[17] Z. Drezner (2013) The Fortified Weiszfeld Algorithm for Solving the Weber Problem, *IMA Journal of Management Mathematics*, published online. DOI: 10.1093/imaman/dpt019

[18] Z. Drezner and H. Hawacher (2004) *Facility location: applications and theory*, Springer-Verlag, Berlin, Heidelberg.

[19] Z. Drezner and G. O. Wesolowsky (1978) A Trajectory Method for the Optimization of the Multifacility Location Problem with lp Distances, *Management Science*, 24, pp.1507-1514

[20] P. Drineas, A. Frieze, R. Kannan, S. Vempala and V. Vinay (1999) Clustering Large Graphs via the Singular Value Decomposition, *Machine learning*, vol. 56(1-3), pp. 9–33

[21] B. S. Duran, P. L. Odell (1977) *Cluster Analysys: a Survey*, Springer-Verlag Berlin–Heidelberg–New York

[22] R. Z. Farahani and M. Hekmatfar, editors (2009) *Facility Location Concepts, Models, Algorithms and Case Studies*, Springer-Verlag Berlin Heidelberg.

[23] V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell and J. French (1999) Clustering large datasets in arbitrary metric spaces, *Proc. 15th Int. Conf. Data Engineering*, pp. 502–511

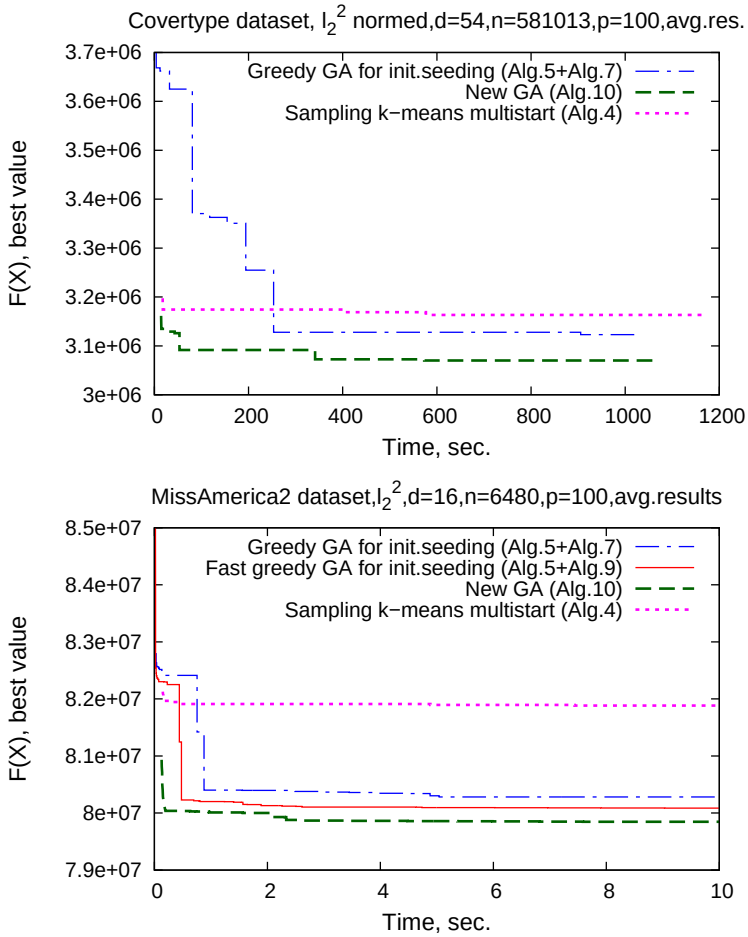


Figure 4: Average results of running new GA in comparison with other algorithms for k-means problems.

[24] S. L. Hakimi (1964) Optimum Locations Of Switching Centers and the Absolute Centers and Medians of a Graph, *Operations Research*, 12(3), pp. 450-459

[25] P. Hansen, J. Brimberg, D. Urošević, N. Mladenović (2009) Solving large p-median clustering problems by primal-dual variable neighborhood search, *Data Mining and Knowledge Discovery*, vol. 19, issue 3, pp 351–375

[26] P. Hansen, E. Ngai, B. Cheung, N. Mladenović (2005) Analysis of Global k-Means, an Incremental Heuristic for Minimum Sum of Squares, *Journal of Classification*, vol. 22, issue 3, pp 287–310

[27] C. M. Hosage and M. F. Goodchild (1986) Discrete Space Location–Allocation Solutions from Genetic Algorithms, *Annals of Operations Research* 6, 35–46.

[28] C. R. Houck, J. A. Joines and M. G. Kay (1996) Comparison of Genetic Algorithms, Random Restart, and Two-Opt Switching for Solving Large Location-Allocation Problems, *Computers and Operations Research*, Vol 23, pp. 587–596

[29] R. Jaiswal, A. Kumar and S. Sen (2013) A Simple D2-Sampling Based PTAS for k-Means and Other Clustering Problems, *Algorithmica*, DOI: 10.1007/s00453-013-9833-9

[30] L. Kazakovtsev (2013) Wireless Coverage Optimization Based on Data Provided by Built-In Measurement Tools, *WASJ*, vol.22, spl. issue Tech. and Technol., pp. 8–15

[31] K. Krishna, M. Narasimha Murty (1999) Genetic K-Means Algorithm, *IEEE Transactions on Systems, Man, and Cybernetics – part B: Cybernetics*, Vol. 29, No. 3, pp. 433–439

[32] K. Liao and D. Guo (2008) A Clustering-Based Approach to the Capacitated Facility Location Problem, *Transactions in GIS*, vol. 12 (3), pp.323–339

[33] S. P. Lloyd (1982) Least Squares Quantization in PCM, *IEEE Transactions on Information Theory*, Vol. 28, pp. 129–137

[34] J. B. MacQueen (1967) Some Methods of Classification and Analysis of Multivariate Observations, *Proceedings of the 5th Berkley Symposiom on Mathematical Statistics and Probability*, vol. 1, pp. 281–297

[35] S. Masuyama, T. Ibaraki and T. Hasegawa (1981) The Computational Complexity of the m-Center Problems on the Plane, *The Transactions of the Institute of Electronics and Comunication Engineers of Japan*, 64E, pp. 57–64

[36] U. Maulik, S. Bandyopadhyay (2000) Genetic Algorithm-Based Clustering Technique, *Pattern Recognition,* Vol. 33, pp. 1455–1465

[37] L. A. A. Meira and F K. Miyazawa (2008) A Continuous Facility Location Problem and Its Application to a Clustering Problem, *Proceedings of the 2008 ACM symposium on Applied computing (SAC '08)*. ACM, New York, USA, pp.1826–1831. doi: 10.1145/1363686.1364126

[38] N. Mishra, D. Oblinger and L. Pitt (2001) Sublinear time approximate clustering, *12th SODA*, pp. 439–447

[39] N. Mladenović, J. Brimberg, P. Hansen, J. A. Moreno-Perez (2007) The p-median problem: A survey of metaheuristic approaches, *European Journal of Operational Research*, Vol. 179, issue 3, pp.927–939

[40] J. G. Morris (1981) Convergence of the Weiszfeld Algorithm for Weber Problem Using a Generalized "Distance" Function, *Operations Research*, vol. 29 no. 1, pp. 37–48

[41] I. A. Osinuga and O. N. Bamigbola (2007) On the Minimum Norm Solution to the Weber Problem, *SAMSA Conference proceedings*, pp. 27–30

[42] M. N. Neema, K. M. Maniruzzaman and A. Ohgai (2011) New Genetic Algorithms Based Approaches to Continuous p-Median Problem, *Netw. Spat. Econ.*, Vol. 11, pp. 83–99, DOI: 10.1007/s11067-008-9084-5

[43] P. Phoungphol and Y. Zhang (2011) Sample Size Estimation with High Confidence for Large Scale Clustering, it Proceedings of the 3rd International Conference on Intelligent Computing and Intelligent Systems, http://www.cs.gsu.edu/ pphoungphol1/paper/icis2011.pdf

[44] A. W. Reza, K. Dimyati, K. A. Noordin, A. S. M. Z. Kausar, Md. S. Sarker (2012) A Comprehensive Study of Optimization Algorithm for Wireless Coverage in Indoor Area, *Optimization Letters*, September 2012, published online, doi 10.1007/s11590-012-0543-z, http://link.springer.com/article/10.1007 %2Fs11590-012-0543-z?LI=true

[45] M. G. C. Resende (2008) Metaheuristic hybridization with Greedy Randomized Adaptive Search Procedures, *in TutORials in Operations Research, Zhi-Long Chen and S. Raghavan (Eds.)*, INFORMS, pp. 295–319

[46] M. G. C. Resende, C. C. Ribeiro, F. Glover and R. Marti (2010) Scatter search and path-relinking: Fundamentals, advances, and applications, *Handbook of Metaheuristics, 2nd Edition, M. Gendreau and J.-Y. Potvin, Eds.*, Springer pp. 87–107

[47] X. Rui and D. Wunsch (2005) Survey of Clustering Algorithms, *Neural Networks, IEEE Transactions on*, vol. 16, no. 3, pp.645–678, doi: 10.1109/TNN.2005.845141

[48] S. Still, W. Bialek and L. Bottou (2004) Geometric Clustering using the Information Bottleneck method, in: *Advances In Neural Information Processing Systems 16* Eds.: S. Thrun,L. Saul, and B. Scholkopf, MIT Press, Cambridge, MA.

[49] P.-N. Tan, M. Steinbach, V. Kumar (2006) Cluster Analysis: Basic Concepts and Algorithms, Chapter 8 in: *Introduction to Data Mining*, Addison-Wesley, pp. 487–567

[50] V. A. Trubin (1978) Effective algorithm for the Weber problem with a rectangular metric, *Cybernetics and Systems Analysis*, **14(6)**, DOI:10.1007/BF01070282, Translated from Kibernetika, **6** (November-December 1978), pp. 67–70.

[51] Y. Vardi and C.-H. Zhang (2001) A Modified Weiszfeld Algorithm for the Fermat-Weber Location Problem, *Math. Program., Ser. A*, vol. 90: pp. 559–566, DOI: 10.1007/s101070100222

[52] A. Weber (1922) *Über den Standort der Industrien, Erster Teil: Reine Theorie des Standortes*, Tübingen, Mohr

[53] E. Weiszfeld (1937) Sur le point sur lequel la somme des distances de n points donnes est minimum, *Tohoku Mathematical Journal*, **43** no.1, pp. 335–386.

[54] G. Wesolowsky (1992) The Weber problem: History and perspectives, *Location Science*, **1**, pp. 5–23.

[55] G. O. Wesolowsky and R. F. Love (1972) A nonlinear Approximation Method for Solving a Generalized Rectangular Distance Weber Problem, *Management Science*, vol. 18 no. 11, pp. 656–663

[56] T. Zhang, R. Ramakrishnan and M. Livny (1996) BIRCH: An Efficient Data Clustering Method for Very Large Databases, *Proceedings of the 1996 ACM SIGMOD international conference on Management of data (SIGMOD '96)*, ACM, New York, NY, USA, pp. 103–114, DOI: 10.1145/233269.233324

A Cognitonics Approach to Computer Supported Learning in the Mexican State of Oaxaca

Paul Craig
Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu Province, China
E-mail: p.craig@xjtlu.edu.cn, paulspapers.com

Néna Roa-Seïler
Edinburgh Napier University, Edinburgh, United Kingdom
Universidad Tecnológica de la Mixteca, Huajuapan de León, Oaxaca, México
E-mail: n.roa-seiler@napier.ac.uk

Marcela Martínez Díaz
Universidad Tecnológica de la Mixteca, Huajuapan de León, Oaxaca, México
E-mail: mtz.diaz.marce@gmail.com

Felipe Lara Rosano
Universidad Nacional Autónoma de México, Ciudad de México, Distrito Feder, México
E-mail: flararosano@gmail.com

Keywords: cognitonics, human computer interaction, educational videogames, collaborative learning

Received: July 27, 2014

Cognitonics is a new science which looks at ways to reconcile human socio-spiritual development with increasingly rapid human intellectual development in the new context of technological advances and increased cultural homogeny. This is particularly relevant in areas such as education and informatics where children are found to be increasingly capable to control and adapt to new technological advances yet often suffer from a lack of social development or are unable to engage with aspects of their own cultural heritage. In this study we consider the application of a cognitonics based approach to the problems of the Oaxacan education system, particularly for indigenous children who suffer from a loss of culture and diminished provision of education due to a lack of resources and regular teacher strikes. Specifically, we look at how the introduction of face-to-face collaborative video games can help develop academic, information-technology and social skills together while promoting spiritual well-being and cultural identity.

Povzetek: Predstavljena je kognitonska metoda učenja v mehiški državi Oaxaca.

1 Introduction

Oaxaca State is located in the south west of Mexico bordering Puebla and Veracruz to the north, Guerrero and Chiapas to the East and West, with the Pacific Ocean to the south. The overall population is around 3.5 million with about a third of the population speaking one of sixteen formally recognised indigenous languages. Around half of these do not speak Spanish and the remaining two-thirds of the population are predominantly mixed or indigenous peoples who have lost their language. The rugged terrain and linguistic or cultural differences mean that people often live in small secluded communities. Many of these communities suffer from limited access to services and employment opportunities. Overall, 53% of the population live in rural areas [1] and 67.2% live in poverty [2]. 60% of the population are under 30 and more than a third of the total population are enrolled in the educational system. Around 700,000 of these are children in primary education [3].

Despite their rich cultural heritage and the resilience of Oaxacan people to general hardship, they are currently faced with a number of serious social and economic challenges. Primary education is particularly problematic with Mixtec municipalities accounting for the majority of the 80% in Oaxaca not adhering to minimum requirements set by the Mexican government [4]. Only 5% of indigenous persons in the state attain a grade beyond primary school level and over 21% of the overall state population is illiterate [4]. These problems can be attributed to a number of factors including low family income [5], disruption of family structures due to high rates of migration [4, 5] and the large percentage of the population that live in remote rural areas [6]. There is also the significant problem that many indigenous teachers have not received formal training [7] and a strong sense that the education system is not properly adapted to best serve the indigenous population.

Most of these social, economic and political problems have their roots in, and are exacerbated by, corruption, which is endemic in Mexico as a whole [8]. Corruption at the fundamental level is the illegitimate use of public power to benefit private interest. This is when individuals expect illicit payback for services or preferential treatment. While this occurs on different levels with different grades of severity, the result is that people are denied access to opportunities and official structures fail to function in an efficient or just manner. This leads to inequality that leads to frustration, social unrest and criminal activity.

A lack of altruistic community spirit [9] or, at a very basic level, a lack of common compassion is at the root of corruption. A tendency toward corrupt behaviour is not however a natural attribute of the Mexican people but rather a result of the social conditions such as economic hardship and marginalization caused by factors such as the large scale migration of people to urban areas and the subsequent loss of established community structures.

Our proposal for working toward a partial solution to this problem is to support the provision of innovative education methods such as collaborative learning. Collaborative learning is where two or more students learn together by working on the same problem. This allows the students to learn through shared experience and face-to-face interaction, in effect capitalising on the inherent social nature of learning [10]. Many of the attributes promoted by collaborative learning are important for promoting community spirit and tackling the root causes of corruption. Important aspects of collaborative learning are positive interdependence, individual accountability, face-to-face promotive interaction, social skills and group processing (or self-analysis of the group) [11]. Our hope is that by developing these abilities at an early age, when young people begin to define themselves as individuals and develop intellectually, we can help avoid the development of the converse negative traits later in life.

Specifically, we look at how collaborative educational videogames can be used to help primary school children develop collaborative and proper social skills while achieving learning objectives that would normally be taught in a traditional classroom environment. This combines the teaching of the classroom syllabus with an introduction to information-technology and the development of essential social skills so as to balance the development of the intellectual and socio-spiritual sides of the student in accordance with the philosophy of cognitonics [12].

2 Related work

Cognitonics aims, in particular, to help people adapt to and use technology by improving cognitive mechanisms of processing information and developing the emotional sphere of the personality [12-14]. Examples of cognitonics in education are the use of broadband tele-conferencing to allow young people to interface with public figures [15], using technology for the self-evaluation of history lessons [16] and building a mental model of student online activity in an online e-learning environment [17]. Collaborative face-to-face videogames also fall within the scope of cognitonics, since they aim to promote a more sociable and culturally relevant learning experience [18-20]. While the concept of using collaborative videogames for cognitonics is a relatively new development, there is a long history of technology being used in Mexican education and in many regards collaborative games can be seen as a logical progression from these technologies.

2.1 Distance learning

Over the years various Mexican governments have recognized the problems of delivering education in remote and marginalized areas and sought to remedy these problems through the application of technology [21]. The most successful programs have been in the area of distance learning [22] supported by television networks and courses delivered through the mail. Computer based learning has been somewhat less successful due to problems of IT infrastructure and the cost of equipment.

In the late nineties Ernesto Zedillo introduced a program of Distance Education aimed at bringing quality education to remote areas without the necessity for students to relocate or travel large distances to attend classes. Since then, three distinct projects have been put into place to provide Distance Learning in Mexico. These are the Educational Satellite Television Network (EDUSAT) [23] which provides support for the training and development of teachers, the Red Escolar (Scholar Network) [24] which provides education in information technologies and Telesecundaria (Tele-secondary school) [25] providing general and technical secondary education.

The provision of computer based distance learning material is limited in Mexico and its usefulness is questionable as very few Mexicans stand to benefit from this type of education in its traditional form. Traditionally, computer based distance learning means learning at home and few Mexicans have a suitable computer at home. According to the INEG, as of 2006, only 58.7% of the population have access to a personal computer with only 45% having access to the internet [26]. These figures are likely to be a lot worse for less advantaged sections of the population who stand to benefit the most from distance learning programmes.

2.2 Computer based distance learning

Enciclomedia [27] is the Mexican governments most committed effort in the area of computer based distance learning to date. This was released in the 1990s, incorporating videos, text, virtual visits, sounds and images to complement free textbooks. Later versions of the service incorporated content from Encarta [28] provided by Microsoft and integrated resources, activities and audiovisuals generated by projects such as the Red Escolar [24], Biblioteca Digital, Sec 21 and Sepiensa [29]. However, the project was not generally considered a success. Teachers found the material to be inconsistent

and, according to primary teachers in the nation capital, the program ceased to be used altogether from the 2010 when as part of the Basic Education Reform the content free textbook was changed without the program being updated [30].

Other private initiatives tended to concentrate on the provision of computer equipment in the classroom. In 2009, Mexican billionaire Carlos Slim donated 50,000 laptops for use in Mexican schools from the One Laptop per Child (OLPC) program [31]. Kids on Computers [32] is another important program, working in Oaxaca to recycle used computers for use in rural schools. A general criticism of these types of program is that they can do little to serve the immediate needs of the population such as the supply of food and availability of clean water [33]. This isn't so much of an issue however in nations such as Mexico where the problems of development are more to do with organization and infrastructure rather than resources. In the Mexican context these programs certainly have a great potential to improve conditions for the general population.

2.3 Collaborative learning games

The use of collaborative games is a form of learning that has already shown promise to improve the education of children while at the same time helping to develop positive character traits. This is reflected in the vast majority of research on the subject which indicates that students learn more effectively when they work collaboratively [34]. Several studies point out benefits of using collaborative methods in education [35]. These include that;

- Students learn more.
- Students are more positive about school, subject areas, and teachers or professors.
- Students are more positive about each other regardless of differences in ability, ethnic background, or physical disability.
- Students are more effective interpersonally as a result of working cooperatively: students with cooperative experiences are more able to take the perspective of others, are more positive about taking part in controversy, have better developed interaction skills, and have a more positive expectation about working with others.

These results aren't surprising since we already know that students tend to learn more in social situations [10], and collaborative learning is intrinsically social. Moreover, there is strong evidence that indigenous Mesoamerican peoples in particular have a cultural disposition to collaborative learning rather than the traditional directed approach [36], and children with this background may not adapt well to the more authoritarian European-American classroom model [37]. These are all reasons to assume that collaborative learning games might form part of a successful strategy to improve primary education in Oaxaca.

3 Collaborative face-to-face videogames in the state of Oaxaca

Over-and-above the aforementioned pedagogical benefits of collaborative learning, there are a number of practical reasons why collaborative videogames could form part of a realistic solution to the problems of the Oaxacan education system. Videogames do not depend on supervision or on language, they are also relatively cheap to implement and maintain, and suit the way that children naturally learn. Collaborative games can also help children learn how to work together. This ability to work together is in itself an important life-skill that's often neglected in the traditional model of classroom education where children strive toward individual merit over the good of the group. By encouraging students to develop team working skills at an early age we feel this should better equip them to grow into adults who are more able to work together towards resolving some of the more entrenched problems of Oaxaca and Mexican society.

Developing these skills at the same time as they learn to use new technologies should also allow the students to develop a more wholesome relationship with technology and suffer less from adverse effects such as isolation and retarded social skills [13, 38].

Figure 1. Setup of the games. Left, custom games stand set at 45 degrees and, right, children playing the games with a robot assistant.

Figure 2: Mixtec codices used as characters in the mathematics and languages videogames: Jaguar, eagle, muerte and mixteco.

3.1 Methodology

Our experiments to investigate how collaborative videogames might be used in Oaxaca involved six groups of three children aged eight to ten years. Each group spent two hours in total playing three educational videogames. Games were played on a forty-two inch multi-touch screen angled at forty-five degrees and raised between waist and head height to be ergonomically accessible (see figure 1). The children were observed through two-way glass with audio and video recorded throughout the sessions to give us a permanent record of results. The groups consisted of four groups all female and two groups all male.

Each child was tested immediately before, immediately after and four days after their session. The exams used for testing included three five minute sessions testing mathematics, languages and reading. The students were also asked to fill in questionnaires to provide us with more subjective information relating to how they felt about the games and working as a team. In addition to this, observations made during the tests allowed us to assess the dynamics of the groups telling us how the students interacted and how collaboration strategies evolved.

3.2 Videogame design

The three educational games developed for our experiment supported the learning of mathematics, languages and reading. In order for the games to be both accessible and challenging for children with different levels of learning, we incorporated a gradually increasing level of difficulty for each game. Other key aspects of game design were promotion of inter-student interaction, cultural relevance and age appropriateness. Here we tried to ensure that the games were non-violent and did not enforce gender stereotypes while encouraging the children to identify with elements of their native Mixtec and Mexican culture.

Two of the three educational videogames developed for the project (those designed for mathematics and language learning) make use of graphics based on Mixtec codices (see figure 2). The codices are a form of colourful hieroglyphic used by the early Mixtecos to record their history. While these are no longer used today for writing, they remain a strong symbol of Mixtec culture used in logos, books and t-shirt designs. Parts of the costumes seen in the codices are also used in traditional ceremonies and festivals. The codices used in the games are the jaguar, the eagle, the muerte, and the Mixtec man. Muerte can be literally translated as death, and the character used in our game represents a dead friend or relative returning to visit the living. To western eyes this might seem a morbid character to include in a video game for young children, but the Mixtecos have a somewhat different attitude to the symbolism surrounding death. Mixtecos consider the ‘day of the dead’, when the dead are said to return to visit their loved ones, as a happy occasion to be celebrated with bright colours and loud music.

The mathematics game developed for the project

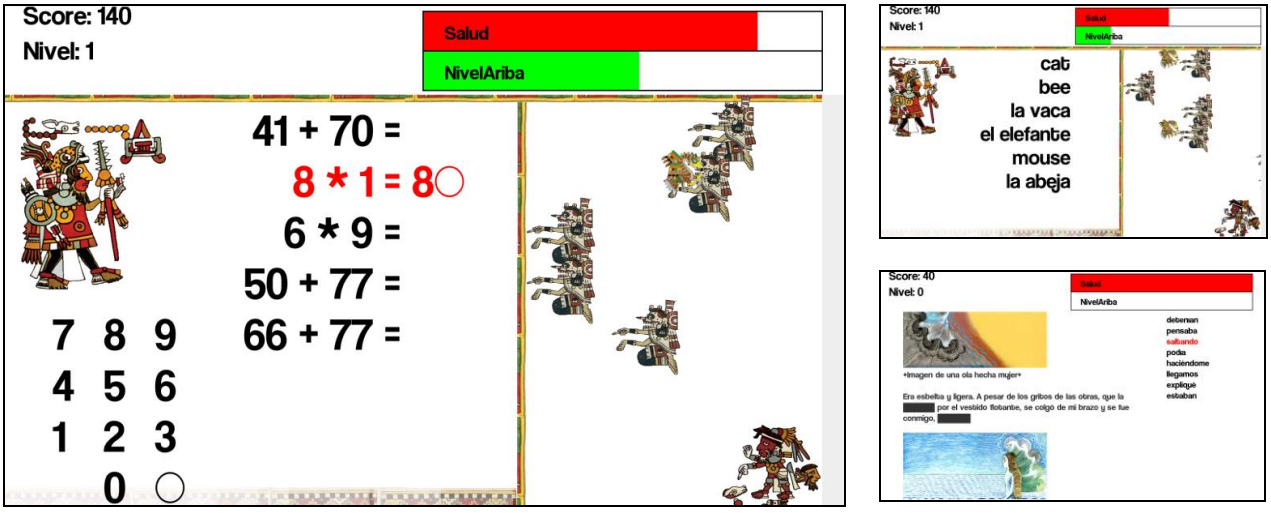


Figure 3. Screenshots of the educational videogames: left mathematics, top right languages and bottom right reading.

(figure 3 left) is a simple 'tower defence' type game where the students have to solve mathematical equations to fire eagles and prevent the muertes from reaching the perimeter wall of their 'tower' and draining their energy. The character of the user, the jaguar, sits at the left hand side of the screen. Below this character is a keypad and to the right a list of sums. To the right of the sums is a vertical wall and beyond the wall are the muertes. Each muerte advances slowly from right to left toward a sum and if a muerte reaches the wall it stops and begins to drain the health of the user. When the health of the user reaches zero, the game is over. In order to stop muertes reaching the wall, the user can answer sums to fire eagles. The user can press on different sums to answer them using the keypad. If a sum is answered incorrectly, health is drained, and if a sum is answered correctly, an eagle is fired from the wall toward the right hand side of the screen. When a muerte is hit by an eagle, it is pushed back away from the wall. Pushing back muertes also adds to the users score and causes the level-up bar to rise. When the level up bar is full, every sum fires an eagle to push all the muertes back, and the user progresses to the next level. As the level increases, the muertes begin to speed up, and it becomes gradually more difficult to do all the sums on time to keep the muertes away from the wall.

The languages game help students to learn the names of animals in English (figure 3 top right). This game is another 'tower defense' type game similar to that used to learn mathematics described above. However, instead of doing sums to fire eagles, the students have to match words in English to their Spanish translations. When words are matched correctly eagles are fired from both words and health is drained when words are matched incorrectly. The game begins with a small number of words for more common animals such as cats and dogs. As the game advances, the difficulty level increases with a wider variety of gradually more obscure animal names. If the children are not already familiar with the names of these animals in English, they can normally find the translations out by trial and error and learn from their mistakes.

The reading game (figure 3 bottom right) aims to help the children with reading by asking them to complete a story by replacing missing verbs. Literature and authors are held in particular esteem in Mexico, and Latin America in general, with popular authors often considered as national heroes. This game encourages the children to explore the Mexican national side of their identity by using an adaptation of the short story 'Mi vida con la ola' by Mexican Nobel laureate for literature Octavio Paz.

3.3 Results

Our experiments provided us with three types of results. Firstly, short exams taken by the students immediately before, immediately after, and four days after the experiments allowed us to assess how the games contributed towards specific short-term learning objectives. In addition to these exams, the students were asked to fill in questionnaires to provide us with more subjective information relating to how the students felt about working as a team throughout the sessions. Finally, observations made during the tests allowed us to observe the dynamics of the groups, and how strategies evolved during the sessions.

3.4 Exams

Table 1 summarizes the improvement in the children's performance in the exams after the session with the educational games. Here we can see that the children's performance did not improve significantly, or deteriorated, immediately after their session with the games. This was most likely due to the children being tired and over-stimulated after playing the games for two hours. When the students were tested again, four days after the tests, there was a significant improvement in their performance. This improvement was particularly marked for mathematics where the student's performance showed an increase of 21.9%. The improvement for languages was 4.3% and the students regressed slightly in their reading (by 2.1%). In order to statistically validate our results and account for inter-sample variance, we performed a single-tailed t-test. This gave a p-value of 0.016 for the second test to indicate that it was highly likely the children's improvement was due to their exposure to the games rather than variation of the children's scores overall. The p-value for improvement in the mathematics test was 0.0018, indicating a greater probability that the children's improvement was due to their exposure to the games. The p-value for the language test was marginal at 0.20 (0.089 for the boys), indicating

Table 1. Improvement in child performance assessed by exams administered after a session with collaborative educational games.

		Improvement in performance							
		Immediately after the test				four days after the test			
		Maths	Languages	Reading	All	Maths	Languages	Reading	All
girls	%	10.00%	0.96%	0.00%	3.65%	22.08%	2.56%	1.04%	8.56%
	p-value	0.111	0.421	0.500	0.243	0.012	0.215	0.500	0.393
boys	%	9.17%	-0.64%	-25.00%	-5.49%	21.67%	7.69%	-8.33%	7.01%
	p-value	0.065	0.468	0.005	0.130	0.012	0.089	0.282	0.016
all children	%	9.7%	0.4%	-8.3%	0.6%	21.9%	4.3%	-2.1%	8.0%
	p-value	0.067	0.458	0.148	0.439	0.002	0.198	0.387	0.031

that there is insufficient evidence to conclude that exposure to the games caused an improvement in the children's results (using the standard p-value threshold of 0.05). Results also indicate that there is insufficient evidence to conclude that the *drop* in reading performance was due to exposure to the reading game (with a p-value of 0.38).

3.5 Questionnaires and observations

The questionnaires filled in by students and observations made during the experiments provided us with more subjective information regarding the benefits of our educational games. The first thing we noticed was the short time the children took to learn how to play the games. On average it took around two minutes for the children to develop an understanding of how each game worked. During the following five or ten minutes the children would develop basic game strategies and continue developing these while playing the games. The students also developed collaboration strategies such as coordinated turn taking, task delegation and thinking aloud. In general, the boys tended to prefer the mathematics game while the girls preferred the reading game. The boys and girls also tended to use different strategies for the reading game. The girls would read larger sections of the text aloud, following the story. Boys tended to use a more direct strategy of reading individual sentences and trying to use grammatical rules to choose a word. The boys also tended to be more competitive, celebrate more when an answer was correct, and argue more over whose turn it would be to operate the interface. All of the student groups spent around about equal time playing each of the different games and tended to spend around twenty minutes or half an hour playing a game before moving on to the next.

The positive feelings the children had toward the games were reflected both in observed behaviour and questionnaire results. At the end of the sessions the children wanted to continue playing even after two hours. In the questionnaires the children told us they enjoyed the gaming sessions and would be very happy to use the games for future learning. The children particularly enjoyed being able to learn together with their fellow students and found the graphical nature of the games stimulating. They recognized the characters from the codices and felt this helped them relate to the games. The story used in the reading game was not familiar to the students but those who followed the text in the session expressed an interest in learning more about the story.

4 Conclusion

The results of the evaluation presented in this paper demonstrate the potential of collaborative games to improve the educational experience of children in the Mixtec region of Mexico. This includes the development of academic abilities and inter-personal skills together so as to realise a more balanced learning experience and prevent more negative personality traits such as isolation and anti-social tendencies that might otherwise be the result of less interactive education models such as private study with a computer [13, 14].

Evaluating exam scores before and after gaming sessions showed a statistically significant improvement of over 20% in results for mathematics. Results for the languages were positive though not conclusive due to the small sample size and natural variation between student grades. While the exam results did not show an improvement for reading, a number of children involved in the study felt encouraged to develop an interest in the story presented during the game.

More importantly, the games allowed the children to develop important team working skills and encouraged them to identify with different aspects of their native culture. While it is normal for students to interact during playtime, interaction during class-time is relatively scarce, and since team working is such an important skill in the modern work-place, we feel that this would make our collaborative videogames a valuable addition to the children's schooling.

It is also important that the games worked well with children working together while using information technology. We believe that this type of working can move children away from the idea of computing as a solitary anti-social activity and towards a more positive outlook where computers are seen as a tool for helping rather that replacing human interaction. The exercise also prepares the students for new computing paradigms such as ubiquitous computing and augmented reality that are likely to involve concurrent collaborative working in the future.

Acknowledgements

The authors of this paper would like to acknowledge the help of the following persons without whom we could not have been able to realise the experiments described in this paper. Firstly, we would like to thank the teachers, students and the parents of students at Escuela Primaria Rural Benito Juárez in Acatlima Oaxaca for participating in and facilitating our experiments and tests. We would also like to thank Carlos Martinez and Mario Moreno for helping us to use the excellent UsaLab facility at the Universidad Tecnológica de la Mixteca, and José Aníbal Arias Aguilar, Ana Delia Olvera Cervantes, Ariadna Benítez Saucedo, Jessica Santos and Mario Alberto Cortes for help during the experiments. Also at the UTM, we would like to thank Rodolfo Palma Guzman and his team at de Taller de Metal Mécanica for the design and construction of the adjustable large screen display stands used in the experiments. Finally, this article would not have been possible without the generous funding of Proyecto Conacyt 152 008 teorías, métodos y modelos de la complejidad social.

References

[1] I. Inegi, "Conteo de población y vivienda 2005," Indicadores del censo general de Población y vivienda, Ed. INEGI, México, 2005.

[2] C. N. d. E. d. l. P. d. D. Social, "Informe Ejecutivo de Pobreza: México 2007," CONEVAL México, 2007.
[3] L. Sørensen, "Report on Education in Oaxaca-The social conflict between the Mexican government and the Teachers Union," 2006.
[4] A. M. A. Juárez, "Migración y pobreza en Oaxaca," El cotidiano, vol. 148, pp. 85, 2008.
[5] D. R. Pioquinto, "Migración y cambios socioeconómicos en la comunidad de Zoogocho, Oaxaca," Estudios Demográficos y Urbanos, pp. 313-345, 1991.
[6] J. P. Schmal, "Oaxaca: A land of diversity". An educational project of the Houston Institute for Culture, 2006; http://www.houstonculture.org/mexico/oaxaca.html; retrieved April 21, 2014.
[7] N. de Bengoechea Olguin, "10+1 ≠1+10 o de cómo los indios cuentan mejor que los otros," La Vasija, vol. 1, pp. 81-90, 1998.
[8] L. Ionescu, "Mexico's pervasive culture of corruption," Economics, Management, and Financial Markets, pp. 182-187.
[9] S. D. Morris and J. L. Klesner, "Corruption and trust: Theoretical considerations and evidence from Mexico," Comparative Political Studies, vol. 43, pp. 1258-1285, 2010.
[10] C. D. Lee and P. Smagorinsky, Vygotskian perspectives on literacy research: Constructing meaning through collaborative inquiry: Cambridge University Press, 2000.
[11] D. W. Johnson and R. T. Johnson, "An educational psychology success story: Social interdependence theory and cooperative learning," Educational researcher, vol. 38, pp. 365-379, 2009.
[12] V. A. Fomichov and O. S. Fomichova, "Cognitonics as a New Science and Its Significance for Informatics and Information Society", Informatica. An International Journal of Computing and Informatics (Slovenia), Vol. 30, pp. 387-398.
[13] V. A. Fomichov and O. S. Fomichova, "An Imperative of a Poorly Recognized Existential Risk: Early Socialization of Smart Young Generation in Information Society," Informatica. An International Journal of Computing and Informatics (Slovenia), 2014, Vol. 38, No. 1, p. 59-70.
[14] O. Fomichova and V. Fomichov, "Cognitonics as an Answer to the Challenge of Time" , Proceedings of the 12th International Multiconference Information Society - IS 2009, Slovenia, Ljubljana, 12 – 16 October 2009. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute, 2009, pp. 431-434; available online at http://is.ijs.si/is/is2009/zborniki.asp?lang=eng; retrieved 27.04.2014
[15] T. B. Kane, "David and Leviathan: Forming Cognitive Tunnels between Classrooms and Artificial People in the Real World," in Gams, M., Piltaver, R., Mladenic, D. et al., Eds. (2013). Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute, pp. 453-459; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 15.04.2014
[16] A. Labus and M. Miljković, "Self-evaluation of history lessons and some related aspects of citizenship education," in Gams, M., Piltaver, R., Mladenic, D. et al., Eds. (2013). Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; pp. 443-446; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 15.04.2014
[17] T. M. Gabriela, M. M. Cristian, and D. D. Burdescu, "Building professor's mental model of student's activity in on-line educational systems," in Gams, M., Piltaver, R., Mladenic, D. et al., Eds. (2013). Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; pp. 472-475; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 15.04.2014
[18] P. Craig, N. Roa-Seïler, M. Martínez Díaz, and F. Lara Rosano, "Assessing the Potential of Colaborative Video Games to Improve Education in La Mixteca Region of Mexico," in Gams, M., Piltaver, R., Mladenic, D. et al., Eds. (2013). Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; pp. 413-446.; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; retrieved 15.04.2014
[19] P. Craig, N. Roa-Seïler, F. L. Rosano, and M. M. Díaz, "The Role of Embodied Conversational Agents in Collaborative face to face Computer Supported Learning Games," in Proc. 26th International Conference on System Research, Informatics & Cybernetics. Baden-Baden, Germany, 2013.
[20] P. Craig, N. Roa-Seïler, M. M. Díaz, and F. L. Rosano, "Evaluating the Case for Computer Supported Face to Face Collaborative Learning to Suppliment Traditional Primary Learning in the Mexican State of Oaxaca," in INTED. Valencia, Spain, 2014.
[21] J. Batista and G. Rumble, "Educación a distancia en América Latina: análisis de costo-efectividad," Documento técnico del Instituto de Desarrollo Económico del Banco Mundial (disponible en: http://www.wds.worldbank.org/servlet/WDS_IBank_Servlet, 1992.
[22] L. G. Aretio, "La educación a distancia," De la Teoría a la Práctica. Barcelona, Editorial Ariel, 2001.
[23] G. Vega, "La educación continua a distancia en México: transformaciones y retos," Revista de la Educación Superior, vol. 34, pp. 133, 2005.
[24] M. Herrera and J. Díaz, "Descripción especifica del estudio evaluativo acerca de la gestión y uso óptimo

de los recursos de la red escolar Fe y Alegría," Documento inédito, Caracas, Unesco-Cendes-Cice, 1991.
[25] A. B. Heldt, Cien años en la educación de México: Ed. Pax-México, 1972.
[26] J. G. Sánchez, "La falacia de la ampliación de la cobertura educativa mediante la utilización de las NTIC y la educación a distancia en la educación superior en México," Revista iberoamericana de educación, pp. 123-140, 2007.
[27] A. M. P. Hernández, A. E. Huerta, and F. J. P. Ochoa, "Enciclomedia. Un programa a debate," Revista Mexicana de Investigación Educativa, vol. 11, pp. 209-224, 2006.
[28] N. Cohen, "Microsoft Encarta dies after long battle with Wikipedia," The New York Times, vol. 30, 2009.
[29] F. Díaz Barriga, "en López Portillo, SEPIENSA, México [sepiensa.org.mx]," Disponible en, 1998.
[30] R. Ramirez-Velarde, D. Dolan, and J. Perez-Cazares, "Strategies for Sustainable E-Learning Projects," Technological Advances in Interactive Collaborative Learning, pp. 203, 2012.
[31] B. Einhorn, "Intel inside the third world," BusinessWeek Online, vol. 9, 2007.
[32] "Kids on Computers. Setting up computer labs for kids worldwide", 2014; http://www.kidsoncomputers.org/; retrieved April 27, 2014..
[33] M. Warschauer and M. Ames, "Can One Laptop per Child Save the World's Poor?," Journal of International Affairs, vol. 64, 2010.
[34] T. Roger and D. W. Johnson, "An overview of cooperative learning", Originally published in: J. Thousand, A. Villa and A. Nevin (Eds), Creativity and Collaborative Learning; Brookes Press, Baltimore, 1994.
[35] http://teachers.henrico.k12.va.us/staffdev/mcdonald_j/downloads/21st/comm/BenefitsOfCL/OverviewOfCoopLrng_Benefits.html; retrieved April 17, 2014.
[36] D. W. Johnson and R. T. Johnson, Learning together and alone: Cooperative, competitive, and individualistic learning: Prentice-Hall, Inc, 1987.
[37] R. Paradise, "El conocimiento cultural en el salón de clases: Niños indígenas y su orientación hacia la observación," Infancia y aprendizaje, pp. 73-86, 1991.
[38] S. U. Philips, The Invisible Culture: Communication in Classroom and Community on the Warm Springs Reservation: ERIC, 1983.
[39] V. A. Fomichov and O. S. Fomichova, "A Contribution of Cognitonics to Secure Living in Information Society," Informatica: An International Journal of Computing and Informatics (Slovenia), vol. 36, No. 2, pp. 121-130, 2012.; www.informatica.si/vol36.htm#No2.

A Chaotic Charged System Search Approach for Data Clustering

Y. Kumar and G. Sahoo
Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India
E-mail: yugalkumar.14@gmail.com, gsaho@bitmesra.ac.in

Keywords: chaos theory, clustering, Coulomb law, charge particles, Gauss Law, Newton Law

Received: February 19, 2014

Data clustering is a key technique in the field of data mining, pattern recognition, bioinformatics and machine learning which concerns the organization and unexplored relationship between the huge amounts of data. It can analyse the data without knowing the size and distribution. Thus, this paper presents a new approach based on the charged system search (CSS) with chaotic map for partition clustering. The aim of this method is to achieve the global optimum solution by minimizing an objective function. The chaotic charge system search algorithm (CCSAA) utilizes the concept of CSS algorithm and chaos theory to obtain the desired results. The quality of the proposed algorithm is evaluated on seven datasets and then compared with other well-known algorithms in data mining domain. From the simulations results, it is observed that the proposed algorithm delivers more efficient and effective results than the other methods being compared.

Povzetek: Razvita je nova metoda gručenja s pomočjo kaotičnega učenja.

1 Introduction

The aim of clustering is to discover a subset of items in a given dataset which are more alike than others using similarity measures. The various authors have applied different criteria or similarity measures to identify the items in the clusters. But, the Euclidean distances is widely accepted similarity measure for clustering problems. Cluster analysis has proven its significance in many areas such as pattern recognition [47, 49], image processing [41], process monitoring [45], machine learning [1], quantitative structure activity relationship [8], document retrieval [17], bioinformatics [15], image segmentation [34] and many more. Due to wide area of clustering in diverse domains, a large number of algorithms have been developed by various researchers and applied successfully for clustering task. Generally, clustering algorithms can be classified into two groups- Hierarchical clustering algorithms and Partition based clustering algorithms [4, 5, 26, 30]. In hierarchical algorithms, a tree structure of data is formed by merging or splitting data based on some similarity criterion. In partition based algorithms, clustering is achieved by relocating data between clusters using some clustering criterion i.e. Euclidean distance. From the literature, it has been found that partition based algorithms are more efficient and popular than hierarchical algorithms [18]. The most popular and commonly used partition based algorithm is K-means algorithm [29]. It is easy, fast, and simple to implement. In addition to it, also have one more characteristic that is linear time complexity [10, 19]. In K-means algorithm, a dataset is divided into k number of predefined clusters and the clustering objective is to minimize the intra-cluster distance [18]. Nonetheless, this algorithm has some limitations such as the results of k-means algorithm is highly dependent on the initial cluster centers and also get stuck in local minima [20, 38]. Thus, in order to overcome the pitfalls of the K-means algorithm, several heuristic algorithms have been developed. K-harmonic mean algorithm was also proposed for clustering task instead of K-Means [48]. A simulated annealing (SA) based approach has been developed in [39]. A tabu search (TS) based method was introduced in [2, 42]. A genetic algorithm (GA) based methods was presented in [6, 16, 27, 31, 32]. A comparison has been performed to investigate the computational performance of K-Means, SA, TS and GA for data clustering [3]. Fathian et al., in 2007 have developed a clustering algorithm based on honey-bee mating optimization (HBMO) [9]. An ant colony optimization (ACO) based approach for clustering is proposed in [40]. Particle swarm optimization (PSO) is applied for clustering in [46]. Whereas Hatamlou et al. employed a big bang-big crunch algorithm for data clustering in [12]. Karaboga and Ozturk presented a novel clustering approach based on artificial bee colony (ABC) algorithm in [21]. A data clustering based on gravitational search algorithm was presented in [13, 14]. For the effective clustering, a teacher learning based optimization method is applied in [35, 36]. But every algorithm has some drawbacks, for example, K-Means algorithm sucks in local optima, convergence is highly dependent on initial positions in case of genetic algorithm, in ACO the solution vector has been affected as the number of iterations increased etc.

So,the aim of this research work is to investigate the capability of chaotic charged system search (CCSS) algorithm for data clustering. This algorithm is the combination of the chaotic map and charge system search algorithm. The chaotic maps are required for chaotic search. Several chaotic maps have been reported in the literature such as the logistic map, tent map and so on. These maps are used to produce a sequence of numbers

to substitute the random parameters r_i, $rand$, $rand_1$ and $rand_2$ of the CSS algorithm. Thus, in this paper a logistic map is considered for chaotic search with charge system search algorithm (CSS) and called it chaotic charge system search algorithm (CCSSA). On the other side, the CSS algorithm is the latest meta-heuristic optimization technique developed in [22]. This technique is built on three principles: Coulomb law, Gauss law and Newton second law of motion. Every meta-heuristic algorithm contains two unique features i.e. exploration and exploitation. The exploration is referred to generate promising searching space while the exploitation can be defined as determination of the most promising solution set. Thus, in CCSS, the exploration process is carried out using Coulomb and Gauss laws while Newton second law of motion is applied to perform exploitation process. The performance of the proposed algorithm has been tested on two artificial datasets and several real datasets from the UCI repository and compared with some existing algorithms in which the quality of the solution is improved using CCSS algorithm.

2 Background

2.1 Charge system search (CSS) algorithm

CSS is the latest meta-heuristic algorithm developed by the Kaveh et al. based on movement of charged particles in D-dimensional search space [22]. The position of the i^{th} charged particle in search space is described as $X_i = (X_{i,1}, X_{i,2}, X_{i,3}, \ldots\ldots\ldots X_{i,D})$. The velocity of the i^{th} charged particle is represented as $V_i = (V_{i,1}, V_{i,2}, V_{i,3}, \ldots\ldots V_{i,D})$. It is suggested that the position and velocities of the charged particle should be in the range of $[X_{min}, X_{max}]^D$ and $[V_{min}, V_{max}]^D$ respectively. Initially, it is assumed that the velocities of all charged particles are set to 0. The main steps of the CSS algorithm are as follows.

Begin

- Randomly initialize charge particle in D-dimensional search space using equation 1.

$$X_{i,j} = X_{min,i} + r_i * (X_{i,max} - X_{i,min}) \quad (1)$$

Here, $X_{i,j}$ represents the initial value of the i^{th} variable for the j^{th} CP; $X_{min,i}$ and $X_{i,max}$ are the minimum and the maximum values for the i^{th} variable; r_i is a random number in the interval [0,1] and the initial velocities of charged particles are zero.

$$V_k = 0, \qquad k = 1, 2, \ldots . K$$

While (The stopping criterion is not met)

- Evaluate fitness of each charged particle, compare the values of fitness function and sort them into increasing order.
- Store the positions of initial charged particles (C_k) into a variable, called CM.

For n = 1 to number of charged particles

- Compute the value of moving probability for each charged particle using the following equation.

$$p_{ik} = \begin{cases} 1 & \frac{fit(i) - fit(best)}{fit(j) - fit(i)} \\ 0 & otherwise \end{cases} \quad (2)$$

If, $\frac{fit(i)-fit(best)}{fit(j)-fit(i)} > rand\ V\ fit(j) > fit(i)$,

Here, fit (i) represents the fitness of i^{th} instance of dataset while fit (best) represents the best fitness value and fit (j) represents fitness value of j^{th} data instance of the dataset and rand is a random number in between [0, 1].

- Determine the value of actual electric force (F) using the equation 3.

$$F_j = q_j \sum_{i,i \neq j} \left(\frac{q_i}{a^3} * i_1 + \frac{q_i}{r_{ij}^2} * i_2 \right) * p_{ij} * (X_i - X_j),$$

$$\begin{cases} j = 1, 2, 3, \ldots . K \\ i_1 = 1, i_2 = 0 \ \leftrightarrow \ r_{ij} < a \\ i_1 = 0, i_2 = 1 \ \leftrightarrow \ r_{ij} \geq a \end{cases} \quad (3)$$

Here, q_i and qj represents the fitness of i^{th} and j^{th} CP, $r_{i,j}$ represents the separation distance between i^{th} and j^{th} CPs, i_1 and i_2 are the two variables whose values are either 0 or 1, 'a' represents the radius of CPs and it is assumed that each CPs has uniform volume charge density but changes in every iteration. The value of q_i, $r_{i,j}$ and 'a' are evaluated using the equation 12, 13 and 14.

End for

For d = 1 to number of dimensions of charged particles

- Update the positions and velocities of charged particles using equation (4) and (5) respectively.

$$X_{j,new} = rand_1 * Z_a * \frac{F_j}{m_j} * \Delta t^2 + rand_2 * Z_v * V_{j\,old} * \Delta t + X_{j\,old} \quad (4)$$

$$V_{j\,new} = \frac{X_{j\,new} - X_{j\,old}}{\Delta t} \quad (5)$$

where, $rand_1$ and $rand_2$ are the two random functions whose values lie in between [0, 1], Z_a and Z_v are the control parameters which control the influence of actual electric force and previous velocities, m_j is the mass of j^{th} CPs which is equal to the q_k and Δt represents the time step which is set to 1.

- If charged particles cross its boundary limit then correct its position using HS based method.

End for

- Determine the fitness function of new generated charged particles and compare it to the charged particles stored into CM.
- Exclude the worst charged particles from CM.

End while

- Obtain the desired results

End

In short span of time, the CSS algorithm is applied in various research domain which shows its potency [22, 23, 24, 25, 44, 28].

2.2 Chaos theory

Recently, chaos has fascinated researchers and academicians from all over the world from different fields of sciences like chaos control, synchronization, pattern recognition, optimization theory and many others. Chaotic optimization algorithm (COA) can be defined overall as the one which uses chaotic variables in place of random variables [37]. From the extended literature survey, it came to revelation that chaos has unpredictable irregular behavior in the long term.

Whereas, nature inspired algorithms viz. ABC [21], CSS [22], GSA [13], SA [39], GA [16] and so on, as their name indicates are motivated by the biological living systems, swarms characteristics and natural phenomena's where order (regular and self-organized) replaces the disorder (irregular and unorganized) [43]. But, the chaos and nature inspired optimization algorithms shares two basic characteristics like self-organization and evolution, so, amalgamation of these two may enhance the performance of the optimization algorithm by replacing the random number with the chaotic variable. In 1976, May brought the concept of the logistic map, which is popular one among the chaotic maps [33]. It appears in nonlinear dynamics of biological population evidencing its chaotic behavior. The logistic map is a well-known polynomial chaotic map which can be described as:

$$cm_{a+1} = \alpha \times cm_a(1 - cm_a) \quad (6)$$

Where cm_a is the a^{th} chaotic number and "a" denotes the current iteration number. Clearly, $cm_a \in (0, 1)$, it is simple to indicate that if $0 < \alpha \leq 4$ then the interval (0, 1) is mapped into itself i.e. $cm_0 \in (0, 1)$, then $cm_a \in (0, 1)$. Mathematically, it is already established that when α=4, then all the values generated by the chaotic map will lie in the range of 0 -1 except that $cm_0 \notin$ (0, 0.25, 0.50, 0.75, 1). In this paper, α=4 has been used to perform the experiment.

3 Proposed CCSS algorithm for clustering

This section describes the working of CCSS algorithm for clustering task. A chaotic (logistic) map is applied to tune the CSS parameters in three different ways to enhance the performance of the CSS algorithm as well as global search especially for clustering problems. These modifications are described as:

CCSSA-1: The quality of the clustering algorithms depends on the initial cluster centers. In CSS algorithm, the initial positions of charged particles (CPs) are determined using the equation 1 which contains a random parameter (r_i). The Parameter (r_i) of equation 1 is updated using the chaotic map (cm) during iterations and the modified equation is rewritten as:

$$C_k = X_{min,i} + cm_a * (X_{i,max} - X_{i,min}),$$
$$where\ i = 1, 2 \ldots n\ and\ k = 1,2 \ldots K \quad (7)$$

In the standard CSS algorithm, r_i presents a random number in between 0 and 1. In CCSSA-1, the r_i is replaced by the chaotic number (cm_i) whose values also lies in between 0 and 1,C_k represents the number of cluster centers, $X_{min,\ i}$ and $X_{max,\ i}$ represent the minimum and maximum value of the i^{th} attribute in the dataset respectively; K is the total number of cluster centers in the given dataset.

CCSSA-2: The selected chaotic map is introduced in equation 2 which is used to compute the value of the moving probability of particles towards the charged particle. The modified equation can be given as follows.

$$p_{ik} = \begin{cases} 1 & \frac{fit(i) - fit(best)}{fit(k) - fit(i)} \\ 0 & otherwise \end{cases} \quad (8)$$

$$\text{If, } \frac{fit(i) - fit(best)}{fit(k) - fit(i)} > cm_a\ V\ fit(k) > fit(i),$$

In the standard CSS algorithm, $rand_i$ is the random number between 0 and 1. In CCSSA-2, it is replaced by chaotic number cm_i which ranges in between 0 and 1.

CCSSA-3: The parameters $rand_1$ and $rand_2$ of the position update equation 4 is altered using selected chaotic map and the modified equation can be given as.

$$C_{k,new} = cm_{a,1} * Z_a * \frac{F_k}{m_k} * \Delta t^2 + cm_{a,2} * Z_v * V_{k\ old} * \Delta t + C_{k\ old} \quad (9)$$

In the standard CSS algorithm, $rand_1$ and $rand_2$ present the random number in between 0 and 1. The chaotic map $cm_{a,1}$ and $cm_{a,2}$ are used instead of $rand_1$ and $rand_2$.

3.1 Algorithm details

In this section, CCSS algorithm is explained to solve the clustering problem. The aim of this algorithm is to find the optimal cluster points to assign N numbers of items to K cluster centers in R^n. Euclidean distance is taken as the objective function for clustering problem and items are assigned to a cluster center with minimum Euclidean distance.

The algorithm starts with defining the initial positions and velocities of K number of charged particles (CPs). The initial positions of CPs are defined in random manner. Here, CPs are assumed to be a charged sphere of radius 'a' and its initial velocity is set to zero. Thus, the algorithm starts with randomly defined initial cluster centers and ended with optimal cluster centers. Consider Table 1 which illustrates a dataset used to explain the working of CCSS algorithm for clustering with N=10, n=4 and the number of cluster centers (K) is 3. To obtain the optimal cluster centers, the CPs uses resultant electric force (attracting force vector), mass and moving probability of particles and cluster centers. After the first iteration, the velocities of CPs are determined and locations of CPs are also moved. The objective function is calculated again using the new positions of CPs and also compared with the old CPs positions that are stored in the memory pool, called as charged memory (CM). The CM is now updated with the new positions of CPs and excludes the worst CPs from CM. As the algorithm grows, the positions of CPs are updated along with the

content in CM is also updated. This process progresses until the maximum iteration or no better position of CPs will be generated.

As described earlier, the algorithm starts with the identification of the initial positions and velocities of CPs in random fashion. Thus, the equation 7 is used to initialize the initial positions of randomly defined CPs (cluster centers) and the initial positions of CPs are given in Table 2. The initial velocities of these CPs are set to zero. The CPs are described as charged spheres that contain some mass. Thus, the mass of each CP is calculated using the following equation.

$$m_k = \frac{fit(k) - fit(worst)}{fit(best) - fit(worst)} \quad (10)$$

Where, fit (k) represents the fitness of k^{th} instance of dataset while fit (best) represents the best fitness value and fit (worst) represents worst fitness value of dataset. The mass of initial positioned CPs are 1.2576, 1.714 and 0.94176.

Table 1: Illustrate a dataset to explain the CCSS algorithm for clustering with N=10, n=4 and K=3.

N	*n*			
	1	*2*	3	4
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1

Table 2: shows the position of initial CPs using equation.

K	*i*			
	1	*2*	3	4
1	4.459	3.281	1.5889	0.12854
2	5.0672	3.1964	1.5394	0.14556
3	4.8364	2.9127	1.3916	0.1791

7.

Euclidean distance is used as objective function to find the closeness of particles to CPs and assigned the particles to CPs with minimum objective value. Table 3 provides the value of objective function of initial positioned CPs for our example dataset. Euclidean distance can be given as.

$$d_{i,k} = \sum_{k=1}^{K} \sum_{j=1}^{N} \sum_{i=1}^{n} \sqrt{\left\| X_{j,i} - C_{k,i} \right\|^2} \quad (11)$$

Table 3: Normalized value of objective function.

N	*K*		
	1	*2*	*3*
1	0.70682	0.34008	0.64416
2	0.56056	0.29821	0.11036
3	0.39144	0.44174	0.33163
4	0.25624	0.48175	0.32117
5	0.65969	0.43567	0.70688
6	1.1639	0.83447	1.1985
7	0.31484	0.55047	0.55501
8	0.56552	0.22469	0.52577
9	0.43529	0.74525	0.43714
10	0.48574	0.20218	0.23904

The information contained in the given string is used to arrange the items into different clusters which are given below.

2	3	3	1	2	2	1	2	1	2

From the above string, it is observed that the first, fifth, sixth, eighth and tenth particles belong to the cluster 2^{nd}; second and third and tenth particles belong to cluster 3^{rd} ; fourth, seventh and ninth particles belong to cluster 1^{st} . Hence at this step, the dataset is divided into three different clusters and store the values of positions of CPs in a new variable called charge memory (CM) which can be used to memorize the positions of CPs. Later on, these CPs positions will be used for comparisons with newly generated CPs positions and the best positions are included in the CM and excluded the worst positions from CM. Here, the size of CM is equal to the N/4. The main work of CM is to keep track of the number of good positions of CPs which are obtained during the execution of CSS algorithm and after the execution of algorithm; the optimal number of CPs positions (i.e. K number of CPs) are determined using minimized objective function values. The above discussion relates to the initialization of the proposed algorithm for clustering problem.

As the study of various meta-heuristic algorithms, it is found that every meta-heuristic algorithm contains two approaches i.e. exploration and exploitation in which one approach initiated local search while the other approach carried out the global search. The local search tends to the exploration of random search space such that the most promising solution space can be obtained while the global search refers to the exploitation of good solution vectors from the promising solution space. Hence in case of CCSS algorithm, the local search i.e. exploration is initiated using Coulomb and Gauss laws while the global search i.e. exploitation is performed by Newton second law of motion. The local search of CCSS algorithm starts by measuring the electric force ($E_{i,k}$) generated by CP. The electric force ($E_{i,k}$) generates at a point is either inside the CPs or outside CPs. So, this direction of force is described as moving probability ($P_{i,k}$) of CPs. While

the Coulomb and Gauss laws are applied to measure the total electric force generated on a CP , called it actual electric force (F_k). The moving probability $P_{i,k}$ for each CPs can be determined using the equation 8. The value of moving probability ($P_{i,k}$) is either 0 or 1 and it gives the information about the movement of CPs. Table 4 shows the moving probability ($P_{i,k}$) values for each particles to each cluster centers.

The Coulomb and Gauss laws are employed to determine the value of actual electric force (F_k) generated on CPs. The Coulomb law is used to calculate the force outside the CP and Gauss law is used to calculate the force inside the CP. The equation 3 is used to determine the actual electric force (F_k).

$$q_i = \frac{fit(i) - fit(worst)}{fit(best) - fit(worst)}, i = 1, 2, \dots N \quad (12)$$

$$r_{ik} = \frac{\|X_i - X_k\|}{\|(X_i + X_k)/2 - X_{best}\| + \epsilon} \quad (13)$$

$$a = 0.10 * max\{x_{i,max} - x_{i,min} | \, i = 1, 2 \dots n\} \quad (14)$$

Table 4: Moving probability P_{ik} of each CPs with each items of dataset.

N	*K*		
	1	*2*	*3*
1	0	0	0
2	0	1	0
3	1	1	0
4	1	1	0
5	0	0	0
6	0	0	0
7	0	1	0
8	0	0	0
9	1	0	1
10	0	1	0

Table 5: values of magnitude of charge (q_i) of each CPs and separation distance ($r_{i,k}$).

N	q_i	*Separation Distance ($r_{i,k}$)*		
		K		
		1	*2*	*3*
1	1.8833	1.991	0.61468	2.4197
2	1.1833	0.2648	1.3881	1.7242
3	0.93333	0.80968	1.0094	1.352
4	1.2333	0.45338	1.4435	1.1904
5	1.8833	2.0875	0.59338	2.2831
6	4	3.1051	1.9508	3.3476
7	1.6167	1.3312	0.55961	1.3142
8	1.9333	1.8999	0.42761	2.3886
9	0.58333	4.1443	4.1107	3.9615
10	1.2	0.80915	0.78601	1.9357

Tables 5,6 and 7 provide the values of q_i, $r_{i,k}$, i_1, i_2, "a" and (X_i-C_k) variables which are used to calculate the actual electric force (F_k), applied on the CPs and the value of 'a' is 0.4459, 0.9526 2 and 1.4363 which is calculated using equation 14. Thus, the values of the actual electric force (F_k) generated on initial CPs (cluster centers) is 0.32832, 1.1605 and 0.20691.The values of electric force (F_k) are used with Newton second law of motion to determine the new position of CPs and velocities of CPs.

The Newton second law of motion is employed to get the new positions and velocities of CPs. This is referred to as exploitation of solution vectors from the random space search.

Table 6: values of I_1 and I_2.

N	*K (I1)*			*K (I2)*		
	1	*2*	*3*	*1*	*2*	*3*
1	1	0	1	0	1	0
2	0	0	1	1	1	0
3	0	0	0	1	1	1
4	0	1	0	1	0	1
5	1	0	1	0	1	0
6	1	1	1	0	0	0
7	0	0	0	1	1	1
8	1	0	1	0	1	0
9	1	1	1	0	0	0
10	0	0	1	1	1	0

Table 7: values of X_i-C_k for each cluster center K.

N	*K*		
	1	*2*	*3*
1	0.74245	0.25143	0.88023
2	0.042448	0.44857	0.18023
3	0.057552	0.54857	0.080228
4	0.057552	0.54857	0.080228
5	0.74245	0.25143	0.88023
6	1.9424	1.4514	2.0802
7	0.24245	0.24857	0.38023
8	0.64245	0.15143	0.78023
9	0.55755	1.0486	0.41977
10	0.14245	0.34857	0.28023

Thus, the new positions of CPs and velocities are obtained from equation 9 and 5 respectively. Z_a and Z_v act as the control parameters which are used to control the exploration and exploitation process of the algorithm. These parameters also affect the values of previous velocities and actual resultant force generated on a CP. These values may be either increased or decreased. Thus, Z_a is the control parameter belongs to the actual electric force F_k and controls the exploitation process of CSS algorithm. The large value of Z_a increases the

convergence speed of the algorithm while small value increases the computational time of the algorithm. Z_v is the control parameter for the exploration process and acts with the velocities of CPs. Here, it is noted that Z_a is the increased function parameter while Z_v is the decreased function parameter. Table 8 provides the new positions of CPs which are evaluated using CCSS algorithm. The new positions of CPs are mentioned in Table 8 and the values of control parameters Z_a and Z_v are determined using the equation 14. The new velocities ($V_{k,new}$) of each CPs is 0.21599, 0.46049 and 0.10449.

$$Z_a = (1 - iteration/iteration\ max),$$
$$Z_v = (1 + iteration/iteration\ max) \quad (14)$$

Table 8: New position of CPs.

K	i			
	1	2	3	4
1	4.513	3.335	1.6429	0.18253
2	5.1823	3.3115	1.6545	0.26068
3	4.8625	2.9388	1.4177	0.20523

Hence from the above discussion, the process of the algorithm can be categorized into three sections: Initialization, Search and Termination condition. The initialization section deals with the CCSS algorithm parameters; positions and velocities of the initial CPs; determine the value of objective function and rank them; store the positions of CPs into CM. In the search section, the new positions and velocities of CPs are determined using moving probability ($P_{i,k}$) and actual electric force (F_k). The value of objective function is evaluated using newly generated CPs, compared with previous CPs; rank them and store the best CPs in CM. The termination condition of the algorithm is either maximum number of iterations or repeated positions of CPs.

3.2 Pseudo code of CCSS algorithm for clustering

Step 1: Load the dataset and specify the number of initial cluster centers (K).

Step 2: * initialize the initial positions and velocities of charged particles (CPs) * /
For each charged particles k=1 to K
For each j=1 to m
Determine the value of initial position of CPs (C_k) using equation 7;
Calculate the value of mass for each C_k using equation 10;
End for
End for
V_k=0; * velocities of each CP*/
Iteration =0;

Step 3: Evaluate the value of objective function using eq. 11 and assign the items to the clusters with minimum objective function values.

Step 4: Store the positions of initial charged particles (C_k)
into a variable, called CM.

Step 5: while (the termination conditions are not met), do
\ * Calculate the value of moving probability $P_{i,k}$ for each charged particle C_k * /
For each charged particles k=1 to K
For each i=1 to n
Determine the value of fitness of each instance (q_{ik}) with each CP (C_k) using eq. 8;
End for
If (fit (q_{ik}) > fit (k))
$P_{ik} \rightarrow 1$;
Else
$P_{ik} \rightarrow 0$;
End if
End for

Step 6: \ * Determine the value of actual electric force (F_k)*/
Determine the value of mass for each instance q_i using eq.12;
Calculate the value of radius 'a' using eq. 14;
For each charged particles k=1 to K
Calculate the value of separation distance ($r_{i,k}$) using eq. 13;
If ($r_{i,k} < a$)
$i_1 \rightarrow 1$
Else
$i_2 \rightarrow 0$
End if
If ($r_{i,k} \geq a$)
$i_2 \rightarrow 1$
Else
$i_1 \rightarrow 0$
End if
End for
For each charged particles k=1 to K
Determine the value of actual electric force (F_k) using eq. 3;
End for

Step 7: Calculate the new positions and velocities of CPs using eq. 9 and 5 respectively;

Step 8: Compute the value of objective function using new positions of CPs;

Step 9: Compare the value of objective function of newly generated CPs to CPs reside in CM;

Step 10: Memorize the best solution achieved so far
Iteration= Iteration +1;
End while

Step 11: Output the best solution obtained.

4 Experimental results

This section describes the results of the CCSS algorithm for data clustering problem. To assess the performance of CCSS algorithm, it is applied on ten datasets. These datasets are ART1, ART2, iris, wine, CMC, glass, breast cancer wisconsin, Liver disease (LD), thyroid and vowel in which iris, wine, CMC, glass. Liver disease (LD), thyroid, vowel and breast cancer wisconsin datasets are real that are downloaded from the UCI repository while rest of two datasets are artificial i.e., ART1 and ART2. The characteristics of these datasets are discussed in Table 9. The proposed algorithm is implemented in

Matlab 2010a environment on a core i5 processor with 4 GB using window operating system. For every dataset, the algorithm is run 100 times individually to check the effectiveness of the proposed algorithm using randomly generated cluster centers. The parameter settings for CCSS algorithm are mentioned in Table 10. The sum of intra cluster distance and f-measure parameters are used to evaluate the quality of solutions for clustering algorithm. The sum of intra cluster distances can be defined as distances between the instances placed in a cluster to the corresponding cluster center. The results are measured in terms of best case, average cases, worst case solutions and standard deviation. The quality of clustering is directly related to the minimum of the sum of distances. The accuracy of clustering is measured using f-measure. To ensure the effectiveness and adaptability of CCSS algorithm in clustering domain, the investigational results of CCSS algorithm are compared with K-Means, GA, PSO, ACO and CSS algorithms which are given in Table 11.

4.1 Datasets

4.1.1 ART1

It is two dimensional artificial dataset, generated in matlab to authenticate the proposed algorithm. This dataset includes 300 instances with the two attributes and three classes. Classes in dataset are disseminated using μ and λ where μ is the mean vector and λ is the variance matrix. The data has generated using μ1 = [3, 1], μ2 = [0, 3], μ3 = [1.5, 2.5] and λ1 = [0.3, 0.5], λ2 = [0.7, 0.4], λ3 = [0.4, 0.6]. The Figure 1(a) depicts the distribution of data into ART1 and figure 1(b) shows the clustering of same data using CSS method.

4.1.2 ART2

It is three dimensional artificial data which includes 300 instances with three attributes and three classes. The data has generated using μ1 = [10, 25, 12], μ2 = [11, 20, 15], μ3 = [14, 15, 18] and λ1 = [3.4, -0.5, -1.5], λ2 = [-0.5, 3.2, 0.8], λ3 = [-1.5, 0.1, 1.8]. The Figure 2(a) represents the distribution of data in ART2 dataset and Figure 2(b) shows the clustering of same data using CCSS method.

4.1.3 Iris Dataset

Iris dataset contains three variety of the iris flower which is setosa, versicolour and virginica. The dataset contains 150 instances with three classes and four attributes in which each class contains of 50 instances. The attributes of iris dataset are sepal length, sepal width, petal length, and petal width.

4.1.4 Wine Dataset

It contains the chemical analysis of wine in the same region of Italy using three different cultivators. This dataset contains 178 instances with thirteen attributes and three classes. The attributes of dataset are alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and proline.

4.1.5 Glass

This dataset consists of six different types of glass information. The dataset contains 214 instances and 7 classes. It contains nine attributes which are refractive index, sodium, magnesium, aluminum, silicon, potassium, calcium, barium, and iron.

4.1.6 Breast Cancer Wisconsin

This dataset characterizes the behavior of cell nuclei present in the image of breast mass. It contains 683 instances with 2 classes i.e. malignant and benign and 9 attributes. The attributes are clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. Malignant class consists of 444 instances while benign consists of 239 instances.

4.1.7 Contraceptive Method Choice

It is a subset of National Indonesia Contraceptive Prevalence Survey data that had been performed in 1987. This dataset contains information about married women who were either pregnant (but did not know about pregnancy) or not pregnant. It contains 1473 instances and three classes i.e., no use, long term method and short term method. Each class contains 629, 334 and 510 instances respectively. It has nine attributes which are Age, Wife's education, Husband's education, Number of children ever born, Wife's religion, Wife's now working, Husband's occupation, Standard-of-living index and Media exposure.

4.1.8 Thyroid

This dataset contains the information about the thyroid diseases and classifies the patient into three classes-normal, hypothyroidism and hyperthyroidism. The dataset consists of 215 instances with five features. The features are the medical tests which have been used to categorize the patients. The features are Tresin, Thyroxin, Triiodothyronine, Thyroidstimulating and TSH value.

4.1.9 Liver Disorder

This dataset is collected by BUPA medical research company. It consists of 345 instances with six features and two classes. The features of the LD dataset are mcv,alkphos, sgpt, sgot, gammagt and drinks.

4.1.10 Vowel

This dataset consists of 871data instances of Indian Telugu vowel sounds with three features which correspond to the first, second, and third vowel frequencies and six classes.

4.2 Performance measures

4.2.1 Sum of intra cluster distances

It is sum of distances between the data instances present in one cluster to its corresponding cluster center. Minimum the sum of intra cluster distance indicates the better the quality of the solution. The results are measured in terms of best, average and worst solutions.

Table 9: Characteristics of datasets.

Dataset	*Classes*	*Features*	*Total instances*	*Instance in each classes*
ART 1	3	2	300	(100, 100, 100)
ART 2	3	3	300	(100, 100, 100)
Iris	3	4	150	(50, 50, 50)
Glass	6	9	214	(70,17, 76, 13, 9, 29)
LD	2	6	345	(145, 200)
Thyroid	3	3	215	(150, 30, 35)
Cancer	2	9	683	(444, 239)
CMC	3	9	1473	(629,334, 510)
Vowel	6	3	871	(72, 89, 172, 151, 207, 180)
Wine	3	13	178	(59, 71, 48)

4.2.2 Standard Deviation

This parameter gives the information about the dispersion of data present in the cluster from its cluster center. The minimum value of standard deviation indicates that the data are close to its center while a large value indicates the data are far from its center points.

4.2.3 F-Measure

F-measure is calculated by the recall and precision of an information retrieval system [7, 11]. It is weighted harmonic mean of recall and precision. To determine the value of F –measure, every cluster describes a result of the query while every class describes as a set of credentials for the query. Thus, if each cluster j consists a set of n_j data instances as a result of a query and each class i consists of a set of n_i data instances require for a query then n_{ij} gives the number of instances of class i within cluster j. The recall and precision, for each cluster j and class i is defined as

$$Recall\ (r(i,j)) = \frac{n_{i,j}}{n_i} \text{ and}$$

$$Precision\ (p(i,j)) = \frac{n_{i,j}}{n_j} \qquad (13)$$

The value of F-measure (F (i, j)) is computed as

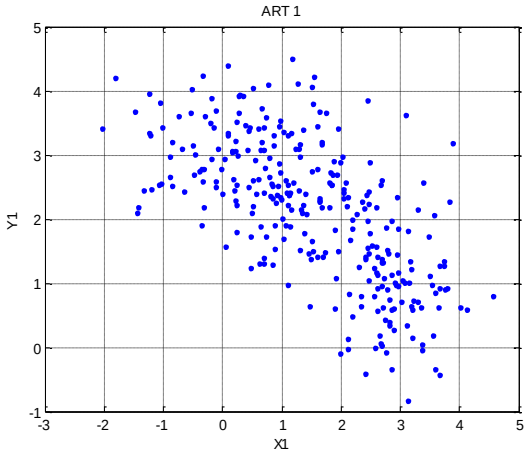


Figure 1(a): Distribution of data in ART1 dataset.

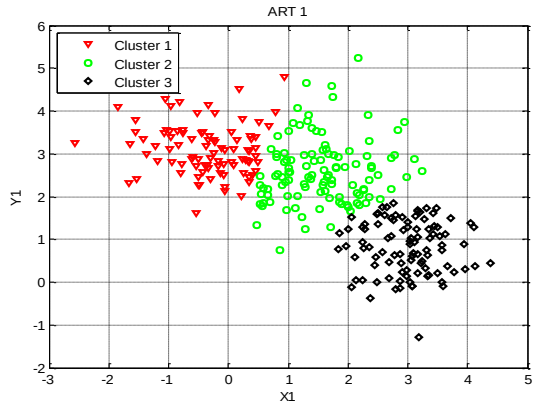


Figure 1(b): Clustered the ART1 dataset using CCSS.

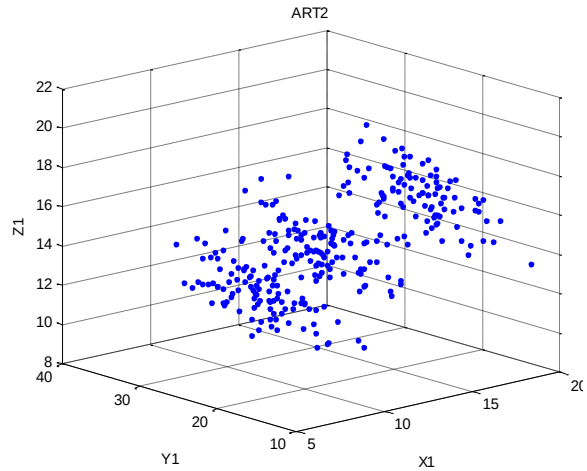


Figure 2 (a): Distribution of data in ART2 dataset.

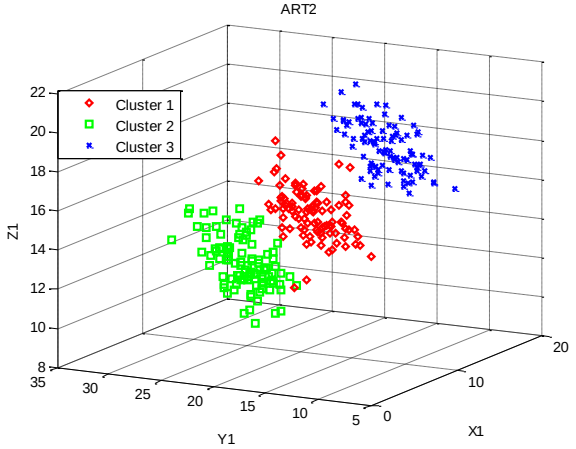


Figure 2(b): Clustered the ART2 data using CCSS.

$$F(i,j) = \frac{2 * (Recall * Precision)}{(Recall + Precision)} \quad (14)$$

Finally, the value of F-measure for a given clustering algorithm which consist of n number of data instances is given as

$$F(i,j) = \sum_{i=1}^{n} \frac{n_i}{n} . \max_i(F(i,j)) \quad (15)$$

Table 10: Parameters setting for CCSS algorithm

Parameters	Value
No. of CPs	No. of Clusters
c	0.1
∈	0.001

Table 11: Comparisons of different clustering algorithm with CCSS algorithm.

Dataset	*Parameters*	*K-means*	*GA*	*PSO*	*ACO*	*CSS*	*CCSS*
ART 1	Best	157.12	154.46	154.06	154.37	153.91	152.69
	Average	161.12	158.87	158.24	158.52	158.29	157.32
	Worst	166.08	164.08	161.83	162.52	161.32	159.86
	Std.	0.34	0.281	0	0	0	0
	F-Measure	99.14	99.78	100	100	100	100
ART2	Best	743	741.71	740.29	739.81	738.96	737.91
	Average	749.83	747.67	745.78	746.01	745.61	745.36
	Worst	754.28	753.93	749.52	749.97	749. 66	748.78
	Std.	0.516	0.356	0.237	0.206	0.209	0.182
	F-Measure	98.94	99.17	99.26	99.19	99.43	99.54
Iris	Best	97.33	113.98	96.89	97.1	96.47	96.38
	Average	106.05	125.19	97.23	97.17	96.63	96.52
	Worst	120.45	139.77	97.89	97.8	96.78	96.69
	Std.	14.631	14.563	0.347	0.367	0.14	0.11
	F-Measure	0.782	0.778	0.782	0.779	0.787	0.791
Wine	Best	16555.68	16530.53	16345.96	16530.53	16282.12	16183.94
	Average	18061	16530.53	16417.47	16530.53	16289.42	16218.42
	Worst	18563.12	16530.53	16562.31	16530.53	16317.67	16268.73
	Std.	793.213	0	85.497	0	10.31	43.16
	F-Measure	0.521	0.515	0.518	0.519	0.529	0.535
LD	Best	11397.83	532.48	209.15	224.76	207.09	205.98
	Average	11673.12	543.69	224.47	235.16	228.27	223.87
	Worst	12043.12	563.26	239.11	256.44	242.14	238.84
	Std.	667.56	41.78	29.38	17.46	18.54	13.29
	F-Measure	0.467	0.482	0.493	0.487	0.491	0.494
Cancer	Best	2999.19	2999.32	2973.5	2970.49	2946.48	2937.56
	Average	3251.21	3249.46	3050.04	3046.06	2961.16	2954.41
	Worst	3521.59	3427.43	3318.88	3242.01	3006.14	2978.34
	Std.	251.14	229.734	110.801	90.5	12.23	19.23
	F-Measure	0.829	0.819	0.819	0.821	0.847	0.856
CMC	Best	5842.2	5705.63	5700.98	5701.92	5672.46	5663.86
	Average	5893.6	5756.59	5820.96	5819.13	5687.82	5692.23
	Worst	5934.43	5812.64	5923.24	5912.43	5723.63	5727.18
	Std.	47.16	50.369	46.959	45.634	21.43	24.76

	F-Measure	0.334	0.324	0.331	0.328	0.359	0.367
Thyroid	Best	13956.83	10176.29	10108.56	10085.82	9997.25	9981.48
	Average	14 133.14	10218.82	10149.7	10108.13	10078.23	10054.32
	Worst	146424.21	10254.39	10172.86	10134.82	10116.52	10102.74
	Std.	246.06	32.64	27.13	21.34	49.02	47.28
	F-Measure	0.731	0.763	0.778	0.783	0.789	0.792
Glass	Best	215.74	278.37	270.57	269.72	203.58	241.08
	Average	235.5	282.32	275.71	273.46	223.44	256.45
	Worst	255.38	286.77	283.52	280.08	241.27	265.29
	Std.	12.47	4.138	4.55	3.584	13.29	8.38
	F-Measure	0.431	0.333	0.359	0.364	0.446	0.453
Vowel	Best	149422.26	149513.73	148976.01	149395.6	149335.61	149326.74
	Average	159242.89	159153.49	151999.82	159458.14	159128.19	159012.39
	Worst	161236.81	165991.65	158121.18	165939.82	164537.08	164286.97
	Std.	916	3105.544	2881.346	3485.381	3128.023	3316.58
	F-Measure	0.652	0.647	0.648	0.649	0.649	0.656

Table 12: Cluster center generated using CCSS method for ART1 and ART2 dataset.

Dataset	Center1	Center2
ART1	0.7036	3.1243
	1.8296	2.4327
	2.9504	0.9212

Dataset	Center1	Center2	Center3
ART2	10.8915	20.0376	15.7924
	14.9639	14.9693	17.8362
	9.2168	24.9836	12.9167

Table 13: Cluster center of Iris, Wine and CMC dataset using CCSS algorithm.

Dataset	Center1	Center2	Center3
Iris	5.01476	5.92624	6.75143
	3.36827	2.73618	3.06485
	1.46534	4.43953	5.63867
	0.24169	1.41694	2.12363
CMC	24.96782	33.57356	43.75427
	3.04023	3.15648	2.86549
	3.87451	3.56795	3.47264
	1.76986	3.65193	4.60256
	0.96258	0.78947	0.79624
	0.79371	0.69826	0.77486
	2.30346	2.10543	1.82644
	2.97296	3.29489	3.43257
	0.03826	0.05987	0.09278

Dataset	Center1	Center2	Center3
Wine	13.94261	12.62849	12.85965
	1.86138	2.34257	2.573642
	2.43564	2.33089	2.38526
	18.06218	20.95385	19.86478
	106.47693	98.69536	92.87932
	2.87896	2.14357	2.08431
	3.24652	1.78643	1.47298
	0.27138	0.43738	0.42846
	2.01853	1.44086	1.41791
	5.63729	4.34576	5.79516
	1.07814	0.95393	0.86541
	3.02437	2.47285	2.26429
	1138.0563	465.3951	688.0637

Table 14: Cluster center of Glass dataset using CCSS algorithm.

Dataset	Center1	Center2	Center3	Center4	Center5	Center6
	1.55643	1.53278	1.52736	1.51864	1.51961	1.52875
	12.98963	13.16383	13.09752	14.9486	14.10576	15.86713
	3.46978	0.25842	3.53789	0.06579	2.43568	3.73632
	1.03678	1.41572	1.34384	2.23528	2.63459	2.98763
Glass	72.01356	72.9382	72.59836	73.6853	71.26418	74.45752
	0.27906	0.31646	0.57812	0.05127	2.59642	5.1922
	9.46573	12.01694	8.3824	8.7243	6.02568	14.63761
	0.03548	0.05182	0.003929	1.03263	1.3356	2.6271
	0.05593	0.05673	0.05904	0.01861	0.01268	0.43456

Table 15: Cluster center of Cancer dataset using CCSS algorithm.

Dataset	Cancer								
Center1	2.89264	1.12476	1.21938	1.16893	1.99784	1.12856	2.06953	1.15432	1.07583
Center2	7.12836	6.64249	6.62812	5.61532	5.27268	8.23549	6.12637	6.10794	2.37681

Table 16: Cluster center of Thyroid dataset using CCSS algorithm.

Dataset	Thyroid				
Center1	0.9386	-1.1452	-0.4298	0.9786	4.1683
Center2	1.6832	-1.6678	-1.0207	2.3346	0.8634
Center3	-0.1723	0.1591	0.2596	-0.3010	-0.2206

Table 17: Cluster center of Vowel dataset using CCSS algorithm.

Dataset	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6
	506.47598	407.6284	624.21692	356.82783	376.54276	437.90257
Vowel	1839.1372	1016.2586	1308.9836	2291.01353	2153.6867	990.76184
	2555.97546	2314.362	2332.9563	2976.8596	2676.4656	2661.53974

Table 18: Cluster center of LD dataset using CCSS algorithm.

Dataset	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6
LD	87.92621	69.86842	25.85635	22.04346	27.10362	2.89347
	91.23956	75.06531	59.18264	38.92736	129.86542	5.96174

From Table 11, it can be seen that the results obtained from the CCSS algorithm are better as compared to the other algorithms. The best values achieved by the algorithm for iris, wine, cancer, CMC, glass, LD, thyroid and vowel datasets are 96.38, 16183.94, 2937.56, 5663.86, 241.08, 205.98, 9981.48 and 149326.74. The CCSS algorithm gives better results with most of the datasets. From the simulation results, it is also observed that CCSS algorithm achieve minimum value to the best distance parameter for the LD dataset and worst distance parameter for vowel dataset among all methods being compared. The standard deviation parameter shows how much the data are far from the cluster centers. The value of standard deviation parameter for CCSS algorithm is also smaller than other methods. Moreover, the CCSS algorithm provides better f-measure values than others which show higher accuracy of the said algorithm. To prove the viability of the results given in Table 11, the best centers obtained by the CCSS algorithm are given in Tables 12–18.

5 Conclusion

In this paper, a chaotic charged system search algorithm is applied to solve the clustering problem. In the

proposed algorithm, Newton second law of motion is used to get the optimal cluster centers but it is the actual electric force (Fk) and chaotic map which plays a vital role to obtain the optimal cluster centers. Hence, the working of the proposed algorithm is divided into two steps. First step involves the tuning of the CSS parameters using chaotic map. In the second step, the optimal cluster centers using Newton second law of motion are obtained. The CCSS algorithm can be applied for data clustering when the number of cluster centers (K) is already known. The performance of the CCSS algorithm is tested on the several datasets and compared with other algorithms; in which proposed algorithm provides better results and the quality of solutions obtained by the proposed algorithm is found to be superior in comparison to the other algorithms.

References

[1] Alpaydin, E. (2004) *Introduction to machine learning*. MIT press.

[2] Al-Sultan, K.S. (1995). A Tabu search approach to the clustering problem. *Pattern Recognition*, vol. 28, pp. 1443–1451.

[3] Al-Sultana, Khaled S., and Khan M. M. (1996). Computational experience on four algorithms for the hard clustering problem. *Pattern Recognition Letters*, vol. 17, no. 3, pp 295-308.

[4] Barbakh, W., Wu, Y., Fyfe, C. (2009). *Review of Clustering Algorithms*. Springer, Berlin/ Heidelberg, pp. 7–28.

[5] Camastra, F., Vinciarelli, A. (2008). Clustering Methods. *Machine Learning for Audio, Image and Video Analysis. Springer*, London, pp. 117–148.

[6] Cowgill, M.C., Harvey, R.J., Watson, L.T. (1999). A genetic algorithm approach to cluster analysis. *Comput. Math. Appl*., vol. 37, pp. 99–108.

[7] Dalli, A. (2003). Adaptation of the F-measure to cluster based lexicon quality evaluation, Association for Computational Linguistics, *In Proceedings of the EACL*, pp. 51-56.

[8] Dunn III, W. J., Greenberg, M. J., and Soledad S. C. (1976). Use of cluster analysis in the development of structure-activity relations for antitumor triazenes. *Journal of medicinal chemistry*, vol. 19, no. 11, pp. 1299-1301.

[9] Fathian, M., Amiri, B., Maroosi, A. (2007). Application of honey-bee mating optimization algorithm on clustering. *Appl. Math. Comput*., vol. 190, pp. 1502–1513.

[10] Forgy, E.W. (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, vol. 21, no. 2, pp. 768 - 769.

[11] Handl, J., Knowles, J., & Dorigo, M. (2003). On the performance of ant-based clustering. Design and Application of Hybrid Intelligent System. *Frontiers in Artificial Intelligence and Applications*, vol. 104, pp 204–213.

[12] Hatamlou, A., Abdullah, S., Hatamlou, M. (2011a). Data clustering using big bang-big crunch algorithm. *In the proceeding of innovative computing technology*, Springer Berlin Heidelberg publisher, Tehran, Iran pp.383–388.

[13] Hatamlou, A., Abdullah, S., Nezamabadi-pour, H. (2011b). Application of Gravitational Search Algorithm on Data Clustering. In the proceeding of the 6th international conference (RKST 2011), Banff, Canada, Springer Berlin Heidelberg publisher, pp. 337–346.

[14] Hatamlou, A., Abdullah, S., Nezamabadi-pour, H. (2012). A combined approach for clustering based on K-means and gravitational search algorithms. *Swarm and Evolutionary Computation*, vol. 6, pp. 47-52.

[15] He, Yi, Pan, W and Jizhen L. (2006). Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. *Computational statistics & data analysis*, vol. 51, no. 2, pp. 641-658.

[16] Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press.

[17] Guobiao, Hu, Shuigeng Z., Jihong G., and Xiaohua Hu. (2008). Towards effective document clustering: A constrained< i> K</i>-means based approach. *Information Processing & Management*, vol. 44, no. 4, pp. 1397-1409.

[18] Jain, A.K., Murty, M.N., Flynn, P.J. (1999). Data clustering: A Review. *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323.

[19] Jain, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letter*, vol. 31, pp. 651–666.

[20] Kao, Y.-T., Zahara, E., Kao, I.W. (2008). A hybridized approach to data clustering. *Expert Systems Appl*., vol. 34, pp. 1754–1762.

[21] Karaboga, D., and Ozturk C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computing*, vol. 11, no. 1, pp. 652-657.

[22] Kaveh, A., and Talatahari, S. (2010 a). A novel heuristic optimization method: charged system search. *Acta Mechanica*, vol. 213, no. 3-4, pp. 267-289.

[23] Kaveh, A., and Talatahari, S. (2010b). Optimal design of skeletal structures via the charged system search algorithm. *Structural and Multidisciplinary Optimization*, vol. 41, no. 6, pp. 893-911.

[24] Kaveh, A., and Laknejadi, K. (2011a). A novel hybrid charge system search and particle swarm optimization method for multi-objective optimization. *Expert Systems with Applications*, vol. 38, no. 12, pp. 15475-15488.

[25] Kaveh, A., and Talatahari S. (2011b). An enhanced charged system search for configuration optimization using the concept of fields of forces. *Structural and Multidisciplinary Optimization*, vol. 43, no. 3, pp. 339-351.

[26] Kogan, J., Nicholas, C., Teboulle, M., Berkhin, P. (2006). A Survey of Clustering Data Mining

Techniques- *Grouping Multidimensional Data*. Springer Berlin Heidelberg, pp. 25–71.

[27] Krishna, K., Murty, MN. (1999). Genetic K-means algorithm. *IEEE Transactions on Systems, Man, Cybernet. Part B: Cybernet*. Vol. 29, pp. 433–439.

[28] Kumar, Y, and Sahoo G. (2014). A charged system search approach for data clustering. *Progress in Artificial Intelligence*, vol. 2, no. 2-3, pp. 153-166.

[29] MacQueen, James. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 281-297, p. 14.

[30] Maimon, O., Rokach, L. (2010). *A survey of Clustering Algorithms, Data Mining and Knowledge Discovery Handbook*. Springer, pp. 269–298.

[31] Maulik, U., Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition*, vol. 33, pp. 1455–1465.

[32] Murthy, C.A., Chowdhury, N. (1996). In search of optimal clusters using genetic algorithms. *Pattern Recognition Letter*, vol. 17, pp. 825–832.

[33] May RM. (1976). Simple mathematical models with very complicated dynamics. *Nature*, vol. 261, pp. 459–474.

[34] Pappas, Thrasyvoulos N. (1992). An adaptive clustering algorithm for image segmentation. *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 901-914.

[35] Satapathy, S. C., & Naik, A., (2011) Data clustering based on teaching-learning-based optimization. *In the proceeding of the second international conference (SMECOO),* Springer Berlin Heidelberg publisher, *Vishakhapatnam, Andhra Pradesh, India* pp. 148-156, 2011.

[36] Sahoo, A. J., & Kumar, Y. (2014). Modified Teacher Learning Based Optimization Method for Data Clustering. *In Advances in Signal Processing and Intelligent Recognition Systems*, Springer Berlin Heidelberg publisher, Kerala, India, pp. 429-437.

[37] Schuster, H. G., & Just, W. (2005) *Deterministic chaos: An introduction.* Wenham: Wiley-VCH Verlag.

[38] Selim, S.Z., Ismail, M.A. (1984). K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 81–87.

[39] Selim, S.Z., Alsultan, K. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, vol. 24, pp. 1003–1008.

[40] Shelokar, P. S., Jayaraman, V. K., and Kulkarni, B. D., (2004). An ant colony approach for clustering. *Analytica Chimica Acta*, vol. 509, no. 2, pp. 187-195.

[41] Sonka, Milan, Vaclav Hlavac, and Roger Boyle. (1998). *Image processing, analysis, and machine vision*. Champion and Hall.

[42] Sung, C.S., Jin, H.W. (2000). A tabu-search-based heuristic for clustering. *Pattern Recognition*, vol 33, pp. 849–858.

[43] Talatahari S., Farahmand A.B., Sheikholeslami R., and Gandomi AH. (2012) Imperialist competitive algorithm combined with chaos for global optimization. *Commun Nonlinear Sci Numer Simul.*, vol 17, pp. 1312–1319.

[44] Talatahari, S., Kaveh, A. and Sheikholeslami, R. (2011). An efficient charged system search using chaos for global optimization problems. *Int J Optim Civil Eng*., vol. 1, no. 2, pp. 305-25.

[45] Teppola, P., Mujunen, S-P and Minkkinen, P. (1999). Adaptive Fuzzy C-Means clustering in process monitoring. *Chemometrics and intelligent laboratory systems*, vol. 45, no. 1, pp. 23-38.

[46] Van der Merwe, D. W., and Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization." *In Congress on Evolutionary Computation CEC-03*, vol. 1, pp. 215-220.

[47] Webb, A. (2002). *Statistical pattern recognition*. New Jersey: John Wiley & Sons, pp. 361–406.

[48] Zhang, B., Meichun H., and Dayal, U., (1999). K-harmonic means-a data clustering algorithm, Hewlett-Packard Labs *Technical Report* HPL.

[49] Zhou, H., and Yonghuai L. (2008). Accurate integration of multi-view range images using k-means clustering. *Pattern Recognition*, Vol. 41, No. 1. Pp. 152-175.

Using Cognitive Tunnels in a New Approach to Building Social Elevators in the Information Society

Thomas B. Kane
Dept of Management, Work and Organisation, Stirling Management School
University of Stirling, UK FK9 4LA
E-mail: t.b.kane@stir.ac.uk

Keywords: telepresence, Turing test, cognitonics, education, artificial intelligence

Received: June 25, 2014

We are on the verge of developing artificial intelligences that may dwarf the capabilities of human intelligence. How we will interact and thrive alongside such intelligence will be a pressing societal problem. This paper addresses the question of the intelligence of "artificial persons" (organisations of people) that already exist in our modern world. A particular issue – social exclusion from the artificial persons of top professions – is explored. The paper shows how cognitonics, using telepresence, cognitive tunnelling and an advanced dialogic framework, based on the viva voce form of the Turing test, can support society in opening up pathways to its advanced professions for all of its youngsters; and at the same time train its artificial persons to work harmoniously with the other members of society. Learning how to have meaningful social dialogue with artificial persons, may be of societal value as we prepare to live among artificial persons that employ their own artificial intelligence.

Povzetek: Opisan je nov pristop v informacijski družbi s pomočjo kognitivnih tunelov.

1 Introduction

Over the last few decades, it has become possible to observe a large number of distortions in the development both of the personality and of society caused by the rapid development of information and communication technologies (ICT), globalization processes, and the preponderance of commercialized values. As a reaction to this situation, a new scientific discipline (and simultaneously a branch of the humanities) called *cognitonics* has emerged [1 - 6].

The principal aim of cognitonics is to combine the efforts of scholars working in various fields in order to find systemic solutions to compensate for the distortions and to establish preconditions for the harmonic, well-balanced development of the personality.

This paper explores an issue in the field of education. When a teenager is close to graduation from high school, he/she needs to possess well developed cognitive skills of processing information and to possess a broad mental outlook in order to understand what is his/her calling. Following Fomichova and Fomichov [7], calling could be described as the most important work the person could do, in which he/she would be most difficult to replace.

However, by the end of high school, a very considerable number of teenagers have a rather narrow mental view of the possibilities before them, and don't consider many professions where they may be able to find a calling. There are complex reasons for this, both personal and societal.

In the case of youngsters from state education systems, a contributing factor might be that societal tools of our modern cultures, the institutions of our advanced professions, are inadvertently colluding with complex education systems to exclude talented youngsters from fully developing while they are at school, thus rendering them unfit for tertiary education, or for reaching the height of their natural potential within the advanced professions. The result of which, being to contribute to social immobility throughout society.

Within Artificial Intelligence is the notion of Singularity [8], the point at which computer intelligence will finally exceed human intelligence. Ray Kurzweil, who has predicted that the moment of this achievement will be in 2045, has also predicted that in that world, it will be possible for humans to decide how long they want to live, merge with technology, and participate in unimaginable adventures of the imagination.

Although the Singularity is still some way off, we already live in a world of artificial persons - organisations of people. The artificial persons of our modern world (national authorities, local authorities, education systems, businesses, affiliations) are tools of society. They can be composed of a few, or thousands, or even millions, of people at the same time. They have their own reasons for existing, have relations with natural and artificial persons, they offer services and have codes of conduct with which their members must conform. These persons operate a functional, artificial intelligence, in our societies.

This paper presents a *dialogic* framework, an adaptation of the Turing Test[9], suitable for establishing Turing Test-like dialogue between artificial persons and

natural persons in a communications-rich world of social networking and *telepresence*. The framework sees the intelligence test lodged in the real-world, and put in the hands of stakeholders who are involved in the issues of intelligence being explored and for whom the dialogue explores contextual and time-critical issues. If the experience is recorded, then the record can be analysed by multiple stakeholders.

We are also in a philosophical era that questions whether we have wholly "brainbound" minds, or extended minds[10] which overlap between the organic being and cognitive tools in the real-world around us. Andy Clark compellingly argues that human beings are natural-born cyborgs[11]. He also notes that as mind extension technologies develop, "It is simply up to us, in these critical years, to try to guarantee that *human*-centred technology really means what it says: that human means all of us and not just the lucky few."

Artificial persons of our society are supra-cognitive tools capable of working with many natural minds for a range of functional purposes. Perhaps these artificial persons have become adept at exploiting multiple forms of intelligence, by successfully inhabiting the minds of other beings for their own intelligent, self-serving, purposes. Dealing respectfully with such artificial persons, and the difficult societal questions they raise - exploring them and their relationships with natural persons; might help us to deal with the Kurzweillian forms of intelligence that are evolving artificially. This paper focuses on probing intelligence that already exists within an artificial person, and exploring the organisational extended-mindedness that sees natural and artificial persons use many minds in pursuing their goals.

Telepresence technologies can now connect functioning parts of the world together in a way that was unimaginable even 5 years ago. The proposal in this paper is that societal functions actively seek to employ these technologies to cognitively link operational parts of society with the educational parts of society so that all pupils have a chance to develop higher language skills, feel their way into professions, network, and understand the world around them as it is. Such work will open real-world opportunities to pupils at school and could help to challenge the negative effects of social immobility in an immediate way and with intelligence.

2 Telepresence and cognitive tunnelling

Minsky [12] introduced the word ``*telepresence*" into the language. He imagined people at work physically controlling apparatus that is far away from them (perhaps it could be roving mining equipment that is present on Mars; perhaps it could be control functions that are present in a nuclear power station that is flooded with radiation) with a sense of connectedness. Today, *telepresence* is a term that is sometimes used for *immersive videoconferencing*. Types of videoconferencing that could be described as telepresence in Minsky's terms include situations where young people need to be protected from the overwhelming nature of an event, such as giving evidence at a criminal prosecution, and are allowed to give evidence from abroad. In such cases, the important part of the link is that the young person is both present and absent at the same time.

A major task of such telepresence links is to keep the contributing environments within their normal work routine, and to construct protocols of expression that allow the fullest possible means of communication. In terms of telepresence, a great deal of parallel effort, operating at different organisational levels, is employed to produce the necessary conditions. Too much of educational videoconferencing is currently showcase activity - where dialogue is made inferior to the importance of broadcast; questions are allowed in time slots at the end of a presentation; events describe issues not of the moment, but of the past; where presentation professionals rather than educationalists or the professionals who do the work are involved in the link. In order to guard against the inauthenticity of such issues we focus on creating a telepresence at events as they are happening, and with the people who are involved.

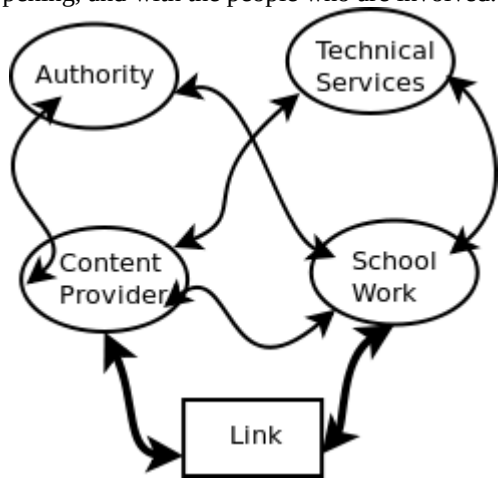


Figure 1: Producers of a Real-world Link.

Real-world educational videoconferencing falls into two types of activities: production activities, and educational activities. A successful real-world link involves a partnership between four artificial persons: educational authority, a content provider, a school group, technical services provider, and a separate programme co-ordinator (who may be a natural or artificial person).

Note that all of these organisational levels are contributory stakeholders in the link. They have different points of view on the importance of different aspects of the link. All of the contributing factors can be explored *dialogically* by persons within their own organisational silo, and by other persons who are outside of it.

In the simplest videoconferencing setup two environments are linked. At each site there is a camera, an eye, and a microphone (an ear) that can be used to focus the attention of the other, and a number of ICT items that can be used to deliver semiotic images to the other side.

A communal communication model that takes into account a variable number of people in each contributing

point, and the multiple ways of sending, receiving and contemplating the utterances of the dialogues makes it clear that the members of the school classroom form a single artificial person. If we imagine that each natural person is given a hat for each artificial person that they can represent, we see that each pupil has, for example, a school hat, a particular classroom hat, a person hat, a social background hat. Each professional in dialogue with the class has at least the following hats: a unique natural person hat, one hat for each Artificial Person that they are a member of (as many as are appropriate to the link), a personal environmental hat, etc. The task of analysis is to be aware that there are multiple natural and artificial persons being represented in each real-world link. The educational task is to navigate the pupil through the process of learning about the real-world in a way that is consistent with pupils pursuing the requirements of the school curriculum.

The communication model needs to encapsulate the artificial person that is the classroom, and the artificial person that is the real-world content provider, in such a way as to allow these to unfold into the various natural people who are participating. That model needs also to allow parallel transmission and reception of all forms of communicable message that can be captured by the telepresence equipment -- and to allow participants to acknowledge and record impressions.

Figure 3 shows two artificial persons joined in a link. Each has

- a receiving function that receives all inputs that have been transmitted,
- a collegiate group function that produces the utterance that represents the dialogic response to received signals,
- a production function, that sends a fresh set of messages back.

Many messages can be sent at the same time, and anything that can be detected and transmitted can be reckoned as part of a potential communication.

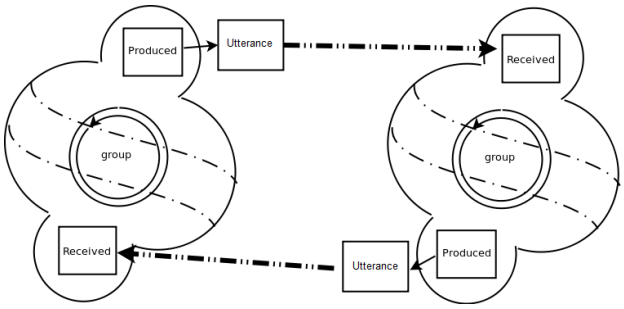


Figure 2: Communications Between Artificial Persons.

In examining all communicatable signals to be read in parallel as part of the communication, we bring a very interesting problem to Bakhtinian analysis. With appropriate questions, we make the task open to the extended Turing Test much more valuable to those who seek to understand the world of artificial entities of multiply contributed intelligences.

2.1 Cognitive tunnelling

Telepresence allows us to link pupils to a place of work where cognitive capacities are called upon, and for a while to join the classroom environment with that external environment. (For some young people the cognitive distance between their living environment and these external working spaces might as well be the distance to the moon.) The task is to focus on material that minimises the cognitive obstacles that lie between work pupils are currently capable of doing in class and the work they would be expected to perform if they were in the external place of employment.

Classroom learning areas transform to a stage upon which pupils can perform, and from which they can draw on school objects and achievements. Similarly, the working environment becomes a stage, and working materials of an expert can be brought into a link. Each stage setting is slightly removed from the working environment to satisfy the requirements of the link.

Figure 3: Art Gallery with Art-Classroom.

The challenge is to develop appropriate educational/real-world scaffolding [13]: to allow the contact to be an authentic engagement between pupils and the real-world, to remain in the pupil's mind, and to familiarise the young people with professional activities that they themselves might perform later in life as a result of their education and professional development. We describe the work in this area as "*cognitive tunnelling*".

With suitable educational scaffolding, young people can be asked to engage with the real-world as it is (and, even, as it changes), and for the real-world to engage with pupils as they are developing. A key issue in working with socially excluded pupils is in joining the environmental languages of the pupils -- their local accents, idioms, dialects - and the professional languages -- technical terms, idioms, jargon - of working spaces using Bakhtinian [14] techniques of hybridization and dialogue.

3 Anatomy of a real-world link

The city of Glasgow has linked all of its 29 state high schools into a shared videoconference service network that allows business quality links to any classroom in any high school in the city. The link we are dissecting was \ part of a programme called ``Listening to Young People". The programme engaged young people in schools throughout Glasgow with the Scottish Parliament and with the Civil Service. This school involved in this

link was Holyrood High School – one of the largest high schools in Europe.

Educational Authority -- Glasgow City Council Education Department had the vision, relationships, power and financial ability to promote the endeavour and make a link viable (within all social obligations) and interesting. This is a level of Leviathan which is below city government level.

Real-World Content Provider -- The Scottish Parliament, an artificial person, provided a real-world scenario for the schools as an educational experience. Understanding the parliament forms part of the school curriculum. The link was with a Member of the Scottish Parliament, Frank MacAveetie, who had been a teacher at Holyrood High School; and with one of the parliamentary education officers - who brought the parliamentary mace.

The link sought to explore real-world knowledge which is experiential, drawn from the content provider's daily activities and of their operational procedures, which is open, which is harmonised with classroom learning, which can afford interaction and allow pupils a real-world experience.

Figure 4: Modern Studies Class to Scottish Parliament.

Active Curricular Schoolwork—The class was a first year secondary school (aged 11) Modern Studies class, who were studying the Scottish Parliament. The published Curriculum[15] requires the following capacities:

I can investigate the features of an election and the work of representatives at a local, national or European level to begin to develop my understanding of how democracy works.

SOC 2:18a

I can evaluate the impact which decision making bodies have on the lives of people in Scotland or elsewhere

SOC 4:18a

I can debate the reasons why some people participate less than others in the electoral process and can express informed views about the importance of participating in a democracy.

SOC 4:18b

Technical Services Providers -- provider of link from where exactly in the school to where exactly in the real-world. Technical services also have to support users in how to set up camera positions and how to present themselves effectively for transmission. This area of coordination ranges across the educational Internet/communications service provider, the technical support officers at education services, schools and the external places of interest. Smooth coordination here is the technical achievement of the link. This link was between the pupils classroom and a Parliamentary committee room.

Educational Event Producer, or Programme Coordinator -- These four areas mentioned above have occasional need to be coordinated during the period of the educational activity.

Because the link is scheduled to happen at a particular time and there is a matching of the working day with the school day, there are often small problems that need to be addressed in each of the four areas, and there is always a need for communication between one area of expertise and another (e.g. local authority and real world provider). A fifth category of expertise that co-ordinates the real-world educational intervention, sympathetic to the real-world time and content constraints between classroom, service-provider, content producer and educational authority, is introduced here as the Educational Event Producer, or Programme Co-ordinator. Any tool or person involved in this kind of work needs to be concerned with the quality of the communications and needs to ensure that co-operation between parties takes place appropriately.

It may be that the creation of this space of dialogues may be a plausible example of how a Clarkian extended mind space is produced, in which cognitive activities operate dialogically and mental contents can reside inside, outwith and alongside a natural person.

3.1 The link

The link was for 30 minutes and followed the format:

1. General Welcomes -- welcome to Parliament, Frank MacAveetie; welcome to Holyrood, by deputy head, Bernie Pollock
2. Classroom teacher, Mrs Brady engages with contributors, introduces class
3. Presentation from MSP on Parliament
4. Questions from Class to MSP and Education Officer
5. Questions to class from MSP
6. Goodbyes and round up

One interchange with Mr MacAveetie, regarding homelessness solicited a professional response that crossed all the languages involved in the link -- the local, the educational, the personal, and the professional. The question regarded finding a political solution to homelessness, and Mr MacAveetie spoke of a member of his own family, a 71 year old, who was a street vagrant. And then he said:

"But it's not just 71 year olds that are on the streets. I've even seen former pupils of mine who have got heroin problems, actually begging on the streets of Glasgow and it actually breaks your heart, because those young boys in particular used to play on your school football team. They were as fit as a fiddle, and then at 16 and 17 they suddenly get into drugs, and there they are at 23, begging on the streets of Glasgow from their former teacher."

The live experience of such a link is very powerful. This videoconference was recorded, as where constructive dialogues between contributors at all levels of the cognitive tunnel. These recorded objects are available and can be constructively, subjectively examined (for example, by pupils involved, or stakeholders elsewhere), and the humanity of the contributions can be debated, with a view to making improvements. Such materials, when shared across all state schools, could make it possible for all schools to benefit from every link. Once a complete real-world programme has been co-ordinated for the first time, it can then be improved upon from within any of the contributing areas of expertise.

In any links between schools and professionals, there is a need to separate out the voices that represent the organisations from the voices of natural persons who are involved in them. In this regard, it is important to focus on how natural persons work with artificial persons.

4 Meetings with artificial persons in the information society

Thomas Hobbes [16] introduced us to Leviathan, which he described as a commonwealth, or state, "which is just an artificial man---though bigger and stronger than the natural man, for whose protection and defence it was intended", supplying analyses of the social contract natural people accede to in joining up with Leviathan, the civic responsibilities attendant upon the social contract, and how natural people within Leviathan should be governed -- by sovereignty, democracy or aristocracy. The frontispeace of Hobbes book showed a crowned monarch whose body and arms were composed of many human heads, all looking up to him.

Figure 5: Leviathan - None Like Him.

What Hobbes did not foresee was a time when organisations of people within Leviathan's body, would become lesser artificial persons, and would grow to Hobbesian commonwealths themselves, each with its own goals and social contracts. Figure 1 shows what the frontispiece might look like today.

Organisational behaviour [17] is the study of the behaviour of people involved in organisations, and the organisations themselves. There are three main areas:

- the study of individuals in organisations (micro);
- the study of work-groups in organisations (meso), and
- the study of how organisations, themselves (macro), behave.

It is worth noting that even in an Artificial Person, the intermediate level between micro and macro levels allows for the existence of sub-artificial person recursion within a single artificial person. Artificial Persons develop their own cultures and seek to thrive in the culture around them [18].

With such definitions we see Artificial Persons as complex, constructed societal objects, cognisant insofar as they borrow the cognitive abilities of members of the commonwealth. They are cognisant tools of society, dispersed across space, capable of being present in different ways, capable of defining themselves axiomatically, capable of organising themselves recursively, knowledgeable about their rights and history; and able to self-reflect in a communal way, able to take care of themselves in the present and choose policy and goals for the future.

Both Artificial Persons and Natural Persons have narrative history. Insofar as any person is part of more than one artificial person. For each of these that we slip between on a daily basis, there is its own narrative, the shared narrative between individual and group, and any sense of Leviathan relationships or injunctures. Each person in the multiple-lines of real-life storytelling, moving from language to language as they represent other artificial persons and occasionally their own self.

In searching for a definition for an artificial persons that could be a realistic player in the imitation game, this paper proposes that any person that conforms to a narrative account of personhood could be considered as one that can participate in the new Turing Test. Such a person needs to be capable of seeing itself in its own story and being able to tell its own tale. Following Bruner [19], e describe such a person as one who is:

- teleological and agentive
- sensitive to obstacles
- responsive to judgements
- capable of selective remembering
- oriented towards reference groups and significant others
- possessive and extensible
- able to shed values and possessions as required
- experientially continuous over time, despite striking transformations
- sensitive to where and with whom it finds itself
- accountable
- moody, affective, labile and situation-sensitive
- seeking and guarding coherence

Such a definition of person covers natural and artificial persons. The framework will allow any artificial person that encapsulates these traits to participate in the dialogue. Such persons can include legalised and informal groupings of persons, whether those be familial, social, or functional in any way.

A thesis of this work is to recognize that complex relationships exist between natural and artificial persons, see if artificial persons are as capable of being deceptive and self-serving as natural persons. Human intelligence in this complex world is not simply a matter of seeking consistent behavior: indeed, much intelligence might be involved in the judicious applications of non-consistent behaviour over a period of time. We seek to explore the true intelligence, or multiple intelligences of artificial persons through what they say, what they do, how they do it; and analyse the material with the forensic effort of that French[20] applied to the Turing test[9].

Today, a natural person might be a member of any number of different artificial persons at the same time, and may be torn in behavioural terms between differing social contracts. Some of these artificial persons might partner one other, some may be at odds with one another.

We can also explore extended mindedness from the point of view of the artificial person towards the use of natural persons -- that natural persons are extended mind tools of artificial persons. In seeing an artificial person as a person, constituted of many natural (and possibly some lesser artificial) persons; the artificial person is not so much "extended minded" as "many-minded". That is, its cognition is the result of composite, parallel activities, being contributed by many minds (owned by a number of other natural and artificial persons).

The framework presented here aims to allow higher and higher cognitive levels of Leviathan to examine and direct the policies of its members in acceptable forms of social behaviour.

5 Questions for artificial persons

Suitable questions should engage the artificial person as it goes about its business, just as Socrates engaged with citizens in their element while he wandered the streets of Athens. Questions could be asked that address complex societal issues that probe natural persons and artificial persons alike. For example, in social mobility.

The Organisation for Economic Co-operation and Development (OECD) has described[21] very troubling social mobility issues in the United Kingdom: e.g. more than 50% of youngsters will grow up to have the same salary as their father. The Sutton Trust [22] shows that 53% of the UK's most influential people were independently educated, including 24% of university vice-chancellors, 32% of Members of Parliament, 51% of medical consultants, 54% of top journalists, 70% of High Court judges… when only 7% of the UK population are.

Top Universities in Scotland have been criticised for the low numbers of students being taken from the most disadvantaged areas. In response to criticism of the University of St Andrews in this regard [23], the principal of the University highlighted the difficulty of finding sufficient numbers of plausible students with the appropriate grades in deprived parts of society. The worry being that underqualified students would not be capable of doing well at St Andrews. Response from a government advisor, acknowledged that there is a real problem of social inequality, and that social origins of inequality go to "differences in language and brain development".

And yet, we know that academic ability is randomly dispersed across all socio-economic groups and that advanced societies endeavour to seek out the best talents within the population to fill top professions. Runco has found that creative projects[24] which do not focus on cultural tropes are especially suited to helping talented youngsters from disadvantaged backgrounds to develop their unique creative talents. Links with telepresence are known to be immersive and focused on personal effort and creativity. Such educational work that had direct links with direct activities of the real-world could raise the consciousness of the pupils and establish a route into the professions.

Currently, there is a barrier before youngsters in deprived areas in reaching tertiary education; and there is a further barrier before fully educated qualified youngsters in reaching the top of the professions if they have not been privately educated. Such issues suggest that we may have already arrived at a Singularity-like situation in modern-life. This is a question whose scope covers the space of the societal education system as a whole; and the societal employment institutions as a whole -- both of which can be considered as Artificial Persons within the national state Leviathan, who are in some ways governable by Leviathan.

It is difficult to see which of the two hypotheses,

- collusion and exclusion between educational establishments and employers, or
- impaired cognitive development in all young people from socially deprived backgrounds

is actually the worst. Both hypotheses, whether true or false, certainly merit investigation. What is clear is that an engaged society could undertake steps to rectify any difficulties. The discipline of Cognitonics could be of assistance in this matter.

Schools, particularly those who are involved in educating youngsters from disadvantaged backgrounds, deserve an answer to the questions; "is there something that can be done to help our pupils reach the height of their potential in tertiary education"; and "is there something that can be done to help our pupils rise to the top of the professions when they have completed their education?". Schools could, as part of their curricular activities, seek out answers to such questions. Using telepresence and the dialogic framework schools can question persons, natural and artificial, while they are at work -- so that pupils can see how the world is, in real-time, so that they can learn from it.

With such questions, we have left Turing's parlour game and have started to explore the streets of Athens.

6 Supersizing the Turing test

Alan Turing's [9] test of machine intelligence derives from an imitation game, a parlour game, whereby a man (player A) and a woman (player B) are separated away from an interrogator (player C), and each are asked questions in turn, by means of a note delivered to each, by a messenger who acts on behalf of player C. The man is pretending to be a woman; and the woman is playing to have the truth be known. The fun of the parlour game for each player s to prevail. For player B it is to use truthful statements more effectively than player A is able to use misleading statements; and for player A, it is to overcome the truths of Player B, with a more compelling falseness.

In the imitation game, no-one questions the existence of intelligence in any of the participants; the fun is in watching how well intelligence is applied. In his paper, Turing says "The game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether some one really understands something or has learnt it parrot fashion.'" This version of the game, the basis of the proposed framework, involves an interrogator who examines a player specifically to uncover the completeness of their knowledge of "something".

Turing did not stipulate that the interrogator should be a singular, natural person. Consequently, the interrogator could be a group of individuals working together to perform the interrogation function. Thus, the interrogator may be an artificial person, made up, for example, of a group of guests at a party. Player A could be an artificial person (a group of people working together) in the same way. The dialogic framework allows a version of the Turing Test that supports dialogues between persons (either artificial or natural) as interrogator and player.

Using the *viva voce* approach means player A and player C can be as face-to-face as it is possible to be. And yet, they are one-step removed from the world of their activities - they are in a constructed place, derivative of the true work-space, and they enter a dialogue.

In these real-world meetings, between parties across many social classes who employ many individual languages, and produce new hybridisations of language, the depth of dialogic analysis of such meetings can be guided by the novelistic analysis techniques introduced by Mikhail Bakhtin [25 - 26].

Especially, this means that any semiotic capable of being recorded and transmitted by the telepresence equipment can be interpreted as an utterance; any utterance should be viewed through a chronotopic lens which respects the primacy of context over text; any action of any person can be seen as an act of dialogue between that person and the world [14]; and any professional or social language used for a specific purpose by a specific group can be seen as a professional "speech genre" [27] and can therefore be explored in relation to all other languages with the help of literary analysis.

The case for rigorous forms of meetings that respect such Bakhtinian approach to chronotope, utterance, intonation, dialogue and speech genres is made. If recordings of such links are made, in-depth analysis of the links can be made by many artificial persons, at many levels within Leviathan; and results can be pooled to aid in the development of policy to improve matters that are found to be wanting. This work would be in the cognitonic realm of the humanities.

7 Cognitonics, testing humanity, exploring reality

Topics for *Viva Voce* style Turing-tests that could be applied to the production of a link include:

A) Test of Educational Authority: Did the idea serve a valuable purpose? Was it worth the effort? Was it a diversion?
B) Test of Real-World Provider: Did the external party produce material suited to the pupils and the curriculum? Was this an authentic link with the real-world scenario, or could their real-world contact be shown to be self-congratulatory, cursory, inauthentic or unenlightening?
C) Test of Schoolwork Activity: Is the activity involved in this link related to curricular requirements and suitable to the age and development of the pupils involved? Can recording be used in class? Can learners reflect appropriately, or annotate the experience relevantly?
D) Test of Technical Services: Does the link respect the real-world context of all parties involved?
E) Test of Educational Impact: What impact had the school on the real-world? What token of reality passed from the real-world to the school?

Larger societal questions can be articulated and addressed. For example:

A) Harmony: How smoothly did contributors work together? Are there any societal issues that need to be addressed? Are there policies of artificial persons that need to be altered?
B) Collective intentionality: are particular artificial persons welcoming to all, or partial to a few? Do artificial persons use natural persons as shields?
C) Social practices and agreements between artificial and natural persons: are we tolerating inappropriate behaviours between artificial persons that is to the detriment of society?
D) Understanding: How do we construct the interpersonal dialogues that Socrates might have with the beings of our world where he here today - tailored so that we can understand the world as it is, and the world can know and understand itself?

E) Sustainability: How do we promote good relationships between natural and artificial persons within Leviathan?

Cognitonics [1 - 3, 28] enables us to recognise exclusion and societal under-achievement, reason it out, and then to use tools of ICT itself, to tackle it.

8 Conclusion

Understanding complex societal issues of the real-world as it daily goes about its business, has always proven itself to be a difficult task. A major issue has been in separating out the voices which represent artificial persons and the voices which represent natural persons. Today, communications technologies working alongside cognitive tools embedded within Artificial Intelligence make it possible for schools to ask probing questions of natural and artificial persons in the real world, and to examine how natural people move from artificial person to artificial person in society.

The paper suggests that a type of cognitive telepresence can be achieved by means of cognitive tunnelling, a term that is introduced to describe bringing pupils right to the working methods of natural and artificial people at work. Cognitive tunnelling presents a way of extending Vygotsky's scaffolding technique to places of live activity that are remote from the school.

The paper suggests that Bakhtin's extraordinary analysis tools for working on language, narrative and dialogue can help us to articulate and then to overcome cognitive hurdles presented to talented youngsters from disadvantaged communities. If such encounters are recorded and re-used, they can form material for Turing-Test-like, multi-level, multiply judged subjective assessments of the intelligence(s) within organisations, which could guide organisations in developing respectful relationships with natural persons. The ultimate aim of such work being to make all young people aware of the possibilities in their world and to compete for employment opportunities that present themselves at all levels of society.

Acknowledgement

The author wishes to express grateful acknowledgement, for many fruitful discussions regarding the ideas in this paper with participants of the Third International Conference on Cognitonics, in particular, with its Co-Chairs, Vladimir A. Fomichov and Olga S. Fomichova.

References

[1] V. A. Fomichov and O. S. Fomichova, 'Cognitonics as a New Science and Its Significance for Informatics and Information Society', in *Informatica (Slovenia)*, 2006, Vol. 30, No. 4, pp. 387–398.

[2] V. A. Fomichov and O. S. Fomichova, 'A Contribution of Cognitonics to Secure Living in Information Society', in *Informatica (Slovenia)*, 2012, Vol. 36, No. 2, pp. 121–130.

[3] V. A. Fomichov and O. S. Fomichova, 'An Imperative of a Poorly Recognized Existential Risk: Early Socialization of Smart Young Generation in Information Society', in *Informatica (Slovenia)*, 2014, Vol. 38, No. 1, pp. 59-70.

[4] M. Bohanec, M. Gams, and V. Rajkovic, 'Proceedings of the 12th International Multiconference Information Society – IS 2009, *Slovenia, Ljubljana, 12 – 16 October 2009. The Conference Kognitonika/Cognitonics*. Jozef Stefan Institute', *http://is.ijs.si/is/is2009/zborniki.asp?lang=eng*, pp. 427–470, 2009.

[5] M. Bohanec, M. Gams, and D. Mladenec, 'Proceedings of the 14th International Multiconference Information Society – IS 2011, *Slovenia, Ljubljana, 10 – 14 October 2011. The Conference Kognitonika/Cognitonics*. Jozef Stefan Institute', *http://is.ijs.si/is/is2011/zborniki.asp?lang=eng*, pp. 347–430, 2011.

[6] M. Gams, R. Piltaver, and D. Mladenec, 'Proceedings of the 16th International Multiconference Information Society – IS 2013, *Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics*. Jozef Stefan Institute', *Jozef Stefan Inst.*, http://is.ijs.si/is/is2013/zborniki.asp?lang=eng, pp. 403–482, 2013.

[7] O. S. Fomichova and V. A. Fomichov, 'Cognitonics as an Answer to the Challenge of Time'; *Proceedings of the 12th International Multiconference Information Society - IS 2009, Slovenia, Ljubljana, 12 – 16 October 2009*. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute, 2009, pp. 431-434; available online at http://is.ijs.si/is/is2009/zborniki.asp?lang=eng; retrieved 27.04.2014

[8] R. Kurzweil, *The singularity is near: when humans transcend biology*. London: Gerald Duckworth, 2005.

[9] A. M. Turing, 'Computing Machinery and Intelligence', *Mind*, vol. LIX, no. 236, pp. 433–460, 1950.

[10] A. Clark, *Supersizing the mind: embodiment, action, and cognitive extension*. Oxford; New York: Oxford University Press, 2011.

[11] A. Clark, *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press, USA, 2003.

[12] M. Minsky, 'Teleprescence', *Omni*, vol. 6, Jun. 1980.

[13] H. Daniels, *An introduction to Vygotsky*. London; New York: Routledge, 2005.

[14] M. M. Bakhtin, *The dialogic imagination : four essays / by M.M. Bakhtin ; edited by Michael Holquist ; translated by Caryl Emerson and Michael Holquist.* Austin: University of Texas Press 1982 printing, 1982.

[15] Education Scotland, 'Scottish Curriculum for Excellence Outcomes'.

http://www.educationscotland.gov.uk/Images/all_experiences_outcomes_tcm4-539542.doc.

[16] T. Hobbes, *Leviathan*. Oxford: Oxford University Press, 2008.

[17] J. A. W. III and J. R. Hollenbeck, *Organizational Behavior: Securing Competitive Advantage*, 1 edition. New York: Routledge, 2009.

[18] G. Hofstede, G. J. Hofstede, and M. Minkov, *Cultures and Organizations: Software of the Mind, Third Edition*, 3 edition. New York: McGraw-Hill Professional, 2010.

[19] J. S. Bruner, *Making stories: law, literature, life*. Cambridge, Mass.; London: Harvard University Press, 2003.

[20] P. D. Turney, 'Answering Subcognitive Turing Test Questions: A Reply to French', *arXiv:cs/0212015*, Dec. 2002.

[21] OECD, 'A Family Affair: Intergenerational Social Mobility Across OECD Countries, in OECD Economic Policy Reforms, 2010'. OECD Publishing, 2010.

[22] Sutton Trust, 'Educational Backgrounds Reports'.

[23] S. Carrell, Scotl, and correspondent, 'Top Scottish university says poorer students don't make the grade', *The Guardian*, 27-Jun-2012.

[24] M. A. Runco, *Creativity: Theories and Themes: Research, Development, and Practice*, 1 edition. Amsterdam ; Boston: Academic Press Inc, 2007.

[25] M. M. Bakhtin and C. Emerson, *Problems of Dostoevsky's poetics*. Minneapolis: University of Minnesota Press, 1984.

[26] M. M. Bakhtin, *Rabelais and his world*. Bloomington: Indiana University Press, 1984.

[27] M. M. Bakhtin, M. Holquist, V. McGee, and C. Emerson, *Speech genres and other late essays*. Austin: University of Texas Press, 1986.

[28] V. A. Fomichov and O. S. Fomichova, 'The social responsibility of computer science specialists for the creative potential of the young generation', *Int. J. Artif. Intell. Educ.*, vol. 11, pp. 208–219, 2000.

Use Case of Cognitive and HCI Analysis for an E-Learning Tool

Marian Cristian Mihăescu [1], Mihaela Gabriela Ţacu [2], Dumitru Dan Burdescu [1]
[1] University of Craiova, Faculty of Automation Computers and Electronics, Romania
[2] University of Craiova, Faculty of Economics and Business Administration, Romania
E-mail: mihaescu@software.ucv.ro, mihatacu@yahoo.com, burdescu@software.ucv.ro

Keywords: e-learning, human computer interaction, cognitonics, stemming, concept detection

Received: June 25, 2014

This paper presents the development of a new tool for an e-Learning platform, with implications and analysis from human computer interaction (HCI) and cognitonics perspectives. The goal is to improve the educational process by helping the professor form a correct mental model of each student's performance (capability). Besides this, the developed application is also analyzed from HCI and cognitive perspective with an attempt to offer an effective and highly usable tool. The interplay between cognitive psychology and HCI is emphasized as a fundamental prerequisite for a constructive and argumentative design. The main functionalities offered by the developed application are: evaluation of the level of understanding of the course material by the student and analysis of the difficulty level of the questions proposed by the professor for an exam. The prerequisites for accomplishing the task are a good structure of the on-line educational environment and information about students' activities on the platform. An important part of the process of obtaining the desired result is performing a text analysis and concept extraction on the professor's uploaded courses. The languages supported by this module are both English and Romanian. Using this module, the professor will have the ability to control the difficulty of test and exam questions and make changes accordingly: add, delete or modify test/exam questions. The developed application is thereafter discussed from HCI and cognitive psychology perspectives, such that further analysis and improvements are ready to be achieved. Our approach creates a context in which continuous design, implementation and evaluation has as output a high quality user interface suitable for an Intelligent Tutoring System.

Povzetek: Opisana je nova platforma za učenje, ki omogoča uporabo kognitivnih in HCI analiz.

1 Introduction

During the last years, the interaction between professors and students in on-line educational environments has been considerably improved, especially by developing new tools and implementing different functionalities that integrate intelligent data analysis techniques. An area that still needs further work is the cognitive area, particularly towards helping the professors build more accurate mental models of each student's capabilities.

In regular educational environment, a professor can achieve that mental model by continuously interacting with students and observing their learning skills and capabilities. Online, it is harder to accomplish that, because of the lack of constant and valuable analysis of feedback that is offered by students. That is why the approach to building a professor's mental model of student's activity becomes a tool that can improve an on-line educational environment.

The main purpose of the tool is to help the professor understand and analyze student's activity without having any face-to-face activity.

Besides the design and development of the tool that actually implements the needed functionality, this paper also presents a detailed analysis from human computer interaction (HCI) and cognitive perspective.

The research of HCI and cognitive psychology issues are cornerstone in shifting gears from "technology that solves problems" towards "design that emphasizes the user's needs". These general research areas have a great impact on the field of e-learning due to the wide range media that can produce cognitive affection at various industries. Among the most common options there are simple text, voice, picture, video, or virtual reality.

The final goal of the tool is to extend its usability with respect to the particularities of e-learning environments. That is why the general fundamental usability evaluation formulated by Nielsen [25] needs a proper specific adjustment for e-learning environments. The most important characteristics of a usable e-learning environment should be usefulness, effectiveness, learnability, flexibility and satisfiability [1].

On the other hand, discussing web-based instructional strategies from cognitive point of view requires a different approach that should mainly be constructive (demands for synthesis in the process of building the tool) and argumentative (justify design decisions of the developed tool, critically asses the tradeoffs in alternative designs and conduct usability studies to evaluate prototypes).

2 Related works

Let's regard related works mainly in the following research areas: educational data mining, HCI and cognitonics. Within the educational data mining (EDM) area this paper is closely related to Intelligent Data Analysis performed on educational data.

Among many research problems from EDM, this paper relates to the works that have attempted to represent and analyze educational data with the goal to improve their knowledge level through custom designed data analysis systems and applications. Thus, implementing data mining tasks on educational data (e.g. performed activities, educational resources, messages, etc.) provides the environment in which progress may be achieved.

Many research hours have been allocated to the purpose of extracting key concepts from course materials, messages, questions and finding ways of using them for enhancing the teaching and learning processes [2, 3].

Also, a considerable amount of work has been put into discovering the similarity between concepts. Relevant in this area is the paper that Tara Madhyastha wrote in 2009 called „Mining Diagnostic Assessment Data for Concept Similarity" [4], it presented a method for mining multiple-choice assessment data for similarity of the concepts represented by the multiple choice responses. The result obtained was a similarity matrix that can be used to visualize the distance between concepts in a lower-dimensional space.

The NLP (Natural Language Processing) is another major research area, with a strong focus on documents (text and diagrams). Particularly interesting from our perspective is the research conducted in the domain of linguistics [11, 12]. Important is also the work put into constructing treebanks, both monolingual [13] and parallel. In [9] M. Colhon presents the construction of an English-Romanian treebank, a bilingual parallel corpora with syntactic tree-based annotation on both sides, also called a parallel treebank. Treebanks can be used to train or test parsers, syntax-based machine translation systems, and other statistically based natural language applications.

Agathe Merceron and Kalina Yacef published a case study about how educational data mining algorithms can be used for extracting relevant information from web-based educational systems and how this information can be used for helping teachers and learners [5]. A comprehensive report on the state of the educational data mining was published in 2009 [6] and presented a general view of the EDM field, including the methods used, the papers with the most influence in the field and the key areas of application.

One of the main goals of the educational research is identifying students' current level of understanding. For this purpose, a series of estimates have been used, including DINA model, sum-scores and capability matrix. A comparison between these estimates was presented in „A Comparison of Student Skill Knowledge Estimates" [7].

This paper presents an approach to using concept extraction along with activity monitoring and concept weighting towards constructing accurate models of students' present knowledge and level of understanding of the courses, as well as detecting the difficulty level of each course, course chapter, test question or exam question.

Another scientific discipline whose contribution is necessary is cognitonics [18, 19, 20]. Cognitive aspects are playing an important role in the information society and also in the particular case of e-learning applications. From cognitonics point of view, the developed applications for sustaining on-line courses (or other related activities) should develop creativity, support cognitive-emotional sphere and appreciate the roots of the national cultures. One of the main goals of this quite new research discipline it is the development of a new generation of tools for on-line learning that compensate the broadly observed negative distortions [18].

Among many application domains where cognitonics (cognitive psychology for information society and advanced educational methods) finds a suitable place is e-learning. From this point of view, e-learning tends to need progress from various research domains (e.g. EDM, cognitonics, HCI, etc.) in order to improve its effectiveness and control the main negative side effects regarding linguistic ability, phonological ability, social relationships, etc.

Another research domain that is highly connected with the discussed issues is HCI. Currently, there are numerous research efforts that deal with user interface adaptation in e-Learning Systems, adaptable interfaces featuring multiple views and finally integration of usability assessment frameworks that are designed and refined for the context in which they are applied. HCI issues related to e-learning are user-centered design [21] and user sensitive design [22]. From this perspective, adaptation of knowledge presentation, of interaction style regard specific issues like domain knowledge base generation, user/system interaction modeling, interface evaluation. Poor interaction in various on-line educational activities (e.g. evaluation of exercises after class, quiz games, intelligence analysis, etc.) may find proper solution by employing specific HCI research methodologies related to usefulness, effectiveness, learnability, flexibility and satisfiability.

3 Tools and technologies

The context for which the tool is developed is related to on-line education. From this point of view, the developed tool is actually a web application that gathers various technologies in order to achieve its business goal.

The first step in accomplishing this module's purpose is retrieving the text from documents. For reading .pdf files, we used Apache PDFBox, which is is an open source Java tool for working with PDF documents [10]. For manipulating .doc and .docx files, our choice was Apache POI [14], a powerful Java API designed for handling different Microsoft Office file formats.

Stemming [16] is the process for reducing inflected (or sometimes derived) words to their stem, base or root. The documents written in English were stemmed using the snowball stemmer [15]; as for the Romanian stemmer, we used as a base the PTStemmer implementation [17] and adapted it for the Romanian language by building the corresponding set of grammatical reduction rules: plural reduction, genre reduction, article reduction. The PTStemmer is a toolkit that provides Java, Python, and .NET C# implementations of several Portuguese language stemming algorithms (Orengo, Porter, and Savoy).

The XML processing was done using the Java DOM parser.

4 System architecture and usability assessment

The most important prerequisite for the development of such a tool is an online educational platform that has a proper structure for the educational assets and the ability to integrate proper intelligent data analysis techniques.

The online educational system we have chosen is Tesys Web [8], an e-learning platform used in several faculties from University of Craiova. Tesys has been designed and implemented to offer users a collaborative environment in which they can perform educational activities.

4.1 General architecture

Figure 1 presents the general architecture of the system. In the left part of the figure we can see only persistent data, basically found on the server, and on the right side the core business logic is presented, it includes the concept extraction, activity monitoring and recommender modules.

Starting from the course documents that were previously uploaded by the professor on the platform, the system extracts the concepts, using a custom concept extraction module, which incorporates a stemming algorithm and TF-IDF formulas. The obtained data is then transferred into the XML files. The five most relevant concepts are also inserted into the Tesys database, for further use.

As soon as the professor uploads the test questions and specifies each concept's weight for every question, the student's activity monitoring process can begin.

Afterwards, using the concept-weight association, student's responses to the test questions and taking into consideration the performances of student's colleagues, the system will be able to show relevant statistics to the professor, so he can understand each student's learning difficulties as well as the general level of the class.

The recommender module is designed to review the difficulty of the proposed exam questions and advise the professor on lowering or increasing the exam difficulty. All this process is supervised by the professor, who takes the final decision.

4.2 Concept extraction tool

A key feature of this module is the extraction of the most important concepts from every chapter that belongs to a course.

This part of the module is divided into two steps: stemming and computing TF-IDF values.

Several tools and algorithms have been developed for English word stemming, but for the Romanian language this research area is still at the beginning, therefore we developed our own tool and set of rules to accomplish this task.

After the stemming process we use the TF-IDF formulas for every word in the document and then we store the obtained data into an xml file, which has the following structure:

```
<?xml version="1.0" encoding="UTF-8"
standalone="no"?>
<words>
<regul originalForm="reguli"
tf="1.00" idf="0.47"/>
```

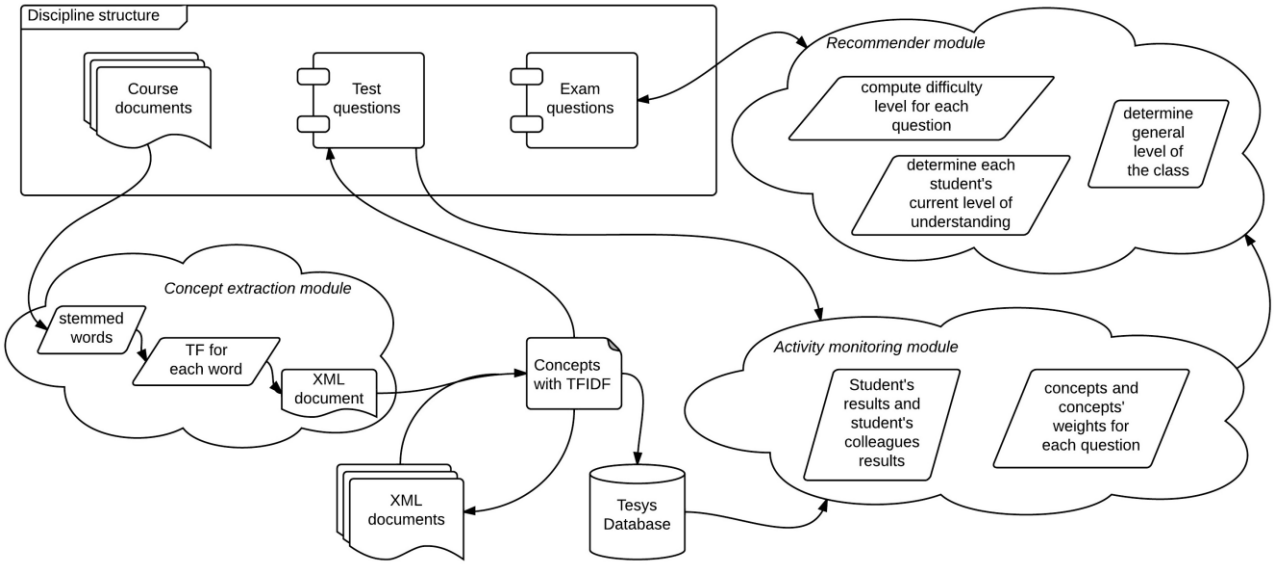


Figure 1: System architecture.

```
<fapt originalForm="fapt" tf="0.74"
idf="0.47"/>
 <atribut originalForm="atribut"
tf="0.45" idf="0.47"/>
 <sistem originalForm="sistem"
tf="0.33" idf="0.47"/>
 <inferent originalForm="inferenta"
tf="0.32" idf="0.55"/>
<algoritm originalForm="algoritm"
tf="0.32" idf="0.55"/>
        . . . . . .
 <absolut originalForm="absolut"
tf="0.01" idf="1.25"/>
</words>
```

Each concept extracted from the course chapter's document is represented as an element in the xml file. The stemmed form of the word is stored as the element name, and its original form, TF value and IDF value are stored in the element attributes.

The first five concepts that have the highest TF-IDF value are inserted in the database, for further use on the platform.

Figure 2 illustrates part of the interface available to the professor for managing the concepts. It is very straight forward, providing the professor with the list of extracted concepts and some additional options for managing them. These options include: the possibility to add new concepts, modify the existing ones in case they were not correctly extracted and delete the irrelevant concepts, if any.

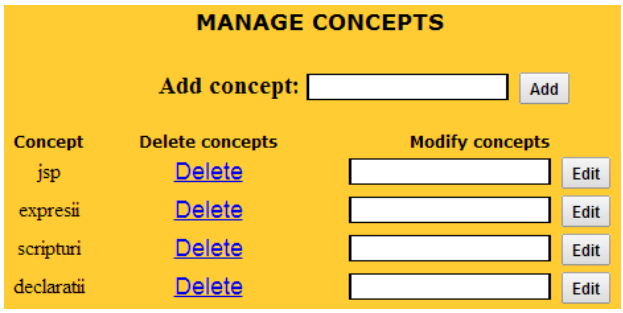


Figure 2: Concepts management.

4.3 Discipline structure

Within the online platform a discipline has the following structure: Chapters, Test Questions, Exam Questions and Concepts.

The chapters are documents uploaded by the professors, which can have one of the following extensions: *.pdf*, *.doc*, *.docx*. These documents are parsed and stemmed, resulting in a list of concepts.

Test questions are the questions used by the students throughout the semester for evaluating their current knowledge. A feature that allows the professor to choose from a list the concepts that are related to the test question and assign them weights was added to the e-learning platform.

The exam questions are the ones from which the students will take the final exam and obtain their final grades.

The concepts are the ones extracted from the chapters' documents which were previously reviewed by the professor.

As presented in Figure 3, for each question it is available the list of concepts extracted from the chapter to which the question belongs. Here is where the professor has the ability to assign the corresponding weights, representing the level of relevance that the concept has to the question.

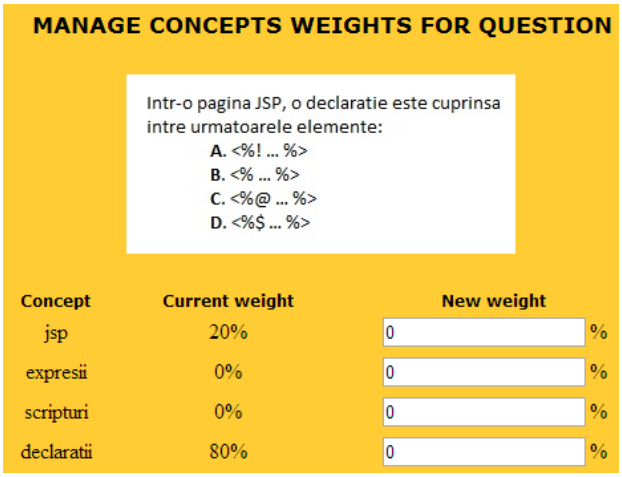


Figure 3: Weights management.

4.4 Activity monitoring

This step is very important because it is decisive for the determination of the student's ability to understand the course material and to highlight his/her progress. The most relevant information can be obtained by evaluating the correctness of the answers from the test questions, taking into consideration the concept associated with those questions and their given weights. This will help the professor figure out which are the concepts that the student has difficulties understanding and how he/she can be helped.

One of the monitoring tools provided to the professor is the graphic presented in Figure 4. By regularly checking this table, the professor can watch the progress of his/her students, be informed of their level of interest in the course material and discover which are the concepts that pose them problems.

4.5 Usability assessment and cognitonics

Usability evaluation is the final and most critical step within the lifecycle of the application, since it may have tremendous implications on the redesign of the user interface and underlying business logic. Applying general heuristics (i.e. with no special tuning to educational context) may be a reasonable option but using approaches that are adapted to e-Learning may offer greater progress [23].

From cognitive perspective, the visual stimuli refer to the following processes: visual search, find, identify,

	First/Last name	Overall understanding level	No. answered questions	Concept 1: jsp	Concept 2: expresii	Concept 3: scripturi	Concept 4: declaratii
1	Student18 Student18	78.37	5	100	73.33	100	100
2	Student17 Student17	77.17	10	100	80	100	100
3	Student14 Student14	56.13	10	100	63.33	100	66.67
4	Student1 Student1	53.7	5	0	60	0	66.67

Figure 4: Activity monitoring.

recognize and memory search [24]. Based on this kind of analysis, specific issues may be obtained specific issues within our e-learning tool. From this perspective, presented educational assets (e.g. quizzes, concepts, etc.) can be characterized as confusing, not findable, etc. and thus reduce the unwanted side effects regarding the distortions in perception of the world caused by information society and globalization.

5 Experimental results

In order to better explain how the system works, we will consider a sample usage scenario. The main steps of the scenario are:

Professor – Concept setup

Let's assume that professor P is a professor on the e-learning platform, and has a course with two chapters. He first has to upload the documents of the chapters. Immediately after he does that, the system parses the documents, applies the stemming algorithm and the TF-IDF formulas and ultimately extracts the most important five key concepts from each file: C11-C15, C21-C25.

This list can be accessed from the course page and is differentiated by the chapter to which they belong. In this moment the professor can review the concepts, he/she can delete the ones that he/she considers irrelevant, or maybe add new ones.

Professor – Test questions setup

This step is performed when professor P loads test questions for the students to answer, and for each question assigns weights of the extracted concepts, denoting the relevance level of every concept for each particular question. The weights have values in the range of [0.0,1.0].

Table 1 presents a possible weight distribution for concepts among questions. The cells corresponding to the concepts that have absolutely no relevance to a question and therefore have a weight of 0.0 are left blank.

	C11	C12	…	C21	C22	C23	C24	C25
Q1		0.1		0.7			0.2	
Q2			0.4		0.6			
Q3		0.7			0.1	0.1		0.1
…								

Table 1: Sample weights distribution.

On the platform, the professor is able to assign the weights as percentages, as previously presented in Figure 3. These weights can be updated at any time, and the progress of the students will be modified accordingly.

Student – Take tests

Let us consider student S1. The first test the student takes contains questions 1, 2, 3, 5, 7. Table 2 presents possible values for the correctness of the answers given by the students that answered these questions.

S1 is the analyzed student, S2 to Sn are the other students that answered the questions.

It is assumed that the tests contain only single choice questions, so the answer can be only evaluated as CORRECT or INCORRECT.

	Q1	Q2	Q3	Q5	Q7
S1	CORRECT	CORRECT	INCORRECT	CORRECT	INCORRECT
S2	INCORRECT	CORRECT	INCORRECT	INCORRECT	CORRECT
S3	CORRECT	CORRECT	INCORRECT	CORRECT	CORRECT
…					
Sn	CORRECT	CORRECT	CORRECT	CORRECT	INCORRECT

Table 2: Sample Answer Data.

After computing the weights and results, the system will provide the professor with the following statistics:

- student's level of understanding of each concept:

$$LU_c = \frac{\sum_{i=1}^{n-p} w(c, q_i)}{\sum_{j=1}^{n} w(c, q_j)}$$

In this formula, LU_c represents the level of understanding of concept c and $w(c, q_i)$ is the weight associated to the concept for question q_i. The numerator of the fraction is therefore the sum of the weights of the concept for the correctly resolved questions, and the denominator is the maximum amount that could be obtained if the responses to all questions were correct.

- student's performance relative to his colleagues;

$$Performance = \sum_{i=1}^{k} \frac{LUc_i}{Avg\,(LUc_i)}$$

- the difficulty level of every test question

$$Difficulty = \frac{No.\,correct\ answers}{Total\ no.\,answers}$$

- the difficulty level of each concept

S1. Professor – Visualize results : build mental model

After analyzing the presented data, the professor will be able to start creating a mental model regarding the student's current level of understanding of the material and his/her place among the other students. Also, if necessary, the professor might decide to modify the course material, for example add some extra information on a particular concept that the students have trouble understanding.

Another action the professor might choose to take, given the reported level of the class, is increase or decrease the general difficulty level for the test questions, as well as deciding which will be the best exam questions.

6 Conclusions and future works

This paper presents a use case of building a tool for Tesys e-Learning platform and analyzing cognitonics and HCI related issues in an attempt to offer a high quality interaction design that minimizes the cognitive side effects.

The developed tool is presented in detail from architectural and technical point of view, with an emphasis on the design of the user interface and on the data processing issues.

The technical challenges that are addressed in this paper regard building a Romanian stemmer, obtaining concepts, designing mathematical formulas for determining concept and quiz weights and overall students' knowledge levels. From this point of view, the future works regard validation of these mathematical formulas and possibly inferring better ones. As a general approach, continuous usage of the tool will provide data evidence for our approach.

Another important issue, discussed in this paper, regards the HCI and cognitonics aspects of the user interface designed for this software tool. From this perspective, usability evaluation (general or e-learning related) using HCI specific methodologies represents the final step in obtaining a high quality user interface. From cognitive perspective, the goal is to minimize (or ideally eliminate) the distortions in perception of the world cost by the developed tool. The cognitive aspects validate the student model that is mentally built by the professor while using the tool.

As future works, there are two main directions. One regards properly the analysis of the underlying data (e.g. concepts, weights, formulas) and the other one regards further analysis from HCI and cognitonics perspective of the developed tool. Once progress is made in these directions, similar e-learning tools may also be analyzed providing a framework for progress in this domain.

References

[1] B. Mehlenbacher, L. Bennett, T. Bird, M. Ivey, J. Lucas, J. Morton, L. Whitman. Usable E-Learning: A Conceptual Model for Evaluation and Design. *Proceedings of HCI International 2005: 11th International Conference on Human-Computer Interaction*. Las Vegas, NV: Mira Digital P, 1-10.

[2] K. Nakata, A. Voss, M. Juhnke, T. Kreifelts. Collaborative Concept Extraction from Documents (1998). *Proceedings of the 2nd Int. Conf. on Practical Aspects of Knowledge management (PAKM 98)*. 1998. Switzerland.

[3] J. Villalon, R. A. Calvo. Concept Extraction from Student Essays, Towards Concept Map Mining. *Proceedings of the 2009 Ninth IEEE International Conference on Advanced Learning Technologies*, Riga, Latvia, 2009, Volume 0: 221-225.

[4] T. Madhyastha, E. Hunt. Mining Diagnostic Assessment Data for Concept Similarity. Journal of Educational Data Mining (JEDM). pp. 72-91. 2010.

[5] A. Merceron, K. Yacef. Educational Data Mining: a Case Study. *Proceedings of the 12th international Conference on Artificial Intelligence in Education AIED*. pp. 467-474. Amsterdam, The Netherlands. 2005.

[6] R.S.J.D. Baker, K. Yacef. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*. pp. 3-17. 2009.

[7] E. Ayers, R. Nugent, N. Dean. A Comparison of Student Skill Knowledge Estimates. *EDM2009: 2nd International Conference on Educational Data Mining*. Cordoba, Spain. 2009.

[8] D. D. Burdescu, M. C. Mihaescu, TESYS: e-Learning Application Built on a Web Platform. *ICE-B*. pp. 315-318. 2006.

[9] Colhon, M.: Language Engineering for Syntactic Knowledge Transfer. Computer Science and Information Systems, Vol. 9, No. 3, 1231-1248. (2012)
[10] B. Litchfield. Making PDFs Portable: Integrating PDF and Java Technology. *Java Developer's Journal*. 2005.
[11] D. Cristea, C. Forăscu, "Linguistic Resources and Technologies for Romanian Language", Computer Science Journal of Moldova, vol. 14, no. 1(40). (2006)
[12] D. Klein, C. D. Manning, "Accurate Unlexicalized Parsing", In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430. (2003)
[13] M. P. Marcus, B. Santorini and M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank". In COMPUTATIONAL LINGUISTICS, vol. 19(2), 313-330. (1993)
[14] Apache POI - the Java API for Microsoft Documents. *http://poi.apache.org/*
[15] Snowball. *http://snowball.tartarus.org/*
[16] Deepika Sharma. Stemming Algorithms: A Comparative Study and their Analysis. *International Journal of Applied Information Systems*. pp. 7-12. 2012.
[17] PTStemmer - A Stemming toolkit for the Portuguese language. *https://code.google.com/p/ptstemmer/*
[18] Fomichov, V. A., Fomichova, O. S. Cognitonics as a New Science and Its Significance for Informatics and Information Society. *Informatica. An International Journal of Computing and Informatics (Slovenia)*, 2006, Vol. 30, No. 4, pp. 387–398; http://www.informatica.si/vol30.htm#No4; retrieved 14.06.2014.
[19] Bohanec, M., Gams, M., Mladenic, D. et al, Eds. (2011). Proceedings of the 14th International Multiconference Information Society – IS 2011, Vol. A, Slovenia, Ljubljana, 10 – 14 October 2011. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; , http://is.ijs.si/is/is2011/zborniki.asp?lang=eng; pp. 347-430; retrieved 15.05.2014
[20] Gams, M., Piltaver, R., Mladenic, D. et al., Eds. (2013). Proceedings of the 16th International Multiconference Information Society – IS 2013, Slovenia, Ljubljana, 7 – 11 October 2013. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute; http://is.ijs.si/is/is2013/zborniki.asp?lang=eng; pp. 403-482; retrieved 15.05.2014.
[21] D. Norman, S.W. Draper (eds.): User Centered System Design. Earlbaum, Hillsdale (1986)
[22] A. Granić, V. Glavinić. Automatic Adaptation of User Interfaces for Computerized Educational Systems. In: Zabalawi, I. (ed.) Proceedings of the 10th IEEE International Conference on Electronics, Circuits and Systems (ICECS 2003), Sharjah, Dubai, pp. 1232–1235 (2003)
[23] D. Squires, J. Preece. Predicting quality in educational software: Evaluating for learning, usability and the synergy between them. Interacting with Computers 11, 467–483 (1999)
[24] Shu-mei Zhang, Qin-chuan Zhan, He-min Du. Research on the Human Computer Interaction of E-learning. International Conference on Artificial Intelligence and Education (ICAIE), 2010.
[25] Jakob Nielsen and Rolf Molich, Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90), Jane Carrasco Chew and John Whiteside (Eds.). ACM, New York, NY, USA, 249-256, 1990.

A Database-Based Two-Phase Algorithm for Efficient and Complete Detection of siRNA Off-Target Homology

Hong Zhou
Department of Mathematical Science, School of Health and Natural Sciences
University of Saint Joseph
1678 Asylum Avenue, West Hartford, CT 06117, USA
E-mail: hzhou@usj.edu

Hong Wang
Department of Chemical Biology and Center for Cancer Prevention Research
The State University of New Jersey
Rutgers, NJ 08854, USA
E-mail: howang@rci.rutgers.edu

Keywords: siRNA, shRNA, off-targets, RNAi, sequence alignment, Smith-Waterman

Received: June 31, 2014

Since the discovery of RNA Interference (RNAi), a cellular phenomenon in which a small double stranded RNA induces the degradation of its sequence specific target mRNA, using a computer-aided software tool to help design functional small interfering RNA (siRNA) or small hairpin RNA (shRNA) has become a standard procedure in applying RNAi to silence a target gene. A critical consideration in siRNA design is to avoid any possible off-target effect, i.e. to avoid sequence homology with untargeted genes. Though BLAST is the most powerful sequence alignment tool, it can overlook some significant homologies. Therefore, Smith-Waterman algorithm is the only approach that can guarantee to find all possible mismatch alignments that may cause off-target effect. However, Smith-Waterman alignment suffers from its inefficiency in searching through a large sequence database. A two-phase search algorithm was previously reported in which the first phase is used to identify local regions where the second phase, a bona fide Smith-Waterman alignment, is absolutely needed. Though such a two-phase homology search can improve the efficiency up to two orders of magnitude over the original Smith-Waterman alignment algorithm, it is still not efficient enough to be used alone for siRNA off-target homology search over a large sequence database. In this paper, we propose several improvements that dramatically speed up the reported two-phase algorithm while still guaranteeing the complete identification of siRNA off-target homologies.

Povzetek: V prispevku je predstavljena računalniška metoda za utišanje ciljnega gena.

1 Introduction

RNA interference (RNAi) is a cellular mechanism in which a small double stranded RNA induces the degradation of its sequence specific target mRNA, thus silencing the function of the target gene. Since its discovery, RNAi has become a powerful technique to knock out/down the expression of target genes for gene function studies in various organisms [3,5,16]. To employ this technique, the first step is to design target-specific small interference RNA (siRNA) or small hairpin RNA (shRNA) that is homologous to the target mRNA. Because of the predictability of RNAi based on its matching target sequence [2, 5, 7, 9–11, 14, 15, 19, 22, 25, 26], quite a few studies have been devoted to computer-guided algorithms to design effective siRNA or shRNA (from here on, this article will only refer to siRNA for simplicity) [4, 6, 12, 15, 20, 25, 26]. However, a critical requirement in siRNA design is to guarantee that the designed siRNA is free of off-target effect. Although the actual mechanism of off-target effect is still unknown, it has been demonstrated that a partial sequence homology between siRNA and its unintended targets is one of the major contributing factors [8,18,21]. It has been suggested that if an introduced siRNA has less than 3 mismatches with an unintended mRNA, it would likely knock down the expression of this mRNA in addition to its intended target which shares 100% sequence homology with this siRNA [11,15]. Unsurprisingly, the Basic Local Alignment Search Tool (BLAST) has been used to identify possible unintended homologous regions for siRNA candidates [1,13,17]. BLAST, although extremely fast, is not the best algorithm designed for this type of task since it overlooks significant sequence homologies [15,24,27]. As an alternative, Smith-Waterman alignment algorithm has been employed together with BLAST by

some design tools to identify all possible off-target sequences [15,27].

Smith-Waterman algorithm utilizes a dynamic programming approach to identify the local optimal alignment between two sequences [23]. It guarantees to locate the existing optimal alignment based on a scoring system with a set of scores assigned to a match, a substitution, a deletion, and an insertion. Given two sequences with length of *m* and *n*, the computational complexity of Smith-Waterman algorithm is $O(mn)$. Since the off-target search for siRNA sequences must be conducted completely through a given sequence database (which is usually large), the Smith-Waterman algorithm alone becomes very time-consuming and impractical for this task. Thus we once developed a two-phase homology search algorithm for siRNA off-target detection [29]. In this two-phase algorithm, the phase 1 procedure is used to identify the local regions where an off-target homology is possible to exist. Upon finding such local regions, the phase 2 procedure, a bona fide Smith-Waterman alignment algorithm, is used to determine if this local region has homology with the given siRNA sequence to cause off-target effect. This two-phase algorithm can be explained as the following.

For a siRNA of length *m*, an off-target homology is defined as a sequence that has less than *x* mismatches (i.e. mismatch cut-off equals *x*) when aligned against the siRNA (a mismatch is defined to be either a substitution, a deletion or an insertion hereafter). Thus, after the siRNA sequence is divided into *x* mutually disjointed and equal substrings (as equal as possible), at least one substring must have a perfect match with the off-target region. For the remainder of this paper, let's assume $m=21$ and $x=3$ unless stated otherwise. Under this condition, an off-target homology can only have a maximum of two mismatches, i.e., 0, 1, or 2 mismatches. When there are a maximum of two mismatches, no matter where the possible two mismatches are, at least one third of the siRNA sequence must have an exact match with the homological region. This concept is shown in Figure 1 which explains the case when the middle substring has the exact match.

Since all the possible off-target homological regions bear a substring of length 7 that has an exact match with the siRNA sequence, it is reasonable to perform the Smith-Waterman alignment only on the regions that have an exact match with at least one substring of the siRNA sequence. Thus, the first phase in the two-phase algorithm is designed to identify the potential regions with which at least one of the substrings of the siRNA sequence has an exact match. Only when such a potential region is identified, the second phase calls for the Smith-Waterman procedure to evaluate the best alignment between the potential region and the siRNA sequence. This algorithm does not construct any lookup table from the whole genome sequences, though it significantly improves the searching efficiency by guiding the most time-consuming core Smith-Waterman alignment on the local regions that need to be further examined.

Though the two-phase algorithm was shown to have efficiency gain of up to two orders of magnitude compared to the original Smith-Waterman algorithm alone [29], it is still not efficient enough to be applied alone for off-target homology search for a large number of siRNA sequences, such as the whole-genome siRNA design and off-target detection. For whole-genome siRNA design and off-target search, this two-phase algorithm must be applied with BLAST being the initial screening tool. In this paper, we present several significant improvements over both the phase 1 and the phase 2 procedures. These improvements dramatically speed up the original two-phase algorithm and make it able to complete off-target homology detection by itself alone for whole genome siRNA design.

2 Materials

The computer used in this study is a Dell notebook computer with Intel Core(TM) i5-2410M CPU. The maximum CPU speed is 2.30GHz. Installed RAM is 8.00 GB with 7.88 GB usable. The operating system is Windows 7 Enterprise (64 bit). The programming language used is Java.

The genome sequence database used in this study is NCBI human mRNA RefSeq gene database (human.rna.fna) downloaded on December 9, 2013. It has 68822 non-redundant sequences for mRNA/protein genes with average length of 3452 nucleotides.

The 1000 sample siRNA sequences used in this study were generated as the following: after 100 genes were randomly selected from the NCBI human mRNA RegSeq database, 10 siRNA were generated randomly from each gene using a computer-aided siRNA design tool [27]. All the siRNA sequences are of length 21 nucleotides (21-nt). One reason to select the length 21 is

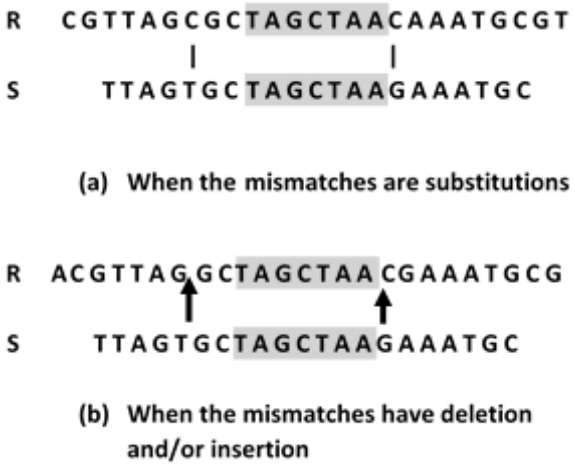


Figure 1: When there are 2 mismatches between the siRNA sequence (S) and the off-target region (R), at least one of the three substrings of S has the exact match with R. The vertical bars mark the substitutions, the arrows mark the deletion and/or insertion in R, and the shaded substrings have the exact match. When the substring in the middle of S has an exact match, the off-target region must be such a region in R that extends from the matched substring to both left and right enough base pairs to completely cover the siRNA

that 21-nt siRNA is the most commonly used siRNA in RNAi applications and the naturally occurring endo-siRNA is of 21-nt [30,31].

3 Improvements on phase 1

In the original two-phase algorithm, the phase 1 is nearly as five times time-consuming as the phase 2. This is shown in Table 1.

Table 1: The time cost (in seconds) analysis of the two-phase algorithm.

# of siRNA	Time Cost (seconds)	
	Phase-1	Phase-2
100	388.49	80.51
200	777.04	160.53
500	1938.97	404.02
1000	3866.47	798.10

The result in Table 1 is obtained by conducting off-target homology search through the whole human mRNA RefSeq gene database using the sample set of the 1000 21-nt siRNA. As the phase 1 is much more inefficient than the phase 2, our first improvement is on the phase 1.

The original two-phase algorithm used the Java's built-in string match algorithm, which is a character-by-charter brutal force algorithm. This algorithm has been shown to be inefficient in English text match. However, our experiment result shows that this brutal force algorithm performed equally efficient compared with both the Knuth-Morris-Pratt (KMP) algorithm and the Boyer-Moore algorithm in the siRNA off-target homology search. This is in fact unsurprising. There are only four different nucleotides in both DNA and RNA sequences, thus repeated sequences can occur frequently. The repeating sequences prohibit the skip-distance in both KMP and Boyer-Moore algorithms from growing, making them unable to achieve the desired efficiency gain.

The fact that there are only four different nucleotides in DNA sequences (let's use DNA as the example as the RefSeq database is for DNA) inspired us to develop a base-4 integer number system to represent DNA sequences. For example, let's define A=0, C=1, G=2 and T=3, then any nucleotide can be represented by a base-4 digit 0, 1, 2 or 3. Though the original two-phase algorithm works with siRNA of different lengths, in this study, siRNA of length 21 is used as the working sample. The reason is that 21-nt siRNA is the most commonly used and the naturally occurring endo-siRNA is of 21-nt [30,31]. In the NCBI probe database which contains thousands of siRNA sequences submitted by different researchers or companies, about 60% of these siRNA sequences are of length 21. However, please note that the concepts introduced in this study work for siRNA of different lengths. A 21-nt siRNA can be divided into three substrings each of size 7-nt. With the base-4 number system, any 7-nt can be represented by 7 digits, which is a base-10 integer between 0 and 16383 inclusively (please observe that $4^7 = 16384$). This means that a siRNA sequence can be represented by three base-4 integers each for a 7-nt subsequence. For example, a siRNA sequence of GCUGCAUCAACACAUGGAGCA is divided to three mutually disjointed 7-nt substrings GCUGCAU, CAACACA, UGGAGCA, which are represented as three integers 10131, 4164, and 14884 respectively. However, a DNA gene sequence of length M nucleotides must be represented by M-7+1 integers. This is because the homology search against the gene sequence is contiguous, shifting a nucleotide at a time. For example, AGCTATCCG is represented as an integer array of {2509, 10037, 7382}.

In the next experiment, we pre-processed the mRNA RefSeq database to convert every gene sequence into an array of integers. With this conversion, the phase 1 string match procedure becomes integer equivalence checking. It is not surprising to observe that the phase 1 procedure is significantly improved by representing the sequences as integers. The result is shown in Table 2.

Table 2: The time cost comparison between the original phase 1 and the modified phase 1 in which character by character comparison is transformed into integer comparison. (o): original Phase 1. (n): the new Phase 1 using integer comparison.

# of siRNA	Time Cost (seconds)		
	Phase1 (o)	Phase1 (n)	Phase2
100	388.49	164.20	80.51
200	777.04	345.81	160.53
500	1938.97	842.18	404.02
1000	3866.47	1760.54	798.10

Table 2 demonstrates that by using a base-4 integer system to represent the DNA nucleotides and thus transforming the string match process into an integer comparison process, the time cost of the original phase 1 can be cut down by more than 50%. The overall efficiency gain of the whole process is about 45%. Though the above experimental result is positive, the improvement is not significant enough. It is clear that dynamically searching for the exact match of a substring is always time-consuming. This motivated us to build a database to index the locations where each siRNA 7-nt substring has an exact match with the DNA gene sequences.

In the RefSeq database, there are 68822 non-redundant gene sequences with an average length of 3452 nucleotides. If we assume that all the four nucleotides have an equal chance to appear through the whole sequence database, then any a 7-nt subsequence has 1/16384 chance to appear, i.e. can show up about 14500 times in the whole gene database. To build the location-indexed database, we generated all the permutations (total 16384) of 7-nt, found the locations of each 7-nt in the RefSeq database and stored their location information in the location-indexed database. By using this location-indexed database, the phase 1 search process is no longer dynamic. Whenever a siRNA 7-nt substring needs to locate its exact matched regions inside the RefSeq gene database, using the integer

representation of the 7-nt substring as the primary key, such needed information is directly provided through this database. Via using such a database, the efficiency of the phase 1 is greatly improved. The result is shown in Table 3.

Table 3: By using a database to store the locations where each 7-nt substring has an exact match in the RefSeq gene sequence database, the phase 1 process is dramatically accelerated. Time-cost values are in seconds. (o): original Phase 1. (n): the new Phase 1 through the location-indexed database.

# of siRNA	Time Cost (seconds)		
	Phase-1 (o)	Phase-1 (n)	Phase-2
100	388.49	0.75	80.51
200	777.04	1.25	160.53
500	1938.97	3.09	404.06
1000	3866.47	5.96	798.10

Table 3 demonstrates that removing the dynamic searching process via a pre-built location-indexed database, the phase 1 process is speeded up by about 600 fold. Table 3 also shows that the phase 2 becomes now the bottleneck in the two-phase algorithm.

Because the phase 2 is now much slower than the phase 1 after using the pre-built database, the overall efficiency gain of the modified two-phase algorithm is only about 5 fold. The challenge becomes now, how to improve the phase 2.

4 Improvements on phase 2

The original phase 2 is a bona fide Smith-Waterman alignment algorithm. As the phase 1 is used to reduce the probability of using Smith-Waterman alignment in phase 2, we then tried to further reduce the use of the phase two operation by adding a pre-phase right before the original phase 2. This pre-phase serves as a filter to further remove unnecessary Smith-Waterman alignment.

The pre-phase dictates that only when the following two conditions are both met, Smith-Waterman alignment is needed.

Precondition: A 21-nt siRNA sequence (S) is equally divided into three mutually disjointed 7-nt substrings, S0, S1, and S3. When S0 finds an exact match with a substring R0 in region (R), the two other substrings of R would be R1 and R2, each corresponding to S1 and S2 separately.

Condition 1: Divide S1 from the middle to generate two sub-substrings. One is 3-nt, and the other is 4-nt. Repeat the dividing for S2. For each of the two corresponding substrings R1 and R2, extend one nucleotide to the direction away from R0 so that both R1 and R2 are of 8-nt. Check if R1 contains (the position does not need to match) either of the two sub-substrings of S1. Repeat the checking for R2. It must be true that the total number of sub-substrings contained in R1 and R2 is no less than 2.

Condition 2: Divide S1 as equally as possible to generate three mutually disjointed sub-substrings. One is 2-nt, one is 3-nt, and the last is 2-nt. Repeat the dividing with S2. For R1 and R2, extend two nucleotides to the direction away from R0 so that both R1 and R2 are of 9-nt. Check if R1 contains (the position does not need to match) any of the sub-substrings of S1, and repeat the checking for R2. It must be true that the total number of sub-substrings contained in R1 and R2 is no less than 4.

Figure 2 illustrates the condition 1.

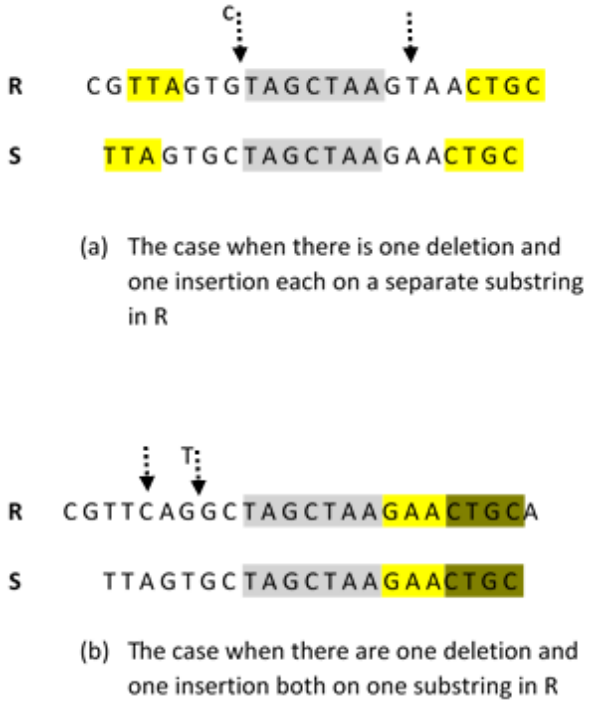


Figure 2: The Condition 1 in the case when the middle substring of siRNA finds an exact match in a region. In condition 1, there are always at least two sub-substrings of S that are contained inside R. Gray-shaded regions have the exact match. Arrows without a letter aside mark the insertion in R, and arrows with a letter aside mark the deletion (the letter indicates the deleted nucleotide in R). Yellow-shaded regions mark the sub-substrings of S contained in R.

The Condition 2 is depicted in Figure 3.

The first critical understanding of both Condition 1 and 2 is that when either S1 or S2 is divided into multiple sub-substrings, one mismatch, no matter what type it is, can only occur inside one sub-substring. Thus, in Condition 1, when there are four sub-substrings, at most two sub-substrings can be changed while at least two others are intact. Though a deletion or insertion can switch the positions of the sub-substrings, their content are not changed if the insertion/deletion are not inside the sub-substrings. A similar idea applies to Condition 2.

The second critical understanding of Condition 1 is that we need only to extend one nucleotide to the direction away from R0 so that both R1 and R2 are of 8-nt. The question raised here is that when R1 has two insertions, theoretically we need to extend two nucleotides so that R1 can fully cover S1. However, if R1 bears two insertions, given a homology between R and S, then S2 and R2 must be an exact match. Thus, there must be two sub-substrings of S2 that are contained

inside R2. It is then unnecessary to extend two nucleotides for R1 anymore. The similar idea can explain why it is necessary to extend only two nucleotides for both R1 and R2 in Condition 2.

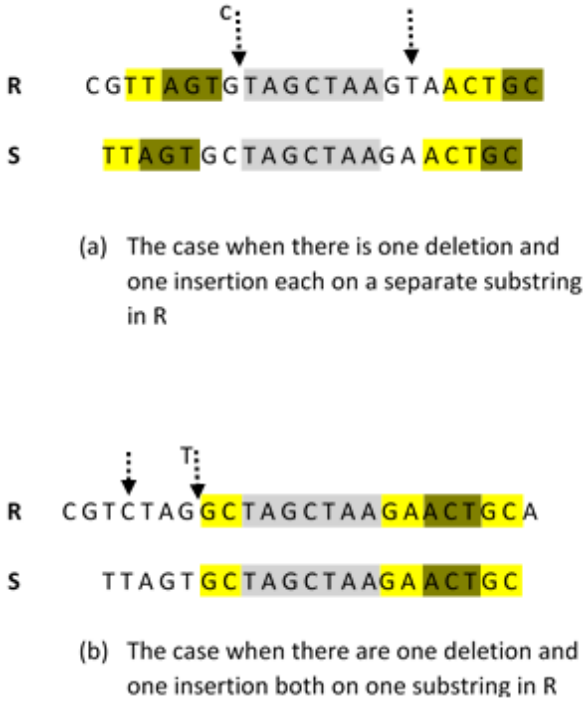


Figure 3: The Condition 2 in the case when the middle substring of siRNA finds an exact match in a region. In condition 2, there are always at least four sub-substrings of S that are contained inside R. Gray-shaded regions have the exact match. Arrows without a letter aside mark the insertion in R, and arrows with a letter aside mark the deletion (the letter indicates the deleted nucleotide in R). Yellow-shaded regions mark the sub-substrings of S contained in R.

With the pre-phase, the use of Smith-Waterman alignment is largely reduced and therefore the phase 2 is dramatically speeded up. The result is presented in Table 4.

Table 4: The pre-phase helps improve the efficiency of the phase 2 by more than 30 fold. (o) the old phase; (n) the new phase.

# of siRNA	Time Cost (seconds)		
	Phase1 (n)	Phase2 (o)	Phase2 (n)
100	0.75	80.51	2.60
200	1.25	160.53	49.46
500	3.09	404.06	12.68
1000	5.96	798.10	24.81

5 Discussion

The drawback of the phase 1 improvement is the necessity of building a database. Roughly speaking, for the 16384 different 7-nt substrings, there would be about 16384 x 14500 = 239018000 integers to store in the database, with each integer marking a position inside a gene for a 7-nt subsequence. In addition, there is other necessary information to store, such as the information of each gene. Depending on the implementation, the database size can be greater than or less than 1 Gb.

The pre-phase for phase 2 further reduces the use of Smith-Waterman alignment by mandating the satisfaction to both Condition 1 and Condition 2. Overall, the modified two-phase algorithm is 150 times more efficient than the original one. However, if only enforcing the satisfaction of one of the two conditions in the pre-phase, the improvement on efficiency is much less. By enforcing Condition 1 alone, the efficiency improvement on phase 2 is about 27 fold, while the efficiency improvement over the original phase 2 is only 11 fold if enforcing Condition 2 alone.

With the 1000 siRNA samples, there are 56402965 match hits in phase 1, indicating 56402965 alignment checking using Smith-Waterman algorithm in the original two-phase algorithm. However, there are only 399962 hits for the pre-phase. This shows that the pre-phase reduces the uses of Smith-Waterman alignment for about 140 fold. Among the 399962 hits, only 21444 of them were found to have true homology by Smith-Waterman alignment. This suggests that there might be additional approaches that can further improve the phase 2 efficiency.

Without considering the insertions or deletions, i.e. when only considering the case of substitutions, Smith-Waterman alignment is not necessary for the off-target homology detection. After the phase 1, for a homology with a maximum of two substitutions, the other two sub-strings in both siRNA and the searching region must have nearly exact matches with less than 3 substitutions. This is shown in Figure 4.

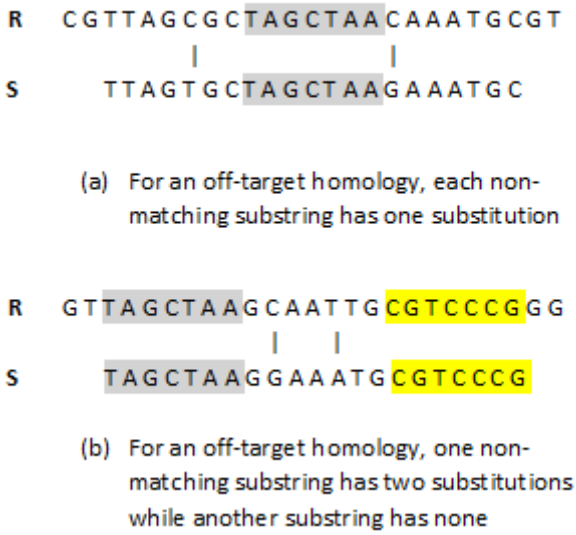


Figure 4: The case when only consider the substitutions in homology check. Gray-shaded regions indicate the exact match. Vertical bars mark the substitutions.

Since the substitutions do not change the positions of nucleotides, a check for the string matching on the two pairs of substrings can be quickly performed. The

experiment results show that it took only 4.00 seconds to complete the phase 2 for 1000 siRNA sequences. In addition, the experiment results disclose that there are only 21364 homologies found with only substitutions. Therefore, only 80 homologies identified for the 1000 siRNA sequences involve either deletions or insertions, a very small portion of the total number of off-target homologies (0.373%).

6 Conclusion

In the siRNA design, designing functional siRNA sequences is a relatively fast process, while the off-target evaluation is much more time consuming. Using the siRNA design tool [27], the time cost to design functional siRNA for all the 68822 human mRNA RefSeq non-redundant genes (an average of 33 siRNA for each gene) is about 400 seconds. With the improved two-phase algorithm (considering deletions and insertions), the time cost to completely check the off-target homology for all the designed siRNA sequences is estimated to be about 19.41 hours, which is acceptable for a process on the whole genome. Thus, after the improvements presented in this paper, the modified two-phase homology search algorithm can complete any off-target checking for functional siRNA design, without the initial use of BLAST.

Acknowledgement

The authors would like to thank Dr. Kevin Callahan and Dr. Joseph Manthey from the University of Saint Joseph for their critical reading of this manuscript.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., "Basic local alignment search tool," J Mol Biol., vol. 215, pp.403–410, 1990.

[2] Bass, B. L., "RNA interference. The short answer," Nature, vol.411, pp.428–429, 2001.

[3] Couzin, J., "BREAKTHROUGH OF THE YEAR: Small RNAs Make Big Splash," Science, vol.298, pp.2296–2297, 2002.

[4] Cui, W., Ning, J., Naik, U. P., Duncan, M. K., "OptiRNAi, an RNAi design tool," Comput Methods Programs Biomed., vol.75, pp.67–73, 2004.

[5] Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., Tuschl, T., "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells," Nature, vol.411, pp.494–498. 2001.

[6] Henschel, A., Buchholz, F., Habermann, B., "DEQOR: a web-based tool for the design and quality control of siRNAs," Nucleic Acids Res., vol.32, pp.W113–W120, 2004.

[7] Holen, T., Amarzguioui, M., Wiiger, M. T., Babaie, E., Prydz, H., "Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor," Nucleic Acids Res., vol.30, pp.1757–1766, 2002.

[8] Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G., "Expression profiling reveals off-target gene regulation by RNAi," Nat Biotechnol., vol.21, pp.635–637, 2003.

[9] Khvorova, A., Reynolds, A., Jayasena, S. D., "Functional siRNAs and miRNAs exhibit strand bias," Cell, vol.115, pp.209–216, 2003.

[10] Kim, D. H., Longo,M., Han, Y., Lundberg, P., Cantin, E., Rossi, J. J., "Interferon induction by siRNAs and ssRNAs synthesized by phage polymerase," Nat Biotechnol., vol.22, 321–325, 2004.

[11] Kim, D. H., Behlke,M. A., Rose, S. D., Chang, M. S., Choi, S., Rossi, J. J., "Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy," Nat Biotechnol., vol.23, pp.222–226, 2005.

[12] Levenkova, N., Gu, Q., Rux, J. J., "Gene specific siRNA selector," Bioinformatics, vol.20, pp.430–432, 2004.

[13] Lipman, D. J., Pearson, W. R., "Rapid and sensitive protein similarity searches," Science. vol.227, pp.1435–1441, 1985.

[14] Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., Tuschl, T., "Single-stranded antisense siRNAs guide target RNA cleavage in RNAi," Cell, vol.110, pp.563–574, 2002.

[15] Naito, Y., Yamada, T., Ui-Tei, K., Morishita, S., Saigo, K, "siDirect: highly effective, targetspecific siRNA design software for mammalian RNA interference," Nuclear Acids Research, vol.32, pp.W124–129, 2004.

[16] Paddison, P. J., Caudy, A. A., Hannon, G. J., "Stable suppression of gene expression by RNAi in mammalian cells," PNAS, vol.99, pp.1443–1448, 2002.

[17] Pearson, W. R., Lipman, D. J., "Improved tools for biological sequence comparison," PNAS, vol.85, pp.2444–2448, 1988.

[18] Persengiev, S. P., Zhu, X., Green, M. R., "Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs)," RNA, vol.10, pp.12–18, 2004.

[19] Reynolds, A., Leake, D., Boese, Q, Scaringe, S., Marshall, W. S., Khvorova, A., "Rational siRNA design for RNA interference," Nat Biotechnol., vol.22, pp.326–330, 2004.

[20] Sætrom, P., Snove, O. Jr., "A comparison of siRNA efficacy predictors," Biochemical and Biophysical Research Communications, vol.321, pp.247–253, 2004.

[21] Scacheri, P. C., Rozenblatt-Rosen, O., Caplen, N. J., Wolfsberg, T. G., Umayam, L., Lee, J. C., Hughes, C. M., Shanmugam, K. S., "Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in mammalian cells," PNAS, vol.101, pp.1892–1897, 2004.

[22] Siolas D., Lerner C., Burchard J, Ge W., Linsley P. S., Paddison P. J., "Synthetic shRNAs as potent

RNAi triggers," Nat Biotechnol., vol.23, pp.227–231, 2005.

[23] Smith, T. F.,Waterman,M. S., "Identification of common molecular subsequences," J Mol Biol., vol.147, pp.195–197, 1981.

[24] Snove, O., Jr., Holen, T., "Many commonly used siRNAs risk off-target activity," Biochem Biophys Res Commun., vol.319, pp.256–263, 2004.

[25] Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., Saigo, K., "Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference," Nucleic Acids Res., vol.32, pp.936–948, 2004.

[26] Yuan, B., Latek, R., Hossbach, M., Tuschl, T., Lewitter, F., "siRNA Selection Server: an automated siRNA oligonucleotide prediction server," Nucl. Acids Res., vol.32, pp.W130–W134, 2004.

[27] Zhou, H., Zeng, X., Wang, Y., Seyfarth, B. R., "A three-phase algorithm for computer aided siRNA design," Informatica (Slovene), 30 (2006) 357–364.

[28] Zhou, H., Zeng, X., Energy profile and secondary structure impact shRNA efficacy. BMC Genomics 2009, **10**(Suppl 1):S9.

[29] Zhou H, Wang Y, Zeng X: Fast and complete search of siRNA off-target sequences**.** In Proceedings of the international conference on bioinformatics & computational biology: 26–29 June 2006; Las Vegas. Edited by: Hamid R. Arabnia and Homayoun Valafar: CSREA Press; 2006:168-171.

[30] Carthew, R. W., Sontheimer, E. J., Origins and Mechanisms of miRNAs and siRNAs. Cell. 2009, 136(4): 642-655.

[31] Chen, L., Dahlstrom, J.E., Lee, S., Rangasamy, D., Naturally occurring endo-siRNA silences LINE-1 retrotransponsons in human cells through DNA methylation. Epigenetics, 7(7): 758-771, 2012.

A Model-Based Framework for Building Self-Adaptive Distributed Software

Ouanes Aissaoui[1], Abdelkrim Amirat[2] and Fadila Atil[1]
[1]LISCO Laboratory, Badji Mokhtar-Annaba University, P.O. Box 12, 23000 Annaba, Algeria
E-mail: aissaoui.ouenes@gmail.com, atil_fadila@yahoo.fr
[2]LiM Laboratory, University of Souk-Ahras, P.O. Box 1553, 41000 Souk-Ahras, Algeria
E-mail: abdelkrim.amirat@yahoo.com

Keywords: self-adaptive system, component-based system, dynamic reconfiguration, distributed reconfiguration, reliable reconfiguration

Received: August 16, 2013

Reconfiguration is widely used for evolving and adapting systems that cannot be shut down for update. However, in distributed systems, supporting reconfiguration is a challenging task since a reconfiguration consists of distributed reconfiguration actions that need to be coordinated and the application consistency must be preserved. To address this challenge, we propose a framework based on a reflexive three layer architecture model for the development of distributed dynamic and reliable component-based applications. The bottom layer of this model is the application layer. It contains the system's application-level functionality. The change management layer is the middle layer. It reacts to changes in state reported from the application layer. The uppermost layer is the self-adaptation layer that introduces the self-adaptation capabilities to the framework itself. It ensures the service continuity of the change management layer and manages the adaptation of this last to the changes which it carries out itself on the application layer. The framework is conceived especially for supporting the distributed reconfigurations. For that, it incorporates a negotiation and coordination mechanism for managing this type of reconfiguration. Moreover, it incorporates a separate system for ensuring the reliability of the application. The paper introduces a prototype implementation of the proposed framework and its empirical evaluation.

Povzetek: Članek predstavlja okolje za gradnjo samo-prilagodljivega porazdeljenega programja.

1 Introduction

Nowadays, more and more of distributed applications run more often in fluctuating environments such as mobile environments, clusters of machines and grids of processors. However, they must continue to run regardless of the conditions and provide high quality services. A solution to this problem is to provide mechanisms allowing the evolution or the change of an application during its running without stopping it [1, 16]. So, we talk about the dynamic adaptation of distributed applications which can be defined as the whole of the changes brought to a distributed application during its running [26].

Software reconfiguration [10] is strongly related to the domain of runtime software evolution and adaptation. In this domain, reconfiguration is used as a means for evolving and adapting software systems that cannot be shut down for update. Reconfiguration actions include component additions and removals; setting the parameter's value of a component; interfaces connections and disconnections; changes of the component state (started or stopped) and additions of new behaviours to component.

Several conditions must be checked by an adaptation operation where the most significant is the application consistency which can be summarized by the following points [19]:

- *Safety*: an adaptation operation badly made should not lead the adapted application to a crash.
- *Completeness*: At the end of a certain time, the adaptation must finish, and must at least introduce the changes necessary to the old version of the application.
- *Well-timedness*: it is necessary to launch the adaptation at the right time. The programmer must specify in advance the adaptation points.
- *Possibility of rollback*: even if we show the correctness of the adaptation, certain errors can escape from the rule. It is necessary to have some means that allow to cancel the adaptation and to roll back the application to its state known before the execution of the adaptation.

Therefore, the preservation of the application consistency is a very significant parameter to evaluate an approach for the dynamic adaptation.

Generally, the existing self-adaptive literature and research which has studied the dynamic adaptation of the distributed software systems provide solutions for adapting such systems, but the adaptation is not distributed (e.g. [29, 30, 31, 21, 9, 34]). In particular, the distribution of the adaptation system itself is rarely

considered. Also, parallel to the need of the dynamic adaptation of applications pose the problem of their reliabilities, which is an important attribute of the functioning safety [6]. In spite of the importance of the application reliability in the adaptation of applications, it was not taken into account in many works [31, 10, 9] particularly those treating the distributed reconfigurations. The works which studied this property in the dynamic adaptation did not reach the required level of coherence; it is only simple mechanisms generally based on the backward recovery technique (e.g. [20, 26]) which consists to roll back the application to a previously consistent state. Also, these mechanisms are incorporated in the components managing the adaptation of the application. So, the code responsible for the adaptation of the application is weaved with that which makes it reliable. Notice that this crosscutting of code prevents the evolution of the two mechanisms managing the reliability and adaptability.

After having identified this problem, we have concentrated on the reliable adaptation of the distributed component-based applications, which is a very topical. Our first objective is to provide a solution for the management of the distributed and coordinated dynamic adaptation. The second objective is to provide a separate solution for managing the fault tolerance of these applications in order to ensure their reliability which helps to lead to reliable reconfigurations, and the third objective is to facilitate the construction of this type of application studied by minimizing the time and the cost of the addition of the self-adaptation capabilities to it.

To achieve the first two objectives, we propose a reflexive three layer architecture model for the development of distributed dynamic and reliable applications. The bottom layer of this architecture model is the application layer which represents the software system. The change management layer is the middle layer. It reacts to changes in state reported from the application layer. The uppermost layer is the self-adaptation layer that manages the adaptation of the change management layer and ensures its service continuity.

In order to minimize the time and the cost of the addition of the self-adaptation capabilities to this type of software studied (distributed and dynamic) we propose a framework based on the proposed architecture model. This framework implements the two uppermost layers of the architecture model. As we deal in this work the distributed applications, we propose that each site must contain two parts; the first represents a sub-system of the application, i.e. components implementing the application's business logic whereas the second represents the proposed framework that controls and manages the adaptation of the first part. Notice that, the management of the adaptation is distributed. This decentralization guarantees the desired degree of fault tolerance required in certain situations.

The remainder of the paper is organized as follows. Section 2 presents the proposed three layer architecture model for building the self-adaptive systems. Section 3 details the design of the proposed framework according to the proposed architecture model. In Section 4, we give the implementation details for a prototype of our framework and we illustrate the validation plan. Section 5 analyses the related proposals found in the literature. Finally, Section 6 concludes the paper.

2 Overview of the proposed architecture model

In this work we propose firstly a three layer architecture model that is used to guide the development of the dynamic and reliable distributed software. Figure 1 summarizes this model.

2.1 Application layer

The bottom layer of the proposed model is the application layer. It consists of a set of components implementing the application's business logic. As we deal the distributed applications these components are distributed on several sites. We propose that each functional component must have a component of type «*ComponentController*» which controls it. This last plays two roles: (1) if the controlled component is active, the «*ComponentController*» intercepts and redirects the incoming calls of service (to the controlled component) to the component «*ApplicationController*» of the fault-tolerant system (see section 3.2). In the contrary case where the controlled component is in a reconfigurable state, i.e. at the time of adaptation, its controller intercepts and saves the incoming calls of service to it in a queue until the end of the launched adaptation operation.

2.2 Change management layer

The middle layer of the proposed architecture model is the change management layer. This layer reacts to changes in state reported from the application layer. For that, it consists of two separate systems; the first is the fault-tolerant system which manages the reliability of the application and the second is the adaptation system which reconfigures dynamically the application. We will present these two systems in detail in the next sections. This separation of the fault-tolerant system from the functional code of the application and the code charged to reconfigure it facilitates the evolution of the reliability mechanism and thus, the development to the developers or integrators of the application which will concentrate on the functional code of the application rather on the non-functional code charged to reconfigure it and make it reliable.

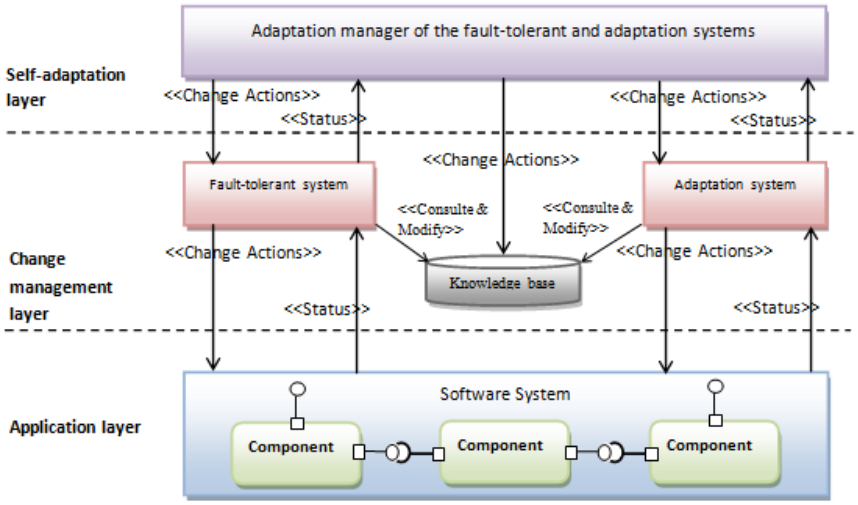


Figure 1: A three layer architecture model for self-adaptation.

2.3 Self-adaptation layer

The uppermost layer of the proposed architecture model is the self-adaptation layer. This layer introduces the self-adaptation capabilities to the framework itself. It controls and manages the change management layer for ensuring its service continuity and adapting its components to the changes that they carry out on the application in order to guarantee its correct operation and also its service continuity because certain changes in the application can lead to the appearance of faults in the execution of the system that manages these changes. For example, an operation of removal of a component in the application leads to the appearance of errors in the change management layer if the non-functional components managing the removed component have not adapted to this change.

Notice that, the proposed architecture is reflexive; the middle layer manages the bottom layer and the uppermost layer manages the middle layer. Also, this decomposition in three layers imposes a clear separation of concerns and facilitates the adaptation management as well as the evolution of the two mechanisms of fault tolerance and adaptation.

In order to facilitate the use of our architectural model we propose a framework implementing the two uppermost layers (self-adaptation and change management layers). So, the framework contains the two systems of adaptation and fault tolerance as well as the manager of these first two systems and which implements the self-adaptation layer. Therefore, an application developed according to our architecture model is made up of a set of functional and non functional components distributed on several sites. At each site we must find a sub-system (level of the application layer) which is a set of functional components representing the application's business logic plus an instance of the proposed framework, which is the responsible for the management of the application context (collection of data, analyses…) and the management of its change. So, the framework represents the hot subject of this paper. Figure 2 shows an overview of our solution for managing the distribution of the adaptation. For reasons of clearness, only two sites are represented.

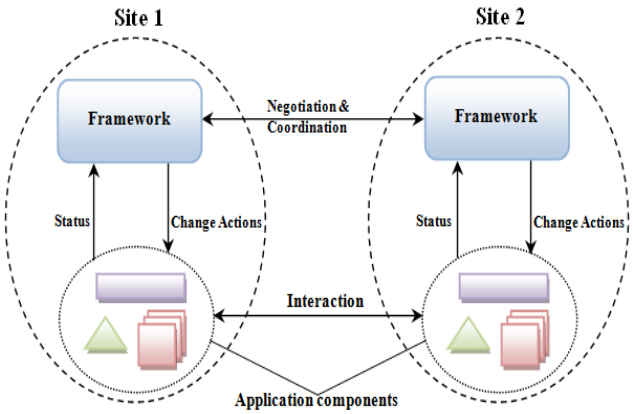


Figure 2: Overview of our solution for the management of distributed reconfigurations.

This organization makes the architecture of the self-adaptive applications developed according to our approach decentralized what avoids the problems of the centralized approaches [11].

In the next sections, we will present in detail the structure and the functioning of the different components of the two uppermost layers in the architecture model through the proposed framework.

3 Design and functioning of the proposed framework

This section describes in detail the various elements of the proposed framework thus that their functioning in order to perform the dynamic adaptation and preserve the consistency of the application. We present these elements according to their order of dependence.

3.1 Knowledge base

The knowledge base is a very important element in our framework since it plays a very significant role to provide reliable dynamic reconfigurations. For that, it is used by the different elements of the framework. It consists of three parts: (1) the description of software architecture, (2) the description of the adaptation policy and (3) the coherence rules. We propose the use of the logic of predicates with the language Prolog [32] for the description of these parts. This choice is justified by:

- Prolog is a language of knowledge representation.
- Prolog can be easily used for the description of the software architecture. We can write an XML tag (<tag>value</tag>) in Prolog as a fact as follows: tag (value).
- The representation of invariants (for the verification of the application consistency) by inference rules eliminates the programming of verification mechanisms of these invariants because this verification is performed by the inference engine of Prolog.
- The existence of Prolog interpreters developed in several languages, which facilitates the use of the prolog formalism.

3.1.1 Description of the software architecture

The description of the software architecture must contain:

- The detailed description of each application component.
- The specification of the component assembly.

3.1.1.1 Component description

component (‘id’, ‘name’).
component_state(‘comp_id’, ‘state’).

State: may be active or quiescent.

component_location(‘comp_id’, ‘ip_site’**).**
required_interface (‘comp_id’, ‘interface_id’).
provided_interface (‘comp_id’, ‘interface_id’).
interface (‘interface_id’, ‘name’).
include_operation (‘interface_id’,’operation _id’).
operation (‘id’, ‘name’, ‘list_param’, ’return_type’).
param (‘operation_id’, ‘name’, ‘type’, ‘value’).
component_property (‘comp_id’, ‘name’, ’value’).

3.1.1.2 Interaction between components

interaction (‘comp_id1’, ‘comp_id2’, ‘oper_id1’, ‘oper_id2’).

The ***interaction*** predicate specifies that the component *"comp_id1"* interacts with the component *"comp_id2"* where the operation *"oper_id1"* is required by the component *"comp_id1"* and the operation *"oper_id2"* is provided by the component *"comp_id2"*.

3.1.2 Application consistency

Parallel to the need of the dynamic reconfiguration of applications pose the problem of their reliabilities which is an important attribute of the functioning safety [6]. In fact, the modifications in a system can leave it in an incoherent state and thus challenge its reliable character. In order to guarantee the reliability of the system following a dynamic reconfiguration, we define the application consistency as the satisfaction of a set of constraints. These constraints are related to the definition of the architectural elements and their assembly and also to the state of the components.

We have used Prolog as a constraint language. So, we use the inference rules to express these constraints:

Example 1: Here is a rule to check if there are two components that have the same identifier:

haveSameID (Comp_name1, Comp_name2):-
Comp_name1 != Comp_name2,
component (Comp_id1, Comp_name1),
component (Comp_id2, Comp_name2),
Comp_id1=Comp_id2.

Notice that, the constraints vary from a component model to another and from an architectural style to another, for example there are models which authorizes the hierarchical structure and others not. The evaluation of these rules is made by the Prolog inference engine. The trigger of the evaluation of these rules is carried out by the two sub-components «*BehaviourChecking*» and «*StructureChecking*» of the component «*VerificationManager*» of the fault tolerant system (see section 3.2). Notice that, an operation of reconfiguration is valid only if the reconfigured system is consistent, i.e. if all the constraints in the knowledge base are satisfied.

3.1.3 Adaptation policy

One fundamental aspect in the software adaptation is the definition of the adaptation policy, i.e., the set of rules which guide the trigger of the adaptation according to the changes of the environment of the application and its components. These rules are in the form ECA, i.e. If (<Event> and <Condition>) then <Action>. The event part specifies the context change that triggers the invocation of the rules. The condition part tests if the context change is satisfied which causes the description of the adaptation (action) to be carried out.

We also propose the use of the inference rules to express the adaptation policy.

Example 2: Assume we have a software component that manages a cache memory. For this, it owns a property “*maxCache*” representing the maximum permitted memory space to save data into memory for faster processing. The following lines show an adaptation policy (described in Prolog) for a possible adaptation of this component.

rule1(Z):- **free_memory**(X), X> **2000**,
component_ property(*‘cacheHandler*, *’maxCache’*, Value),
Value<**10**, Z is "strategy1".

rule2(Z):- **free_memory**(X), X<**1000**,
component _property (*‘cacheHandler’*, *’maxCache’*, Value),
Value>**10**, Z is "strategy2".

strategy (‘strategy1’, “ [localhost] set_Value(‘*cacheHandler’*, *‘maxCache’*, 20) ”).

strategy (‘strategy2’, “ [localhost] set_Value(‘*cacheHandler*’, ‘*maxCache*’,5) ”).

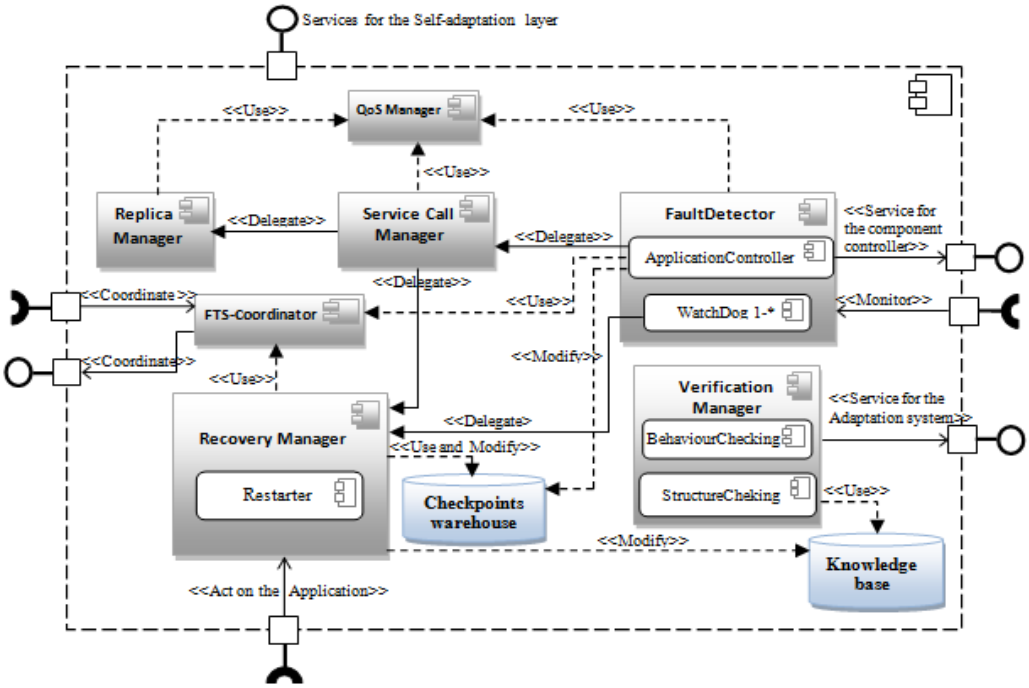


Figure 3: Overview of the fault-tolerant system.

The first rule is triggered only if the memory available exceeds 2Go and the maximum value of the cache is less than 10MB. In this case, the rule returns the string *'strategy1'* which indicates that it is necessary to apply the adaptation strategy number 1. This strategy contains in its action plan a single reconfiguration action which involves increasing the cache value to 20MB. Note that this operation concerns only the local site “*Localhost*”. The second rule is the reverse of the first. It involves decreasing the cache value to 5MB if the memory available is less than 1Go and the max cache value exceed 10MB.

3.2 Fault-tolerant system

For the definition of the fault-tolerant system, we consider a set of constraints which are: (i) modularity and adaptability of the system, (ii) extensibility of the system, (iii) taking into account of the distributed nature of the application to make it reliable as we deal here the distributed software systems.

This system ensures the application service continuity which helps to lead to reliable reconfigurations. We think that this system is very important in the self-adaptive applications because an adaptation operation cannot be executed on a component if it is crashed or in an inconsistent state. Also, the preservation of the application consistency is an important condition in the adaptation of software systems as mentioned in the introduction. We have separated this system to the adaptation system and the application’s business logic in order to integrate more than one fault-tolerance technique for ensuring the application consistency and to facilitate the evolution of this system without influencing either the adaptation system or the application’s business logic.

In order that this system achieves its goal, it contains a component for the management of the service quality, a fault detection component, a recovery component, a component for the verification of the application consistency, a component for the management of the replicas of the functional components, a component for the execution of the call of service plus a component for the coordination of distributed checkpointing and distributed recovery. Figure 3 shows an overview of the fault-tolerant system.

3.2.1 Techniques used in the fault-tolerant system

The proposed fault-tolerant system is based on the following techniques: distributed checkpointing, active replication, distributed backward recovery and message store. Our objective is to use these techniques for providing a fault-tolerant system able to tolerate many types of the software faults.

3.2.1.1 Distributed checkpointing

A common method for ensuring the progress of a long-running application is to checkpoint its state periodically on stable storage [23]. The application can be rolled back and restarted from its last checkpoint which bounds the amount of lost work that must be recomputed [23]. As we deal in this work the distributed applications, the coordination for the distributed checkpointing is a very important operation. In a coordinated checkpointing, processes coordinate their checkpointing activity so that a globally consistent set of checkpoints is always

maintained in the system. For that, we have used in our fault-tolerant system the two-phase commit distributed checkpointing protocol presented in [23].

The algorithm of this protocol is composed of two phases. The next paragraph describes the mapping of this algorithm to the fault-tolerant system of the proposed framework.

The running of this algorithm starts if a service request is launched by a functional component of the application. In this case, the controller of this component intercepts this call and delegates the execution to the sub-component «*ApplicationController*» of the component «*FaultDetector*» in the fault-tolerant system (see figure 3). This last asks the coordinator of distributed checkpointing (sub-component of the component coordinator) to launch a coordination so necessary for saving a checkpoint. Notice that, the checkpointing is performed periodically around an interval of time indicated by the component «*QoS-Manager*». So, if the time is passed the application controller asks the coordinator for the checkpointing to start coordination for checkpointing. In this case, the coordinator according to his policy decides if the safeguard of a checkpoint requires coordination or not. If the two components (client and server) depend on other components installed on other sites the coordination process starts.

In the first phase, the coordinator identifies initially the participants (components installed on other sites and depend on one of the two components client and server) of this coordination operation by using the predicate "*interaction*" presented above. For that, the coordinator asks the question "***? interaction (Cp, 'C_id', _, _).***" for the two components server and client such as 'C_id1' must indicate the identifier of the component concerned of this question, i.e. the client or the server identifier. After, the coordinator broadcasts a checkpoint request message to all participants. Every participant, upon receiving this message, stops its execution, flushes its communication channels, takes a tentative checkpoint and replies "*yes*" to the coordinator, and awaits the coordinator's decision. If a participant rejects the request for any reasons, it replies "*No*". If all participants reply positively, the coordinator's decision is to commit the checkpoints. Otherwise, its decision is the cancellation of the checkpoints. The coordinator's final decision marks the end of the first phase. Note that, the waiting time of the reception of the participants' response by the coordinator is fixed. If the coordinator does not receive a response of a participant for this period of waiting it regards it as "*No*".

In phase II, the coordinator sends its decision to all participants. If its decision is "*Save the checkpoints*," every participant removes its old permanent checkpoint and makes the tentative checkpoint permanent. Otherwise, participants reject the tentative checkpoint previously taken. Finally, each participant resumes its execution. The table 1 presents an overview of this algorithm.

Table 1: Overview of the distributed checkpointing algorithm.

Coordinator	Participants

Begin
If (the coordination for checkpointing is necessary)
Begin
/* Begin Phase I */
determine the participants;
request participants to take tentative checkpoints;
await all replies;
if (all replies = "*Yes*")
decide←"*Save the tentative checkpoints*";
else
decide ←"*Remove the tentative checkpoints*";
/* Begin Phase II */
send the decision to all participants;
end-if
End.
	Begin
	/* Begin Phase I */
	receive the coordinator request ;
	if (accept request)
	begin
	suspend communication;
	take a tentative checkpoint;
	reply *"Yes"* ;
	end-if
	else reply *"No"* ;
	await the coordinator decision ;
	/* Begin Phase II */
	if (decision = *"Save the tentative checkpoints")*
	begin
	remove the old permanent checkpoint;
	make the tentative checkpoint permanent;
	end-if
	else discard the tentative checkpoint ;
	resume communication ;
	End.

3.2.1.2 Active replication

The highly available services can be achieved by replicating the server components and thereby introducing redundancy [24]. If one server fails, the service is still available since there are other servers that are able to process the incoming requests. The active replication also called the state machine approach is one of the techniques allowing achieving such software-based redundancy [24].

In the active replication technique, clients send request to all the servers and it receives the common response to all servers. So, all servers execute all requests and end up in the same final state. Thus, at any given time it is likely that there is at least one server that can accept and process the incoming requests. In the active replication the crash of any server is transparent to the client [24]. We have used this technique in order to tolerate the faults in value in the application.

In the replication technique the components duplicated are generally those that are the more used in the application and these components are generally subject of the dynamic adaptation. So, preserving the continuity of service of these components is a very important task.

The replication has as a consequence a faster recovery of the failed components because the replicas are active and ready to process the incoming requests [28]. We have implemented this technique in the

component *«ReplicasManager»* of the fault-tolerant system (see section 3.2.2).

When the replication technique does not guarantee the masking of faults for the reason of the software crash (for example a problem into a component requires the search of a coherent state to continue processing) or for the reason of hardware problems, the recovery will be the best solution [28].

3.2.1.3 Backward recovery

The backward recovery consists to roll back the application in the case of failure to a previously saved state in order that it continues processing normally [22]. For that, a set of checkpoints must be saved each time that it is necessary. One problem with this technique is that the recursive execution of the backward-recovery process on a component can lead to the domino effect, i.e. that the component could be in its initial state losing all the work performed before the failure [25]. Among the techniques which avoid the domino effect is the coordination of the checkpointing that we have integrated in the fault-tolerant system.

One of the problems which can be posed in the management of the adaptation of distributed systems is the assurance of the message transmission of one process to another. For example, if a message concerning a request for coordination of the execution of an adaptation operation sent by a participant to another is lost, this leads to the cancellation of the adaptation even if the answer of the participant at the other site is positive what prevents the adaptation of the application to the new situation. To overcome this problem, we have used the message store technique described in the next section.

3.2.1.4 Message store

The message store [24] is a technique used for ensuring the message transmission of one process to another. It is a technique used in the mailing systems. According to this technique, the sender does not send the message directly to its destination. It sends it to an intermediate node representing a message queue handler. This latter saves the sender's message in the queue and it takes care of sending it to its destination. The sender is relieved from any additional concerns of message sending. If the recipient is down at the time when the sender sends the message, the message queue handler waits until the server comes up. Moreover, in the case when the message queue handler fails, the sender message remains in the queue and it will be sent to the destination when the message queue has recovered.

We have implemented this technique in the adaptation system for ensuring the message transmission between the negotiators of the adaptation strategy and the coordinators of the reconfiguration execution which are deployed at the different sites (see section 3.3). Also, we have implemented this technique in the fault tolerant system in order to ensure the transmission of messages between all the coordinators of distributed checkpointing and recovery at the different sites.

As the fault-tolerant system is separated from the adaptation system and the application itself, and as the implementation of this system is based on the component paradigm it is easy to add other techniques or to reuse this system or also to evolve it without touching the application or its adaptation system.

3.2.2 Presentation of the fault-tolerant system components

In this section, we present in detail the components of the fault-tolerance system and their functioning.

The component «*VerificationManager*». This component is responsible for the verification of the application consistency. It performs the verification of the conformity of the application components to their component model and architecture style. For that, it has two sub-components *«StructureChecking»* and *«BehaviourChecking»*: the first allows making a structural verification of the application whereas the second allows the verification of the behaviour of the application components. These two sub-components trigger the verification of the coherence rules contained in the knowledge base as explained in the section 3.1. For the verification of the components behaviour we considered only the verification of the component properties.

The component *«FTS-Coordinator».* As we deal in this work the distributed applications, the coordination for the distributed checkpointing and for the backward recovery in the case of faults or crashes is very important. For that, the fault-tolerant system has a component *«FTS-Coordinator»* for such coordination. In order that this component reaches its goal, it is composed of two sub components *«CheckpointingCoordinator»* and «*RecoveryCoordinator*». The first allows the coordination for the distributed checkpointing whereas the second allows the coordination for the distributed recovery. The sub-component *«CheckpointingCoordinator*» implements the protocol of the distributed checkpointing described previously. The protocol of the component *«RecoveryCoordinator»* will be presented in the next sections.

The component *«FaultDetector».* This component is responsible for: (1) the monitoring of the application components for detecting the faults which can appear in the application, and more precisely, the components' crashes and also (2) the reification of the calls of component services (i.e. the service request). For that, this component is composed of two types of component; components of type *«WatchDog»* and only one component of type «*ApplicationController*». The firsts are charge of the monitoring of the application components. They ping periodically the elements that they supervise for detecting the failed ones. If a component «*WatchDog*» detects that the component which it supervises is crashed, it calls the recovery function of the recovery manager for treating this fault.

The *«ApplicationController»* plays an important role in the fault-tolerant system. At the interception of a call

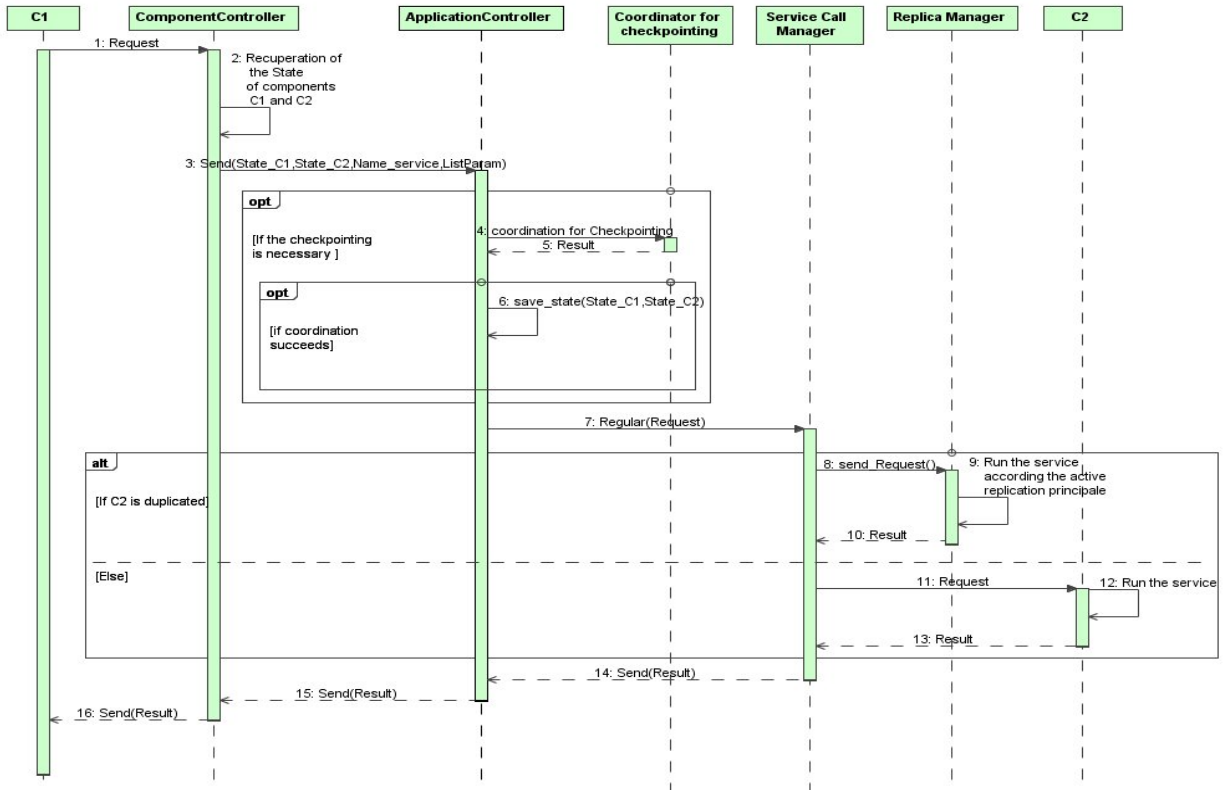


Figure 4: An execution without breakdown of a component C1 request by a component C2.

of service of one component by the corresponding controller, this last extracts the internal state of the two components client and server as well as the different parameters, and then it passes this information to the *«ApplicationController»*. This last, verifies if the checkpointing is necessary by checking if the time has passed compared to the last checkpointing such as this operation is performed periodically around an interval of time indicated by the component QoS-Manager. If the time is passed, the application controller asks the coordinator of distributed checkpointing to launch coordination so necessary for saving a checkpoint. After, it delegates the execution to the manager of the call of service. This last takes care of the processing of the service request.

The component *«ServiceCallManager»*. It is responsible for the management of the execution of the requests (i.e. calls of component services) submitted by the «*ApplicationController*». For the execution of these requests there are two cases: if the component server of the required service is duplicated the «*ServiceCallManager*» passes the execution of the request to the replica manager, which will manage the request processing according to the active replication technique. Otherwise, i.e. if the component server is not duplicated the manager of the execution of the call of service sends directly the request to this server and it awaits the reception of the execution result around a certain time indicated by the «*QoS-Manager*». If it does not receive a response after this waiting period it detects that the component provider of the required service is failed. In this case, it must call the recovery function of the recovery manager for tolerating the application to this fault.

The figure 4 presents a sequence diagram summarizing much more the functioning of the two components «*ApplicationController*» and «*ServiceCallManager*». This diagram explains the running of an execution without failure of a service request sent by a component C1 to a component C2.

The component *«ReplicasManager»*. It implements the active replication technique presented previously. This component is responsible for the execution of the service requests of the duplicated components. It executes the required service according to the active replication technique principle.

The component *«RecoveryManager»*. This component plays a very important role in the preservation of the application consistency. It treats the faults detected by the two components «*FaultDetector*» and «*ServiceCallManager*». The backward recovery technique is implemented in this component.

When a component «*WatchDog*» detects that the component which it supervises is failed or when the «*ServiceCallManager*» detects that the component provider of the required service is failed, it calls the recovery function of the recovery manager, which will carry out the backward-recovery of the failed component and also the components which depend on this last and which exist as well at the site of the failed component, or at the other sites. This distributed recovery is necessary

because we deal in this work the distributed applications, so, there is dependence between the distributed components which requires a recovery of all these components. Therefore, coordination for distributed recovery is necessary. For that, the component «*FTS-Coordinator*» has a sub-component «*RecoveryCoordinator*» that performs such coordination. This component has a specific protocol which we have proposed.

The basic idea behind a protocol for a distributed recovery is to ensure that all components depending on the failed component roll back to their previous coherent states. The set of the realized local recoveries must form a coherent global state of the application.

The algorithm starts with an initiation of a request for coordination to the recovery coordinator by the recovery manager for recovering all the components which are distributed on the other sites and which depend on the failed component (if they exist). In this case, the coordinator according to his policy decides if the recovery requires coordination or not. If the failed component depends on other components installed on other sites, the coordinator invites the participants (which are the recovery managers of the fault-tolerant systems of the instances of the proposed framework and which are installed on the other sites) to perform the rollback towards the last saved checkpoint for the components which depend on the failed component. If a participant rejects the request for any reasons, it replies *"No"*. Otherwise, the participant performs the recovery and it replies "*Yes*". If all the participants' reply "*Yes*", the communication stops and the coordinator announces the success of the recovery.

If there is one or more participant replying by "*No*", the coordinator will wait a certain time, then, it will send again a request for recovery to the participants who replied by "*No*" in order that they perform another time the recovery of the components which depend on the failed component. The basic idea behind this waiting before sending the recovery request again to the components replying by "*No*" is that these components can return to operate (reception of the requests) after this waiting period because they were for example in the process of running a reconfiguration operation or a critical operation (e.g. an operation on the database). The table 2 presents an overview of the proposed distributed recovery algorithm.

Notice that only one operation of recovery coordination can be carried out at the same time and this is for guaranteeing the application coherence.

The component «*QoS-Manager*». It is a component used for managing the service quality level in the application. This component allows to the user to change a set of parameter through a graphical interface in order to increase or decrease the level of the quality of the service in the application. These parameters are: the waiting time of the execution result of a request by the *«ServiceCallManager»*, the interval of time during which a checkpointing is performed, the interval of time during which a component *«WatchDog»* ping the component which it supervises, the interval of time during which the component *«CaptureContext»* supervises the application environment and the max number of replicas of each type of application's component.

Table 2: An overview of the distributed recovery algorithm.

Recovery coordinator	Participants

begin
request participants to perform the recovery of the components depending on the failed component ;
await all replies ;
if (all replies = "*Yes"*) stop the coordination ;
else
begin
await a certain time ;
request the participants who replied by *"No"* to perform the recovery ;
stop the coordination ;
end-else
 	
End.
	Begin
	if (accept request)
	begin
	perform the components recovery;
	reply "*Yes"*;
	end-if
	
	else
	reply "*No"*;
	
	End.

3.2.3 The fault model

The use of the four techniques (active replication, message store, distributed backward recovery and distributed checkpointing) allowed us to propose a powerful fault-tolerant system for the proposed framework able to tolerate many types of failure. For the components crash, the proposed fault-tolerant system is able to detect them via the components «*WatchDog*». Each component in the application sends periodically a heartbeat message to its monitor «*WatchDog*» and this last periodically checks the heartbeat. If the heartbeat message from the supervised component is not received by a specified time, the component «*WatchDog*» assumes that the supervised component is hung. This problem will be treated by the recovery manager. The faults of type omission are treated in our approach via the message store technique which ensures the transfer of messages from an entity to another. The faults of type *"late timing"* are detected by the component *«ServiceCallManager»* such as each type of request has an interval of result waiting indicated by the component «*QoS-Manager*». If time passes and the manager of the calls of service has not received a response, it detects that there is a problem into the component provider of the service. This problem will be treated by the recovery manager as explained in the previous section. The faults in value require for their treatment the existence of several replicas. The active replication technique which we have incorporated in the fault-tolerant system allows treating this type of faults because a client request is sent to all the servers. If a response from a server is different to the majority of servers' response, this server has a fault of value. As we deal in this work the distributed

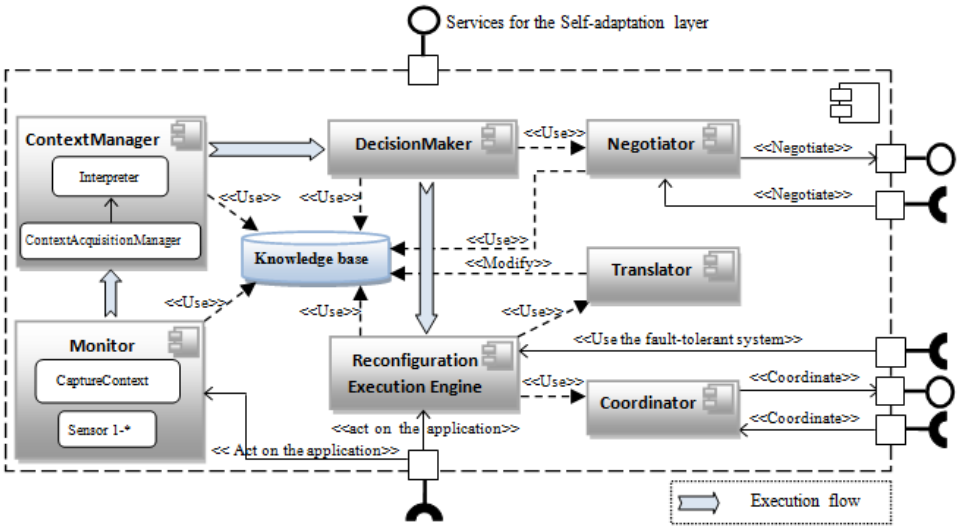


Figure 5: Overview of the adaptation system.

applications which are by nature complexes, the creation of several replicas of each functional component of the application for the treatment of the faults in value is impossible. For that, we propose to duplicate only the components the more used in the application and which are generally a subject of adaptation.

3.3 Adaptation system

This section presents in detail the adaptation system and its functioning for realizing distributed and reliable reconfigurations.

For the definition of the adaptation system, we consider a set of constraints which are: (i) taking into account of the distributed nature of the software to make it adaptable, (ii) reliability of the distributed reconfigurations, and (iii) flexibility and adaptability of the adaptation system.

The proposed adaptation system is designed according to the classical autonomic control loop MAPE-K (Monitoring, Analysis, Planning and Execution) [3], which is the most common approach for self-adaptive systems [5, 33].

So, in our adaptation system we have implemented the elements of this loop as separate components. The monitoring, analysis and adaptations are performed by the MAPE-K control loop. A significant part of the negotiation of adaptation strategy and coordination of reconfiguration execution were externalized from the control loop. Moreover, we have chosen to merge the analysis and plan components because a significant part of these components' logic is externalized from the components and stored in the knowledge base (Prolog script). Therefore, not leaving too much of analysis and planning to be performed within those two components.

Plus the set of components that implement the MAPE-K loop, the adaptation system contains a component *«Negotiator»* which negotiates an adaptation strategy with its similar at the other sites, a component «*Coordinator*» that coordinates the execution of reconfiguration actions, and a component «*Translator*» which executes the reconfiguration actions of the adaptation strategy on the architectural representation of the application. The figure 5 shows an overview of this system.

We propose the implementation of the whole cycle of the MAPE-K loop as a chain of responsibility pattern [13]. We have proposed to use this pattern because the processing is distributed on several objects (components of the adaptation system). When a component finishes its processing, it passes the execution to the next component. Moreover, it is easy to vary the components involved in the processing which makes the adaptation system more flexible.

3.3.1 Monitoring

The «*Monitor*» is the first component in the chain that comprises the control loop. It is responsible for periodically collecting information of the managed elements (i.e., sub-system of the application managed by the adaptation system) and of the execution of the application (CPU consumption, memory usage, bandwidth, service calls per minute). To achieve this goal, the monitor has a sub-component «*CaptureContext*» that collects information about the application execution plus a set of sub-components «*Sensor*» that collect information about the set of the application components at its site. These two sub-components of the monitor pass the collected information to the next object that is part of the execution chain, next to the context manager.

The «*ContextManager*» is the second component in the sequence of the responsibility chain. It interacts with the sensors associated with the execution environment and the application for collecting the information needed to characterize the execution context. For that, it has two sub-components «*ContextAcquisitionManager*» and «*Interpreter*». The context acquisition manager gathers the information collected by the sensors of the «*Monitor*»

and saves them in the knowledge base. After, it delegates the execution to the «*Interpreter*». This last, interprets data provided by the «*ContextAcquisitionManager*» in order to provide a significant contextual data. Notice that, the received data are separately interpreted for each type of measurement in order to provide a significant contextual data. If a suitable context change is detected the «*Interpreter*» notify the decision maker (see next section) of this change as this last subscribes to events near the context manager.

3.3.2 Analysis

The aim of the analysis phase is to see whether a reconfiguration action is required or not. For that, the decision maker component «*DecisionMaker*» is the third component in the sequence of the responsibility chain. It plays the role of the analysis and plan phases in the MAPE-K control loop. This component is responsible for taking decisions on adaptation. It provides in output an adaptation strategy that will be executed in the execution phase of the control loop.

3.3.2.1 Negotiation process

As we deal in this work the distributed reconfigurations, the negotiation is a significant step in the decision-making on adaptation. It is a cooperative process in which a group of adaptation systems reach an agreement on a comprehensive adaptation strategy. We define a global strategy as a set of strategies that the decision makers of the different adaptation systems choose during the negotiation process. Noting that, the negotiation must guarantee the independence in the decision-making of each «*DecisionMaker*» and ensure the global validity of a local decision.

The negotiation is started by the initiating decision maker. This last chooses an adaptation strategy. Then, it asks its negotiator to negotiate this chosen strategy. This negotiator proposes simultaneously to each participant the strategy that the initial decision maker has chosen. The negotiator of each participant receives the strategy and interprets its policy to reason on its applicability. It can then accept, refuse or propose a modification of the strategy; and then, it answers the initiating negotiator. When this last receives all answers, it thinks on the acceptances and/or the applicability of the modifications asked. When all the participants accept the strategy, the negotiation succeeds. Otherwise, it detects and solves the conflicts and can then, in its turn, propose a modification of the strategy. The negotiation process is stopped if one negotiator refuses a strategy or if a stop condition is checked. This condition is in connection to the authorized maximum time of negotiation or with the maximum number of negotiation cycles. If the negotiation succeeds, the initiating negotiator returns to the initiating decision maker the strategy resulting from the negotiation and sends to the negotiator of each participant the final strategy. If the strategy resulting from the negotiation is a new strategy, i.e. not exists in the adaptation policy, the decision maker adds it to the knowledge base and precisely to the adaptation policy part. This operation allows enriching the knowledge base with new adaptation rules in order to better adapt to the new changing situations. At the reception of this strategy, each participant (i.e. negotiator) asks to its decision maker to adopt the strategy resulting from the negotiation and delegates the execution to the next object in execution chain that is the reconfiguration execution engine. In the opposite case (i.e. negotiation failure), the initiating decision maker and participants are informed of the negotiation failure. Otherwise, the adaptation is cancelled and the loop cycle is stopped.

3.3.3 Execution

In order to increase the reliability of the reconfigurations executed by our framework we have used the transaction technique. This technique was originally used in the system managing databases [14]. Their use is widespread in all computer systems where there is a need to maintain the consistency of the information in spite of concurrency and the occurrence of failures. The transactions are thus a means to make systems fault-tolerant. A transaction consists to carry out a coherent computing operation consisting of several actions. The operation will be valid only if all its unit actions are carried out correctly. So, we speak about the commit. Otherwise, all data processed during the running of the operation must be returned to their initial state to cancel the transaction. So, we speak about the rollback.

We have used the transaction technique to define transactional reconfigurations.

According to the transactions principle each transaction is made up of a set of primitive operations. So, in our context an adaptation operation *Adop* is a transaction when its primitive operations are the primitive reconfiguration actions *Prac*. For example, the replacement operation of a component *C1* by another component *C2* is made up by the following primitive actions: stopping the component *C1*, creation of a new instance of the component *C2*, transfer of the *C1* state to the new instance of *C2* for preserving the application consistency and finally the start of the new instance of *C2*. The component replacement in our framework is carried out similarly with the work in [35].

We define the evolution of a component-based system by the transition system < C, Adop, → >:

- $C = \{C_0, C_1, C_2, \ldots\}$ *a set of configurations,*
- *Prac* $\in$ *{Instantiation/Destruction of component, Addition/Removing of component, modification of the component attribute value, modification of the life cycle of a component and adding of new behaviours}*
- *Adop is a set of Prac,*
- $\rightarrow \subseteq C \times Adop \times C$ *is the reconfiguration relation.*

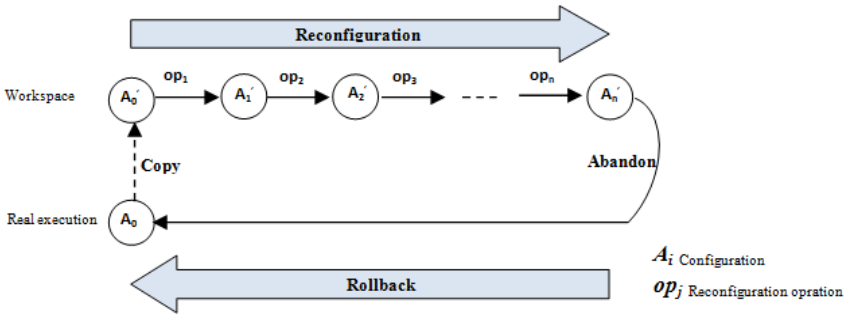


Figure 6: Abandon model of an adaptation operation.

3.3.3.1 Reconfiguration actions of the proposed framework

In our framework the dynamic reconfigurations are based on the following primitive actions:

- *instantiation/destruction of components*
- *addition/removal of components*
- *starting/stopping of components*
- *setting name of components*
- *setting the operating parameters of components or combinations of them.*

Thus, an adaptation strategy consists of a set of adaptation operations, where each operation is composed of at least only one primitive reconfiguration action.

The behaviour of each application component is generally statically encoded. However, changes in the application context, changes in use, changes in resource availability or the appearance of faults in the system, may require further abilities [10]. For this, it is very important to introduce dynamically the ability to add new behaviours to the application's components. The AOP (Aspect-Oriented Programming) and scripting languages are two techniques used for this end. In the AOP and with the runtime weaving, the binding between the logic code and aspect is done during the execution. The advantage of runtime weaving is that the relationship between the functional code and the aspects can be dynamically managed. Nevertheless, the use of the AOP for adding new capabilities to the system has one disadvantage is that the software system could be in an inconsistent and/or unstable state [10].

For the scripting languages, they allow the incremental programming, i.e. the possibility of running and developing simultaneously the scripts [8]. With these languages we can modify the code of a component without stopping it. Therefore, with this technique the addition of a new behaviour is much easier than the use of the AOP technique, but these languages are not powerful as the compiled ones. For that, the developer must find a compromise between the use of scripting languages and the AOP technique in order to improve the performance of the application. The implementation of the mechanism carrying out the addition of new behaviours to the application's components is left to the developer.

3.3.3.2 Quiescence management

Obtaining a reconfigurable state also called quiescent state [17, 18] is a very significant step in the reconfiguration process since it helps to ensure the application consistency in the case of reconfiguration. A reconfigurable state is a state from which a reconfiguration action can be performed without affecting the consistency of the application. For that, in our approach a reconfiguration action is carried out on a component if this last is in a reconfigurable state.

To search for a reconfigurable state, we have integrated in the proposed framework Wermelinger algorithm [18] which extend Kramer/Magee's algorithm proposed in [17]. Wermelinger proposes to block only connections between the components implied in the reconfiguration. An advantage of this algorithm is that interruption time is minimized while only affected connections must be blocked in contrast to whole components.

3.3.3.3 Algorithm of the reconfiguration execution engine

The reconfiguration execution engine is the fourth component in the execution chain. It undertakes the execution of the adaptation strategy proposed by the component «*DecisionMaker*». Firstly it (1) launches a search for a reconfigurable state before running the reconfiguration actions. Then, it (2) triggers the execution of the reconfiguration actions of the strategy. Notice that, we simulate firstly the execution of the reconfiguration on the architectural representation of the application. If no fault is detected, we execute the reconfiguration on the running application. Therefore, the effects of reconfigurations are not directly applied on the system which facilitates the cancellation of the reconfiguration in the case of its execution failure. This operation will be simply the removing of the work copy which has been used for the simulation of the reconfiguration. Figure 6 shows our abandon model of an adaptation operation.

As the reconfiguration of distributed application is a global reconfiguration process composed of distributed local reconfiguration processes, the proposed adaptation system incorporates a component for the coordination of the reconfigurations execution.

Following the running of each primitive reconfiguration action on the architectural representation the reconfiguration execution engine (3) carries out the verification of the consistency of the application structure and the verification of the validity of the behaviour of its components via the component «*VerificationManager*» of the fault-tolerant system. If a constraint is violated, the adaptation operation must be stopped for preserving the consistency of the application. In this case, the reconfiguration execution engine removes the copy of work used for the simulation of the reconfiguration. Moreover, this initiating reconfiguration execution engine notifies its coordinator of the execution failure of the primitive reconfiguration action in question. This last, notifies the other participants (coordinators of the reconfigurations execution deployed at the other sites) of this failure so that they can cancel the adaptation operation at their level in order to preserve the global consistency of the application.

In the opposite case, i.e. where the «*VerificationManager*» does not detect any error following the running of a primitive action of reconfiguration, the reconfiguration execution engine sends a message *"ApplyNextAction"* to the coordinator of the execution of reconfiguration actions. This last awaits the reception of all the participants' responses. Notice that, this waiting time of the participants' responses by the coordinator of the initiating reconfiguration execution engine is fixed. If this coordinator does not receive a message of a participant during this waiting period, it regards it as "*reconfiguration failure*". If one participant replies by "reconfiguration failure" the coordinator announces the failure of the execution of the reconfiguration. Otherwise, it asks for the participants to perform the next primitive reconfiguration action and the process is still repeated. If all the adaptation operations of the strategy are executed on the architectural representation of the application (copy of the work) without faults, the reconfiguration execution engine (4) runs these actions on the running system. Also, the copy of work used for simulating the execution of the reconfiguration is saved as the new architectural representation of the application. Notice that, this operation ensures the conformity of the architectural representation of the application to its system in running and it has many advantages. For example, it facilitates the comprehension of the software through its architecture, and thus its evolution because the architectural representation always conforms to the system. Finally, the reconfiguration execution engine (5) unblocks the connections blocked during the phase of searching for a reconfigurable state. The end of the execution of this operation determines the end of the control loop cycle. The running of the reconfiguration execution engine is summarized by algorithm 1.

Algorithm 1

$pact_j$ is a primitive reconfiguration action

```
1:  Begin
2:  SearchForReconfigurableState();
3:      For all pact_j ∈ startegy do
4:      RunPactOnArchRep(pact_j); /* execute the primitive reconfiguration action pact_j on the architectural representation via the component
5:                                   «Translator» */
6:      if not IsConsistentApplication () then
7:              SendMessageToCoordinator ("reconfiguration failure");
8:              RemoveWorkCopy(); // Removes the working copy used for the reconfiguration simulation
9:              BREAK; // to exit the loop "for"
10:     end_if
11:     else //case where the application is consistent
12:             SendMessageToCoordinator ("ApplyNextAction");
13:             response←coordinator.CoordinationDecision();
14:             if response != "ApplyNextAction" then
15:                 SendMessageToCoordinator("reconfiguration failure");
16:                 RemoveWorkCopy(); // Removes the copy of work used for the simulation of the reconfiguration
17:                 BREAK; // to exit the loop "for"
18:             end_if
19:     end_else
20:   end_for
21: if all actions pact_j in strategy are executed // If the reconfiguration is succeeds
22:     RunAllpactOnSystem(); // execute the primitive reconfiguration actions on the Running system
23:     SaveChanges(); // Save the performed changes on the architectural description
24: end_if
25: ReactivateConnections(); // unblock the blocked connections
26: End.
```

3.4 The adaptation manager of the fault-tolerant and adaptation systems

This manager implements the self-adaptation layer of the proposed architecture model. It manages the two systems of adaptation and fault-tolerance and looks after the adaptation of these two systems to the changes that they carry out themselves on the application components in order to ensure the correct operation of the global application. This manager allows (1) to replace the negotiator of adaptation strategy or the reconfiguration execution coordinator or also the recovery and checkpointing coordinator of the fault-tolerant system by other components if they crash in order to ensure a good management of the application changes. For that, this manager has a set of monitors of the type «*WatchDog*» which monitor these components. Also, it allows (2) in the case where an operation of removing of an application component is carried out, to stop and remove the two components: «*Sensor*» of the adaptation system and «*WatchDog*» of the fault-tolerant system that monitor this component to avoid the introduction of errors into the running of the application. This manager also carries out (3) the update of the two Prolog scripts representing the adaptation policy and the coherence rules by removing the facts representing the adaptation strategies and the coherence rules which are in relation with the removed components.

4 Implementation and validation plan

In this section, we give details and technical choices made to implement a prototype of the proposed framework. We also present the validation plan of this framework.

4.1 Background

For the implementation of the elements of the proposed framework, we have used the two component models EJB (Enterprise Java Beans) and ScriptCOM. So, the implementation of this framework is divided into two parts; one part implemented with EJB and the other with ScriptCOM. EJB [4] is an industrial model; we have used it because it is based on Java that is a powerful programming language that meets our implementation needs (support for AOP, support for native codes via JNI API, Java COM Bridge, support for the remote method invocation via RMI API, an API to access system information like SIGAR[1] and a support for multi-threading).

ScriptCOM [8] is a component model extension of COM (Component Object Model) [2] allowing the dynamic adaptation of the COM components. It allows the development of adaptable scripting components. Notice that, with the scripting languages we have the possibility of developing and running simultaneously the scripts which represents the component's implementation [8]. In this model, the component adaptation is carried out through three controllers which are: the interface controller, property controller and script controller. Moreover, as it is an extension of the COM model it benefited from the advantages of this latter (support of the distributed applications, independence of the programming languages, versioning...). We have used this model in order to facilitate the adaptation of the proposed framework. We think that dynamic inclusion and removal of adaptation management concerns allows the improvement of adaptation to the evolving needs without stopping the entire framework.

4.2 Framework implementation

We have designed a set of software components that implement the different elements of the proposed framework. The implementation of the components of type «*ComponentController*» and the different sensors as well the effectors (part performing the execution of the reconfiguration and backward recovery) are realized with EJB model. The coordination for reconfiguration execution and backward recovery and also negotiation parts are implemented with the two models EJB and ScriptCOM. The rest part of the framework is developed as a set of ScriptCOM components that we can add, remove or update at runtime. This is just one of possible implementations and particularly, this has been designed to provide self-adaptable capabilities to the framework.

For the implementation of the controllers «*ComponentController*» of the functional components of the application we choose the use of the aspect oriented programming. So, the implementation of each controller is based on an aspect. This aspect has a generic pointcut that intercepts all the incoming service calls to the controlled component and treats them as explained in the section 2.1.

For the knowledge base, i.e. the architectural representation of the application, adaptation policy and coherence rules description we have used the language Prolog as explained in the section 3.1. We have used the JPL[2] library which uses the SWI-Prolog foreign interface and the Java JNI interface providing a bidirectional interface between Java and Prolog that can be used to embed Prolog in Java as well as for embedding Java in Prolog. Also, we have used another interpreter Prolog[3] developed with the JavaScript language in order that it will be used with the part of the framework developed with ScriptCOM. The three elements of the knowledge base are contained in separate scripts which facilitate their modifications at runtime. We can then add, remove or change rules or facts in the knowledge base without stopping the framework.

Our framework is independent from particular component models. Therefore, elements of the application layer, i.e. components implementing the business logic of the application can be developed using any component models. The implementation for a

[1] https://support.hyperic.com/display/SIGAR/Home

[2] http://www.swi-prolog.org/packages/jpl/

[3] http://ioctl.org/logic/prolog-latest

specific component model is made with the least effort in the part developed using the model EJB and without changing the main adaptation concepts.

4.3 Validation plan

The objective of the validation in this paper is to test the influence of the proposed framework on the application response time and the adaptation time. These criteria are measured with randomly generated configurations which we have developed using the model EJB. The components of this application execute a few arithmetic operations and they are distributed on two sites. The evaluation test is made by comparing two versions of the same application; one incorporates the proposed framework, the other one without this framework. All the experiments were run on Intel Core 2 Duo CPU T5670 workstations with 1.0 GB DDR2 memory and Windows XP SP3 as the operating system.

The first evaluation consists to test the influence of the proposed framework on the application response time. This test is made by comparing the running of a certain number of requests on the two versions of the application (without and with the proposed framework). The Table 3 shows the response times before and after the incorporation of the framework.

We have calculated the response time increase in the version which implements the proposed framework and we found that the overhead for functional method calls is about 34% of the overall execution time.

Table 3: Increase rate of the response time.

Request Numbers	Response time average : without the Framework	Response time average : with the Framework	Increase rates of the response time
100	16.91 ms	22.87 ms	**35.21 %**
200	33.39 ms	44.48 ms	**33.20 %**
300	50.04 ms	67.31 ms	**34.51 %**
500	83.38 ms	111.18 ms	**33.33 %**
700	115.39 ms	156.1 ms	**35.25 %**
900	150.16 ms	205.67 ms	**36.96 %**
1000	169.83 ms	222.14 ms	**30.80 %**
Average			**34.18 %**

The second evaluation consists to measure the adaptation time, which is calculated as follows:

$$\mathbf{T}_{adaptation} = \mathbf{T}_{rspWAdap} - \mathbf{T}_{rspNAdap}$$

Where: $\mathbf{T}_{adaptation}$ is the adaptation time.

$\mathbf{T}_{rspWAdap}$**:** is the response time with adaptation.

$\mathbf{T}_{rspNAdap}$**:** is the response time with the proposed framework but without adaptation.

Table 4 shows the obtained results. The adaptation time average is approximately 430,95ms. Certainly, this figure is very large compared to the response time of one request, which is approximately 0,22ms (response time of an execution of one request with the framework).

Table 4: Adaptation time.

Request Numbers	Response time average : with the Framework but without Adaptation	Response time average : with the Framework and with Adaptation	Adaptation time
500	111.18 ms	531.39 ms	**420.21 ms**
700	156.1 ms	592.56 ms	**436.46 ms**
800	179.71 ms	639.13 ms	**459.42 ms**
900	205.67 ms	631.78 ms	**426.11 ms**
1000	222.14 ms	634.68 ms	**412.54 ms**
Average			**430.95 ms**

The obtained adaptation time is great, but it is acceptable because the adaptation is distributed (at two sites in this test application) which requires a negotiation of the adaptation strategy and a coordination for the execution of the reconfiguration actions and this influence the adaptation time. Moreover, as we have used the component model ScriptCOM for the development of one part of the proposed framework, this, influence the application response time and also the adaptation time because the implementation of the components of this model is based on the language Jscript[4], which is an interpreted language. So, the execution will be slower than the compiled versions. Notice that, we have used this model in order to facilitate the adaptation of the proposed framework.

Finally, we can say that the obtained results confirm that our framework is very suitable for developing distributed applications where we prefer the reliable dynamic adaptability more than performance.

5 Related work

The problem treated in this paper accosts the domain of research around the dynamic adaptation of the computing systems and in particular the distributed component-based systems.

In terms of model-based approaches Kramer and Magee [15] have proposed layered reference architecture for self-adaptive software. The bottom layer of this architecture is the component control layer which contains the system's application-level functionality. The change management layer is the middle layer. It manages the changes of components state or environment. For that, it contains a set of pre-compiled plans to deal with the different situations encountered by the system. The uppermost layer is the goal management layer which

[4] http://msdn.microsoft.com/fr-fr/library/hbxc2t98%28vs.85%29.aspx

generates new plans if none of the existing plans can address the current situation, or a new system goal is introduced. Also, we have proposed a three layer architecture model where the bottom layer is the application layer similar to the component control layer in the Kramer and Magee model's. Unlike this model, the change management layer of our model contains two systems managing separately the adaptation and the fault tolerance of the application layer. Moreover, to the difference of the uppermost layer of the Kramer and Magee model's, the uppermost layer of our model introduces the self-adaptation capabilities to the framework itself. It ensures the service continuity of the change management layer and manages the adaptation of this layer to the changes which it carries out itself on the application layer. Moreover, our architectural model is reflexive. In the Kramer and Magee model's, the distributed reconfigurations are possible through the decentralized architecture of the change management layer implementation proposed by the authors. From a reliability point of view, Kramer and Magee have expressed that a server failure is a predicted state change and the change management layer must include a procedure for dealing with the change. For that, they propose the use of the repairing strategy of the faults described by Garlan and Schmerl in [27] as a plan executed by the change management layer.

Several research activities [7, 9, 12] implement the autonomic control loop to dynamically reconfigure systems. For example, in [7] the authors use a component-based approach for modelling a framework that provides flexible monitoring and management tasks and allow introducing adaptation to component-based SOA applications. The framework implements the different phase of the autonomic control loop. The main purpose of the authors is to build a framework supporting heterogeneous components implementing the MAPE-k phases as SCA components. This framework supports the development of distributed applications, but it doesn't support to perform distributed reconfigurations while our framework is conceived especially for doing such type of reconfigurations.

A3 [10] is a framework for developing distributed systems that need adaptive features. A3 provides robust mechanisms of coordination that the components can use to share their own knowledge and knowledge of the system to which they belong. The framework itself is self-adaptable. A3 exploits the idea of *group* to organize a system in a set of independent partitions, and reduces the communication problem. From an adaptation point of view, A3 supports the distributed adaptations and it allows indeed interesting adaptations. This framework does not have any mechanism to preserve the reconfiguration reliability. It treats only the fault of type messages omission. Moreover, a reconfiguration action is executed at the system directly, i.e. without reaching a reconfigurable state before the execution of such action.

Huynh et al. [20] propose a platform supporting distributed reconfiguration of the component-based applications. This platform integrates a solution for the management of system states at reconfiguration time. The authors define different system states regarding reconfiguration and ways that the system will act accordingly. This platform allows to correct reconfiguration plans if a disconnection is detected during the reconfiguration in order to continue the reconfiguration if possible, or recover if the reconfiguration fails. It also allows the coordination of the distributed reconfiguration actions. In contrast, to this platform, our framework integrates a negotiation mechanism which allows the negotiation of the adaptation strategy before the coordination of its execution that is a very important point in the distributed reconfiguration process.

In [21], a transactional approach is proposed to ensure reliable reconfigurations in the context of component based systems and particularly in the Fractal component model. To ensure atomicity of reconfiguration transactions, operations performed in transactions must be cancelled if a fault occurs before the end of the reconfiguration. This operation of cancelation of the reconfiguration operations effect is carried out by the execution of the reverse action of each reconfiguration operation performed because the reconfiguration operations are carried out directly on the system. In contrast to this approach, we have proposed to carry out firstly the reconfiguration on the architectural representation of the application which facilitates the cancelation of this operation if there is a problem. From a reliability point of view, the authors propose the use of the integrity constraints to define the system consistency for guaranteeing the respect of these constraints at runtime.

6 Conclusion

In this paper, we have presented a framework for building distributed and dynamic component-based systems. The proposed framework is based on a reflexive three layer architecture model which we have proposed. The bottom layer of this model is the application layer. It contains the system's application-level functionality. The change management layer is the middle layer. It manages the changes of the bottom layer. The uppermost layer is the self-adaptation layer that introduces the self-adaptation capabilities to the framework itself. It ensures the service continuity of the change management layer and manages the adaptation of this last to the changes which it carries out itself on the application layer. The proposed framework implements the two uppermost layers of the proposed architecture model and it is based on a decentralised architecture. It incorporates two separate systems that manage the dynamic adaptation and fault tolerance of the application components and also, an adaptation manager implementing the self-adaptation layer in the architecture model. The proposed framework is designed especially to support the distributed reconfigurations. For that, it incorporates a robust coordination and negotiation mechanisms for managing this type of reconfiguration. The adaptation system of this framework is designed according to the classical autonomic control loop MAPE-K which allows a better

management of the adaptation. As the preservation of the application consistency is an important point in the dynamic reconfiguration, the framework incorporates a separate fault-tolerant system implements four fault tolerance techniques (distributed checkpointing, active replication, message store and distributed backward recovery) which makes it able to tolerate most of faults type. Also, as the adaptation operations in this framework are executed as transactions, this increases the reliability of these operations. A prototype of this framework has been implemented using two component models; EJB an industrial model and ScriptCOM a component model for developing adaptable components, which facilitates the adaptation of the proposed framework.

However, the evaluation of the proposed framework has revealed that the adaptation time is long, for that we plan to improve the adaptation system of the proposed framework in terms of performances.

In the long term, we want to study the possibilities to extend our solution to support dynamic adaptation of other kinds of applications like web services.

References

[1] R. N. Taylor, N. Medvidovic and P. Oreizy (2009). Architectural styles for runtime software adaptation. *In 3rd European Conference on Software Architecture (ECSA)*, pp. 171-180.

[2] Microsoft Corp, "Component Object Model" accessed on July 30, 2013. [Online]. Available : http://www.microsoft.com/COM

[3] IBM. An architectural blueprint for autonomic computing. Autonomic computing whitepaper, 4th edition. 2006.

[4] V. Matena and M. Hapner (1999). Enterprise Java Beans Specification v1.1- Final Release. *Sun Microsystems*.

[5] B. H. Cheng et al. (2009). Software Engineering for Self-Adaptive Systems: A Research Roadmap. *In B. H. Cheng, R. Lemos, H. Giese, P. Inverardi, and J. Magee, editors, Software Engineering for Self-Adaptive Systems*, volume 5525 of Lecture Notes in Computer Science, pp. 1-26. Springer.

[6] M. Léger, T. Ledoux and T. Coupaye (*2010)*. Reliable dynamic reconfigurations in a reflective component model. *In Proceedings of the 13th international conference on Component-Based Software Engineering, pp*. 74-92.

[7] F. B. Cristian Ruz and B. Sauvan (2011). Flexible adaptation loop for component-based SOA applications. *In Proceeding of the Seventh International Conference on Autonomic and Autonomous Systems*, pp. 29–36.

[8] O. Aissaoui and F. Atil (2012). ScriptCOM an Extension of COM for the Dynamic Adaptation. *In Proceedings of IEEE International Conference on Information Technology and e-Services (ICITeS'12)*, pp. 646-651.

[9] Y. Maurel, A. Diaconescu, and P. Lalanda (2010). Ceylon: A service-oriented framework for building autonomic managers. *In Proceedings of the Seventh IEEE International Conference and Workshops on Engineering of Autonomic and Autonomous Systems*, pp. 3 –11.

[10] L. Baresi and S. Guinea (2011). A3: self-adaptation capabilities through groups and coordination. *In Proceedings of the 4th India Software Engineering Conference*, ISEC'11, pp. 11-20.

[11] C. Tan and K. Mills (2005). Performance characterization of decentralized algorithms for replica selection in distributed object systems. *International Workshop on Software and Performance*, pp. 257-262.

[12] M. Zouari, M.T. Segarra, F. André (2010). A framework for distributed management of dynamic self-adaptation in heterogeneous environments. *In the 10th IEEE International Conference on Computer and Information Technology*, CIT 2010, pp. 265–272.

[13] E. Gamma , R. Helm, R. Johnson and J. Vlissides (1995). Design Patterns: Elements of Reusable Object-Oriented Software. *Addison-Wesley Longman Publishing Co., Inc*.

[14] J. Gray and A. Reuter (1992). Transaction Processing: Concepts and Techniques. *Morgan Kaufmann Publishers Inc.*, San Francisco, CA, USA.

[15] J. Kramer and J. Magee (2007). Self-Managed Systems: an Architectural Challenge. *Future of Software Engineering*, pp. 259-268.

[16] P. Oreizy, N. Medvidovic and R. N. Taylor (2008). Runtime software adaptation: framework, approaches, and styles. *In Companion of the 30th international conference on Software engineering (ICSE Companion '08)*. ACM, New York, NY, USA, pp. 899-910.

[17] J. Kramer and J. Magee (1990). The evolving philosophers problem: Dynamic change management. *IEEE Transactions on Software Engineering*, vol. 16, no. 11, pp. 1293–1306.

[18] M. Wermelinger (1997). A Hierarchic Architecture Model for Dynamic Reconfiguration. *In Proceedings of the Second International Workshop on Software Engineering for Parallel and Distributed Systems*, pp. 243–254.

[19] A. Ketfi, N. Belkhatir, P.Y. Cunin (2002). Adaptation Dynamique, concepts et experimentation. *In proceedings of the 15th International Conference on Software & Systems Engineering and their Applications ICSSEA02*, Paris, France.

[20] Ngoc-Tho Huynh, An Phung-Khac and Maria-Teresa Segarra (2011). Towards reliable distributed reconfiguration. *In Proceedings of the International Workshop on Adaptive and Reflective Middleware*, ARM 2011, pp. 36–41.

[21] M. Léger, T. Ledoux, and T. Coupaye (2007). Reliable dynamic reconfigurations in the fractal component model. *In Proceedings of the 6th*

international workshop on Adaptive and reflective middleware, ARM '07.

[22] E. J. Chikofsky and J. H. Cross II (1990). Reverse Engineering and Design Recovery: A Taxonomy. *IEEE Software*, vol. 7, no. 1, pp. 13-17.

[23] Anh Nyguyen-Tuong, Steve Chapin, Andrew Grimshaw, Charlie Viles (1998). Using Reflection for Flexibility and Extensibility in a Metacomputing Environment. University of Virginia, Technical Report CS-98-33.

[24] J. Koistinen (1997). Dimensions for Reliability Contracts in Distributed Object Systems. *Hewlett Packard Technical Report*, HPL-97-119.

[25] M. R. Lyu (2007). Software Reliability Engineering: A Roadmap. *In Future of Software Engineering*, *IEEE Computer Society*, pp. 153-170.

[26] O. Aissaoui, F. Atil and A. Amirat (2013). Towards a Generic Reconfigurable Framework for Self-adaptation of Distributed Component-Based Application. *In the book Modeling Approaches and Algorithms for Advanced Computer Applications, A. Amine et Al. (Eds), series SCI (Studies in Computational Intelligence)*, Vol. 488, Springer Ed., pp. 399-408.

[27] D. Garlan and B. Schmerl (2002). Model-based adaptation for self-healing systems. *In Proceedings of the first workshop on Self-healing systems*, ACM Press, Charleston, South Carolina, pp. 27-32.

[28] O. Aissaoui, A. Amirat, F. Atil (2012). An Adaptable and Generic Fault-Tolerant System for Distributed Applications. *In Proceedings of the International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, pp. 161-166.

[29] S.W. Cheng, A.C. Huang, D. Garlan, B. Schmerl and P. Steenkiste (2004). Rainbow: Architecture-based self-adaptation with reusable infrastructure. *IEEE Computer, vol.* 37, no. 10, pp. 46–54.

[30] T. Batista, A. Joolia, and G. Coulson (2005). Managing Dynamic Reconfiguration in Component-Based Systems. In *EWSA'05: 2nd European Workshop on Software Architecture*, Pisa, Italy, pp. 1-17.

[31] H. Tajalli, J. Garcia, G. Edwards, and N. Medvidovic (2010). PLASMA: a Plan-based Layered Architecture for Software Model-driven Adaptation. *In Proceedings of the 25th IEEE/ACM Int'l Conf. Automated Software Eng, (ASE 10)*, IEEE CS Press, pp. 467-476.

[32] A. Colmerauer and P. Roussel (1992). The birth of Prolog. *In the second ACM SIGPLAN conference on History of programming languages*, pp. 37-52.

[33] Y. Brun, G.M. Serugendo, C. Gacek, H. Giese, H. Kienle, M. Litoiu, H. Müller, M. Pezzè, M. Shaw (2009). Engineering Self-Adaptive Systems through Feedback Loops. *Software Engineering for Self-Adaptive Systems*, Springer-Verlag, Berlin, Heidelberg, pp. 48-70.

[34] J. Floch, S. Hallsteinsen, E. Stav, F. Eliassen, K. Lund, and E. Gjorven (2006). Using architecture models for runtime adaptability. *Software IEEE* vol. 23, no. 2, pp. 62-70.

[35] J. Cano Romero and M. García-Valls (2013). Scheduling component replacement for timely execution in dynamic systems. *Software practice and experince*, doi: 10.1002/spe.2181

Semi-automated Knowledge Elicitation for Modelling Plant Defence Response

Dragana Miljkovic
Department of Knowledge Technologies, Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
E-mail: Dragana.Miljkovic@ijs.si

Thesis Summary

Keywords: systems biology, constraint-driven optimization, natural language processing, triplet extraction, plant defence modelling

Received: July 30, 2014

This article represents a summary of the doctoral dissertation of the author on the topic of knowledge elicitation for modelling plant defence response.

Povzetek: Članek predstavlja povzetek doktorske disertacije, ki obravnava temo zajemanja znanja za modeliranje obrambnega odziva rastlin.

1 Introduction

In systems biology, the growth of experimental data is not uniform for different types of biological mechanisms, hence some biological mechanisms still have few datasets available, like plant defence against pathogen attack. Pathogens are a serious threat to living organisms and can lead to fitness costs, physiological damage or even death [1]. In particular, plants have specially evolved sophisticated mechanisms that can effectively fight-off infections with various pathogens. Upon pathogen recognition, plants trigger a complex signalling network, referred as plant defence response or plant defence signalling (PDS). Biologists have been investigating plant defence response to virus infections; however a comprehensive mathematical model of this complex process has not been developed. One challenge in developing a dynamic model, useful for simulation, are scarce experimental data from which the model parameters could be determined.

2 Methods and results

The thesis [2] describes a novel methodology for the construction of biological models by eliciting the relevant knowledge from literature and domain experts. The methodology has been applied to build the PDS model, and can be used to construct models of other biological mechanisms. The thesis also presents a PDS model.

Most of the plant-pathogen interaction studies are focused on individual interactions or subsets of the whole PDS mechanism. The models that are commonly used are structural models with no information on their dynamics [3]. Several dynamical models of plant defence have been developed. However, they are either simple [4], or do not contain sufficiently detailed information on the pathways of interest in this dissertation [5], or focusing only on one pathway [6].

In order to build the PDS model the standard approach to the construction of dynamic models is enhanced with the following methods: a method for model structure revision by means of natural language processing techniques, a method for incremental model structure revision, and a method for automatic optimisation of model parameters guided by the expert knowledge in the form of constraints.

The initial model structure was first constructed manually by defining the representation formalism, encoding the information from public databases and literature, and composing a pathway diagram. To complement the model structure with additional relations, a new approach to information extraction from texts was developed. This approach, named Bio3graph [7], allows for automated extraction of biological relations in the form of triplets followed by the construction of a graph structure which can be visualised, compared to the manually constructed model structure, and examined by the experts. Using a PDS vocabulary of components and reaction types, Bio3graph was applied to a set of 9,586 relevant full text articles, resulting in 137 newly detected relations. The resulting PDS pathway diagram represents a valuable source for further computational modelling and interpretation of omics data.

An incremental variant of the Bio3graph tool was developed to enable easy and periodic updating of a given model structure with new relations from recent scientific literature. The incremental approach was demonstrated on two use cases. In the first use case, a simple PDS network with 37 components and 49 relations, created manually, was extended in two incremental steps yielding the network with 183 relations. In the second use case, a complex PDS model structure in *Arabidopsis thaliana*, consisting of 175 nodes and 524 relations [7], was incrementally updated with relations from recently published articles, resulting in

an enhanced network with 628 relations. The results show that by using the incremental approach it is possible to follow the development of knowledge of specific biological relations in recent literature.

One obstacle in developing simulation models, are scarce kinetic data from which the model parameters could be determined. This problem was addressed by proposing a method for iterative improvement of model parameters until the simulation results meet the expectations of the biology experts. These expectations were formulated as constraints to be satisfied by model simulations. To estimate the parameters of the salicylic acid pathway, the most important PDS pathway, three iterative steps were performed. The method enabled us to optimise model parameters which provide a deeper insight into the observed biological system. As a result, the constraint-driven optimisation approach allows for efficient exploration of the dynamic behaviour of biological models and, at the same time, increases their reliability.

3 Conclusion

The main results of this thesis are: a new methodology for constructing biological models using the expert knowledge and literature and a PDS model, which was built by applying this methodology. Most notably, the standard approach to constructing dynamic models was upgraded with the following methods: a method for model structure revision by means of natural language processing techniques, a method for incremental development of biological model structures and a method for constraint-driven parameter optimisation.

The thesis also contributes to publicly available biological models and scientific software. The PDS model structure of *Arabidopsis thaliana* in the form of directed graphs is publicly available. Also, the Bio3graph approach is implemented and provided as a publicly accessible scientific workflow.

References

[1] Z. Zhao, J. Xia, O. Tastan, I. Singh, M. Kshirsagar, et al. (2011) Virus interactions with human signal transduction pathways, *International Journal of Computational Biology and Drug Design*, 4: 83 - 105.

[2] D. Miljkovic (2014) *Semi-automated knowledge elicitation for modelling plant defence response*, PhD Thesis, IPS, Jožef Stefan, Ljubljana, Slovenia.

[3] P. E. Staswick (2008) Jazing up jasmonate signaling. *Trends in Plant Science*, 13, 66-71.

[4] T. Genoud, M. B. T. Santa Cruz, J. P. Métraux (2001) Numeric simulation of plant signaling networks, *Plant Physiology*, 126, 1430-1437.

[5] M. Naseem, N. Philippi, A. Hussain, G. Wangorsch, N. Ahmed, et al. (2012) Integrated systems view on networking by hormones in arabidopsis immunity reveals multiple crosstalk for cytokinin, *Plant Cell*, 24, 1793-1814.

[6] A. Devoto, Turner, J. G. (2005) Jasmonate-regulated arabidopsis stress signalling network, *Physiologia Plantarum*, 123, 161-172.

[7] D. Miljkovic, T. Stare, I. Mozetič, V. Podpečan, M. Petek, et al. (2012) Signalling Network Construction for Modelling Plant Defence Response, *PLoS ONE*, 7(12):e51822.

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit a manuscript at: http://www.informatica.si/Editors/PaperUpload.asp. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

QUESTIONNAIRE

☐ Send Informatica free of charge

☐ Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twenty years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name: .....

Title and Profession (optional): .....

.....

Home Address and Telephone (optional): .....

.....

Office Address and Telephone (optional): .....

.....

E-mail Address (optional): .....

Signature and Date: .....

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bošković, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernández, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglarič, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. Čehovin, D. Čerepnalkoski, I. Čosić, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedič, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobrišek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmović, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engström, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipič, I. Fister, I. Fister Jr., D. Fišer, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligorić, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grčar, M. Grgurović, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaçebi, Q.-L. Han, H. Hanping, T. Härder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvärinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobović, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, Đ. Juričić, S. K, S. Kalajdziski, Y. Kalantidis, B. Kaluža, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollár, A. Kontostathis, P. Korošec, A. Koschmider, D. Košir, J. Kovač, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwaśnicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Luštrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marinčič, J. Marques-Silva, A. Martin, D. Marwede, M. Matijašević, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mišić, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Moškon, L. de M. Mourelle, H. Moustafa, M. Možina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Pančur, W. Pang, G. Papa, M. Paprzycki, M. Paralič, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Perš, D. Petcu, B. Petelin, M. Petkovšek, D. Pevec, M. Pičulin, R. Piltaver, E. Pirogova, V. Podpečan, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potočnik, R.J. Povinelli, S.R.M. Prasanna, K. Pripužić, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovič, D. Raković, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robič, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Rožanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. Šajn, R. Šenkeřík, M.R. Šikonja, J. Šilc, I. Škrjanc, T. Štajner, B. Šter, V. Štruc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampuš, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovič, V. Vojisavljević, M. Vozalis, P. Vračar, V. Vranić, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. Žalik, J. Žižka,

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2014 (Volume 38) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Janez Perš)
Slovenian Artificial Intelligence Society (Dunja Mladenić)
Cognitive Science Society (Urban Kordeš)
Slovenian Society of Mathematicians, Physicists and Astronomers (Andrej Likar)
Automatic Control Society of Slovenia (Sašo Blažič)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Vojteh Leskovšek)
ACM Slovenia (Andrej Brodnik)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 38 Number 3 September 2014 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **38** (2014) Number 3, pp. 191–309