Volume 23 Number 4 December 1999 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

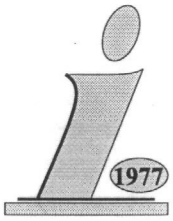
Special Issue:

Information Society and Inteligent Systems

Guest Editors:

Cene Bavec, Matjaž Gams

Informatica 23 (1999) Number 4, pp. 447–579



The Slovene Society Informatika, Ljubljana, Slovenia

Informatica

An International Journal of Computing and Informatics

Basic info about Informatica and back issues may be FTP'ed from `ftp.arnes.si` in magazines/informatica ID: `anonymous` PASSWORD: *<your mail address>*
FTP archive may be also accessed with WWW (worldwide web) clients with
URL: `http://www2.ijs.si/~mezi/informatica.html`

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, Vožarski pot 12, 1000 Ljubljana, Slovenia.
The subscription rate for 1999 (Volume 23) is
– DEM 100 (US$ 70) for institutions,
– DEM 50 (US$ 34) for individuals, and
– DEM 20 (US$ 14) for students
plus the mail charge DEM 10 (US$ 7).
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

LaTeX Tech. Support: Borut Žnidar, Kranj, Slovenia.
Lectorship: Fergus F. Smith, AMIDAS d.o.o., Cankarjevo nabrežje 11, Ljubljana, Slovenia.
Printed by Biro M, d.o.o., Žibertova 1, 1000 Ljubljana, Slovenia.

Orders for subscription may be placed by telephone or fax using any major credit card. Please call Mr. R. Murn, Jožef Stefan Institute: Tel (+386) 61 1773 900, Fax (+386) 61 219 385, or send checks or VISA card number or use the bank account number 900–27620–5159/4 Nova Ljubljanska Banka d.d. Slovenia (LB 50101-678-51841 for domestic subscribers only).

According to the opinion of the Ministry for Informing (number 23/216-92 of March 27, 1992), the scientific journal Informatica is a product of informative matter (point 13 of the tariff number 3), for which the tax of traffic amounts to 5%.

Informatica is published in cooperation with the following societies (and contact persons):
Robotics Society of Slovenia (Jadran Lenarčič)
Slovene Society for Pattern Recognition (Franjo Pernuš)
Slovenian Artificial Intelligence Society; Cognitive Science Society (Matjaž Gams)
Slovenian Society of Mathematicians, Physicists and Astronomers (Bojan Mohar)
Automatic Control Society of Slovenia (Borut Zupančič)
Slovenian Association of Technical and Natural Sciences (Janez Peklenik)

Informatica is surveyed by: AI and Robotic Abstracts, AI References, ACM Computing Surveys, ACM Digital Library, Applied Science & Techn. Index, COMPENDEX*PLUS, Computer ASAP, Computer Literature Index, Cur. Cont. & Comp. & Math. Sear., Current Mathematical Publications, Engineering Index, INSPEC, Mathematical Reviews, MathSci, Sociological Abstracts, Uncover, Zentralblatt für Mathematik, Linguistics and Language Behaviour Abstracts, Cybernetica Newsletter

The issuing of the Informatica journal is financially supported by the Ministry for Science and Technology, Slovenska 50, 1000 Ljubljana, Slovenia.

Post tax payed at post 1102 Ljubljana. Slovenia taxe Percue.

Introduction: Information Society and Intelligent Systems

This "Information Society" special issue of Informatica consists of selected papers, presented at the "Information society – IS'99" international conference held in Ljubljana, Slovenia. The conference papers were modified and typically extended for around 30%. The multiconference included five independent conferences: Information Society and Intelligent Systems, Education in Information Society, Data Mining and Warehouses, Development and Reengineering of Information Systems, Biology and Cognitive Sciences. All of them are related to information society. Most of the papers are dealing with information society in general and with information society in Slovenia.

No doubt we are rushing into the information society. Information, the Internet and intelligent systems characterize this, the most developed era of human civilization. In the information age, everybody has to adapt constantly. Even the mighty Microsoft, the undisputed king of SW-tools generation, is trembling in new circumstances. While big companies are intensively searching for new approaches and markets, an Internet society of millionaires is emerging in developed countries. These nouveau richies do not need enormous investments or decades of hard work. What they need is a good idea to be implemented on the Internet. Risks may be high, but incomes are enormous.

Internet business is growing by 70-80% each year. One year ago we had to observe Europe lagging behind USA. This year, Europe is speeding into information society with the same rate as America. Finland and some European countries are among most developed information countries in the world. Nokia, Finnish telecommunications and information giant, controls over 50% of the mobile telephone market.

Slovenia faces a kind of stagnation in 1999, opposite to the year before. Previous successful introduction of the Internet and information society was strongly correlated with the growth of national providers such as ARNES ("Academic and Research Network of Slovenia") for science and education, and CVI ("Government Centre for Informatics") for governmental institutions. Mobile telephony in Slovenia is growing rapidly because of the end of state monopoly. In other telecommunication areas, state monopoly remains the greatest obstacle for further progress. This in turn decreases the Internet and information society growth in Slovenia. Institutions, politicians and leading managers in Slovenia spend much of their energy fighting for their positions. Instead, they should take a clear position towards progress of information society and its introduction into private enterprises and governmental institutions. For example, unlike many Eastern European countries we have not managed to establish our national ACM chapter or national Information society Forum. Next year we are going to try again.

Information society is the only possibility of development if Slovenia wants to catch up with the most highly developed countries in Europe and elsewhere. The process of becoming a member of the European Union is a unique opportunity for Slovenia to decide clearly in favor for development strategy, and thus to become prepared to enter into the third millennium and into the information society.

The special issue consists of contributions grouped into general papers and technical applications. The first five papers in the special issue deal with information society in general, and often also with its introduction in Slovenia:

The first paper by M. Gams "Information Society and the Intelligent Systems Generation" describes general properties of information society and intelligent systems. It is claimed that information society promotes a primitive network intelligence displayed as connected autonomous intelligent agents. A short history introducing intelligent systems and agents in Slovenia is presented.

The "A New Perspective in Comparative Analysis of Information Society Indicators" paper by P. Sicherl analyses number of hosts in European countries and in particular, relations to Slovenia. Although Slovenia's position is relatively good - above European average, recent years indicate certain stagnation in comparison to most developed European countries like Finland.

V. Vehovar and M. Kovačič similarly to Sicherl conclude that Slovenia is generally on European average with respect to the penetration of the information technologies in society. In the paper "Measuring Information Society: Some Methodological Problems" they analyse different technical indicators and attitudinal measures and not just the number of hosts.

Another modelling and visualisation of information society development is presented in "Modelling of an Information Society in Transition - Slovenia's Position in the CE Countries" by M. Krisper and T. Zrimec. Different techniques are used, from clustering, portfolio, diagramming etc. Six Central European countries associated to the European Union are compared. Results show that Slovenia is often most closely associated with EU.

K. H. Iizuka and M. Wada in "Customer Satisfaction of Information System Integration Business in Japan" describe customer satisfaction as one of the major motivating factors in information society. They analyse and propose elements that affect total customer satisfaction.

The next group of papers describes information society and an additional technical subject, e.g. digital signatures:

First of these papers is the "Digital Signatures Infrastructure" by T. Klobučar and B. Jerman Blažič. They present an infrastructure for the use of digital signatures, technical aspects, and a short overview of several existing legal frameworks.

E. Jereb and B. Šmitek in "Using an Electronic Book in Distance Education" present experience they obtained by forming an electronic book in distant education, and student opinion studying in the new Distance learning centre.

In "Multi-Attribute Decision Modeling: Industrial Applications of DEX" M. Bohanec and V. Rajkovič describe application of a decision-support system DEX.

M. Ankerst, C. Elsen, M. Ester and H.-P. Kriegel describe algorithms and system for data visualisation in the "Perception-Based Classification" paper.

Large networks can be reduced into a smaller comprehensible structure that can be easier interpreted as proposed by V. Batagelj, A. Ferligoj, and P. Doreian in "Generalized Blockmodeling".

Clustering methods are described in the "Adapted Methods For Clustering Large Data-Sets Of Mixed Units" by S. Korenjak Černe.

I. Nančovska, L. Todorovski, A. Jeglič and D. Fefer describe an application of two systems - an equation discovery system and a neural network in "Equation Discovery System and Neural Networks for Short-term DC Voltage Prediction".

Soft computing and neural network methods are applied in "Adaptive On-line ANN Learning Algorithm and Application to Identification of Non-linear System" as presented by D. Sha and V.B. Bajič.

We hope that the special issue will be a contribution to information-society related efforts and that it will, from an independent, scientific aspect, give answers to a number of practical questions appearing in economy and in policy. Information society is based on the civil society of free-thinking individuals and of academic, economic and other groups offering the possibility to release the intellectual potential which will enlighten societies and countries, including Slovenia.

Cene Bavec, Matjaž Gams

Information Society And The Intelligent Systems Generation

Matjaž Gams
Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
Phone: +386 61 1773900; Fax: +386 61 1251038
matjaz.gams@ijs.si, http://www2.ijs.si/~mezi/matjaz.html

Keywords: information age, Internet, intelligent agents, overview paper, viewpoint

Edited by: Cene Bavec
Received: October 12, 1999 **Revised:** December 8, 1999 **Accepted:** December 19, 1999

In this overview paper we analyze basic laws and properties of the information society in general, and its introduction in Slovenia. It is claimed that information society initiated the emergence of primitive network intelligence demonstrated through intelligent assistants on the Internet. One of the key reasons for emergence of the new software generation is the growth of the Internet, and the other information overload. The introduction of intelligent systems, and particularly intelligent agents in Slovenia is analyzed. Finally, the EMA employment agent, one of important intelligent agent applications in Central Europe, is described in detail.

1 Introduction

Information society is often seen as another step in the progress of human civilization. We are moving from post-industrial society and economy into information society and information-technology dominated economy. Changes essentially influence the way we work and live. By 2002, it is predicted that over 80 million Europeans will have access to the Internet, and that 5 % of EU gross domestic product will be affected by the use of digital systems. Great trends are expected in the Internet commerce. In 1998, nearly $8 billion sales were generated in USA by around 9 million American households. In comparison, only $1.2 billion were accounted for online shopping in Europe. Europe was and is significantly lagging behind USA. But in 1999, Europe progressed much faster. Forecasts predict that 500.000 e-commerce-related jobs will be created within the next few years. Each year, Web sales to consumers are expected to grow by 70%. This growth is expected to be exponential for a couple of forthcoming years.

The technological basis of the information age is the Internet with its constant growth (Etzioni 1996; http://www.cio.com/WebMaster/metcalfe1.html).
Another important improvement is the emergence of network intelligence that represents a natural step in the computational evolution heading towards more helpful, adaptive and creative programs. These programs are essential for humans because without intelligent assistants we can not cope with information overload.

At the same time, the pace of progress is so quick and unpredictable in details that we can not determine future in any detail. What we can do is to recognize major information society laws (Lewis 1998; Metcalfe 1997) as described in Section 2. These laws are related to electronics, informatics, and the Internet. In Section 3 and 4 we analyze Slovenian introduction of information society and intelligent systems. The first major application of intelligent agents in Slovenia, the EMA employment agent, is described in Section 5.

2 Global Information Society

Information society is by definition global, however, its implementation is to a large extend dependent on the GNP of a particular country. Therefore, while the state and the pace of progress depend on each country, the basic information society laws stay valid for the global world.

Moore's Law (http://www.whatis.com/mooresla.htm) describes a constant trend in chip properties. The chip capacity doubles in a time span from 1.5 to 2 years depending on the type of particular performance of a chip. The formula is:

Performance(new) = Performance(old) * 1.5 time

where "time" is the number of years.

The basic property of the law - the constant exponential growth - remains unchanged over several decades (Moore 1975, Hamilton 1999).

Metcalfe's Law (http://www.cuug.ab.ca/~branderr/csce/metcalfe.html) says that the value of a network is proportional to the square of the number of nodes, connected by the network:

Value = K * nodes^2

In other words, the bigger the net, the square bigger the value. "K" is a constant.

Sidgemore's Law determines the growth of traffic over nets. The law says that the traffic doubles every three months:

$$\text{Traffic(new)} = \text{Traffic(old)} * 2^{(4\ \text{time})}$$

Andreesen's Law says that the cost of bandwidth is dropping exponentially and inversely proportional to Sidgemore's law:

$$\text{Cost(new)} = \text{Cost(old)} * 1/2^{(4\ \text{time})}$$

Lewis/Flemig's Law describes the network type of capitalism. It denotes nearly "friction-free economy" in the sense that there is small marginal cost and a huge shelf space. The exponential growth indicates that a genuine new market idea will get awarded by huge profits. But in addition to quick rise, an exponential decline is expected when new, more advanced systems appear on the market.

The equation describing the law is:

$$\text{MarketShare(time)} = 1/(1+ K * B * \text{time})$$

where "K" is a constant. The "B" parameter denotes the learning parameter.

Rules of the thumb:

Put on the Internet all your information and information activities. This law means that it is cheaper to put information and information activities on the Web sooner than later (Petrie et al. 1998). Not only it is indeed cheaper and more cost-effective than when done in a standard way, it is also the only way to go along with competition.

The cyber-world doubles fortune. Besides the material world we actually live in, the cyber-copy of our world matures. Since the introduction of the cyber-world in effect tends to double activities and money in circulation, stories of reach youngsters or rich Internet population in the developed countries are well grounded by a general trend. It also guarantees further growth of the developed world despite saturation in other human activities, which are related to classical material world. Another important trend is that our information systems on computers are becoming more and more a cyber-copy of ourselves.

Side-effect of information society is information overload. In infosphere we have to cope with more and more information from one month to another in order just to stay competitive. As a consequence, the information overload causes disappearance of free time, it causes the brain overload and decrease in classical human social life.

Information society demands intensive information knowledge for successful leadership. It is commonly accepted that there is a huge gap between existing knowledge of top executives, politicians and other leaders, and the desired knowledge for successful managing and leadership (O'Leary 1997). The gap is higher in Europe than in USA, and higher in Slovenia than in EU.

Information society belongs to all of us. In a democratic society there are several institutions cooperating in the process of governing and creating strategic directions. Among essential institutions of democratic societies are civil institutions (Borenstein 1998). Information society is by definition a civil society although governmental institutions typically implement it. An example would be Clinton's advocating of information highway or several governmental information society projects in Europe; e.g. Bangemann's reports (http://europa.eu.int/comm/ dg03/speechba.htm). In countries like Slovenia, lacking richness of civil society structures developed in decades of Western democracy, the introduction of Internet is a major inhibitor of faster progress.

The Internet is the most democratic and free media in the world. This was legally established with the American Supreme Court decisions about pornography and free speech on the Internet. In the simplest way it can be observed as a fact that pornography (inside "reasonable" limits) is allowed on the Internet and not on TV. It also means that anarchy and even criminal organizations can exploit this freedom, but the freedom of speech is accompanied by the fact that in such a case sooner or later we are going to hear things we don't like. Whatever the case, the Internet is the most democratic media humans ever had. While countries differ in their social and economic order, the Internet enhances democracy and civil society regardless of their previous level.

The Internet and information society are our hope for the future. There have been many technological innovations that spurred human progress. For example, we speak of the "iron age" historic period. These days we speak of the "information age" or of the "information society" age. Not only new technology changes the way we live in the technical sense; the changes are essential also in the way society functions (Negroponte 1998). At the same time, the world of computer systems we use is rapidly changing due to the massive introduction of information activities. The trend is towards more user-friendly intelligent systems.

3 Slovenia and IS

The introduction of information society and intelligent systems in Slovenia was accompanied with problems of all kinds. Slovenia is one of independent countries that emerged from former Yugoslavia. In those turbulent

transitions, funds for science continued to decrease for at last 5 years. Having in mind that there are only a couple of computer R&D institutions in the country with sufficient critical mass of educated stuff, the introduction of information society looked bleak.

Surprisingly, the progress in turbulent times was faster than anticipated. There might be at least two reasons: first, Slovenia lacked state institutions and by not being burdened by old institutions, new ones could be more up to date. Second, due to the inevitable conflict in the independence days the Internet played substantial role in helping to inform and thus motivate public opinion in the West. The introduction of information society started with the introduction of the Internet at the Jozef Stefan Institute (http://www.ijs.si) as result of a long-term cooperation with the Cern European Laboratory for Particle Physics (www.cern.ch/CERN/Technology/index.html). Soon, the use spread through universities and schools. The major Internet provider was (and still is) ARNES (http://www.arnes.si/).

After transformation into a market society, the economy trends changed to positive, but several state firms still faced substantial problems. Foreign investitures often bought firms and soon afterwards transformed research departments into production units. As a result, 45% of researchers and developers in the Slovenian R&D sector moved into other sectors.

Universities and research institutions were not as heavily stressed because of transition problems. University stuff even increased while for example the major research institute, the Jozef Stefan Institute with over 900 employees declined to 700.

The initial fast growth of the Internet hosts was slowed down after majority of governmental institutions got connected. Private institutions and individuals followed cautiously. Another important indicator is the number of people earning money through the Internet. While Bill Gates and CEOs of major computer companies dominate the list of world's richest, Slovenian computer professionals seldom appear on the list of nation richest. While in the developed countries an Internet generation of rich youngsters emerges, this phenomenon is substantially less present in Slovenia. Unlike in the developed countries, Slovenian political and partially business leaders often belong to the computer hardly literate generation. As a consequence, the real progress of information society is not as fast as it could be.

Overall, unemployment rate in Slovenia has grown up substantially in the last 10 years, especially in the first transition years. In recent years the unemployment in Slovenia is stable - with 125.000 unemployed and 2.000.000 inhabitants the unemployment rate is close to European average.

On the other hand, Slovenia is currently one of the most perspective candidates for joining the European Union based on political stability and economic parameters. It borders Italy, Austria, Hungary and Croatia (see Figure 1). GNP is roughly 10.000 US $. The number of computer connections to the Internet per capita is close to average in Western Europe. In recent years, all economic trends tend to be positive (with the exception of 7% inflation and growing depth). Even science and research, while not as supported and worshiped as they should be, are facing better times.

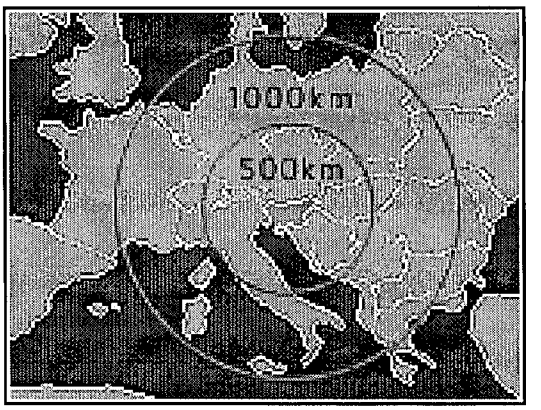


Figure 1: Slovenia's position in Europe.

Today, practically all schools from universities to elementary schools, all research and development institutions, and large majority of other governmental institutions are connected to the Internet. Several R&D conferences are more or less strongly related to the introduction of IS in Slovenia: "Information society" http://ai.ijs.si/is/indexa99.html, "Electronic commerce", Infos, ERK, Informatika etc.

In 1999, the major improvement has been in the mobile phone area. The number of mobile phones is expected to soon overcome the number of classical telephones.

4 Intelligent Systems

Intelligent systems are computer systems aimed at developing advanced user-friendly systems that work in real-life environments (Goonatilake, Treleaven 1995; Bielawski, Lewland 1991). The Internet is the media enabling substantial advantages for intelligent systems (Etzioni 1996; Etzioni 1997). Intelligent systems use a wide variety of artificial intelligence techniques typically implemented on top of classical modules (Bratko, Muggleton 1995): rule-based systems, production systems, expert systems, fuzzy logic systems, neural networks, memory-based reasoning. Advanced systems often combine various methods into one hybrid or integrated system (Gams et al., 1996). The emphasis of intelligent systems design is on combination of AI methods and engineering techniques enabling construction of systems performing practical tasks better than classical systems.

Intelligent agents (Bradshaw 1997; Mueller 1996) are a special branch of intelligent systems, capable of learning,

adapting to the environment, to each specific user, and to each specific situation as much as possible. According to Pattie Maes (Maes at al. 1999) intelligent agents are an important step ahead in humanising computers. Intelligent agents represent personal assistants collaborating with the user in the same environment (Maes, 1994; Minsky, 1987; 1991). Intelligent agents are basically intelligent interfaces providing specific utilities of the system while the core of the system is typically an Internet based query or database system. Unlike passive query languages, agents and humans both initiate communications, monitor events and perform tasks. The essential properties of agents are autonomy and sociability (Bradshaw 1997; Jennings, Wooldridge 1995; 1997; Etzioni, Weld 1995).

Intelligent systems have been developed in a couple of SW centers in Slovenia, among others in the Department of intelligent systems (headed by Prof. Ivan Bratko) at the Jozef Stefan institute. One example is an intelligent system for controlling quality of the Sendzimir rolling mill emulsion (Gams et al. 1996). Practically all national production of rolling steel is manufactured through this machine. The application represents one of major national intelligent system in regular industrial use. In addition to this application, the department has in the last ten years designed around 10 intelligent systems now available on the Internet http://turing.ijs.si/Ales/katalog-a/KATALOG-A.html.

5 The EMA Employment Agent

In 1993, the first agent was designed in Slovenia - IOI, an Intelligent Operating Interface (Hribovsek 1994; Gams, Hribovsek 1996). The basic task of IOI was correcting typing errors and providing help for users communicating with the VAX/VMS operating systems. IOI is an intelligent agent able to learn, adapt, and communicate in a relatively complex environment with human users. Its most important property is self-learning through observing the user performing tasks in the environment. Later, IOI uses accumulated knowledge through user experience to advise new users. The system thus performs a task similar to MS Office Assistant with the essential difference that knowledge in Assistant is coded in advance while IOI constructs most of its knowledge through user observation.

IOI is implemented as a 2000-line program in Pascal with parts of it written in the VAX/VMS command language. The IOI agent was implemented as a research prototype only, however, its flexibility and adaptability as a personal intelligent agent have shown reasonable improvements over classical systems. The most positive properties as observed by users in the testing period, are: IOI is easy to use, it does not demand specific knowledge, is easy to learn and use, and is very transparent. These favorable properties are typical intelligent-agent properties. Two other major intelligent agents developed in Slovenia are Personal WebWatcher (Mladenic 1998) and Ema (Gams et al. 1998).

The EMA project (see Figure 2) started seven years ago as an R&D project "An Integrated Information System for Employment in Slovenia" to provide help regarding unemployment problems. The project was partly funded by the Slovenian Ministry of Science and Technology and partly by the Employment Service of Slovenia (ESS). The system consists of two parts; one is applied at the Employment Service of Slovenia (http://www.ess.gov.si/English/elementi-okvirjev/F-Introduction.htm) where one gets basic information about employment activities in Slovenia, about ESS, and about interesting employment functions. The top part of the system is the EMA agent. For the last three years, the system was further developed as part of the INCO-Copernicus Project: 960154, cooperative research in information infrastructure, CRII (http://www-ai.ijs.si/~ema/proj.html). The intelligent system/agent EMA with a natural language interface consists of several modules.

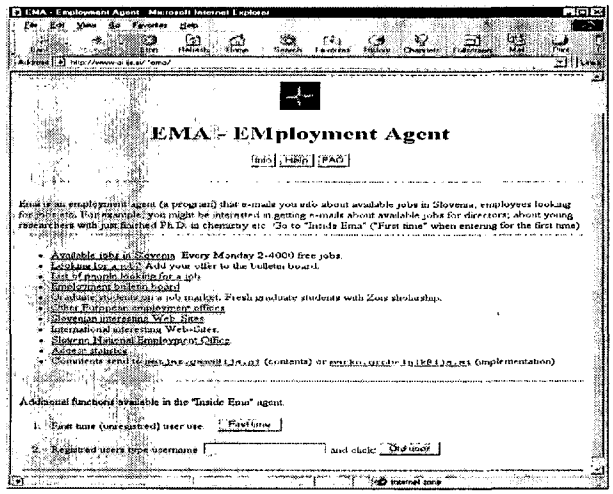


Figure 2: The first version of the EMA employment agent was among the first in the world to offer substantial amount of nationally available jobs on the Internet.

A user has to identify with a username (note that here security is not very relevant) or through a favorite/bookmark list. EMA has four basic functional modules: storing patterns and ordering mails regarding vacant jobs, available workers, it enables storing and observing interesting WEB sites chosen by users, and enables matching jobs and workers. EMA is a "classical" agent providing user-friendly information upon demand or when it notices relevant information for each particular user.

The system is a 15.000 lines program written mainly in C, partially in other languages. Together with text and data it occupies 30 M on a disk.

EMA receives data as limited Slovenian text (with the exception of bulletin boards with language independent

free input) and translates it into English. The translation is based on a dictionary consisting of up to four words observed before in the employment data. New combinations are in the worst case translated as direct word-by-word translation and stored for further overview by humans. Stored combinations are sorted by frequency and translated by humans if reasonable. In addition, the translation system looks into the morphology dictionary to capture different forms of the same words. Finally, a spell-checker submodule corrects spelling errors. The translation is currently not yet at the level performed by systems translating between larger European languages, however, it is sufficiently good to enable understanding since the syntax is quite limited.

In the next stage, the text is transformed into appropriate computer readable forms and HTML forms as outputs. Two speech modules transform the data into speech. The English speech system is based on the Microsoft agents. We have designed our own Slovenian speech module (Sef et al. 1998).

The EMA agent was and still is among most successful applications of intelligent agents in Slovenia. In the first year of its implementation, our country was the third in Europe to offer national employment information through the Internet. At that time, we were the first country in the world to provide over 90% of all nationally available jobs on the Internet.

6 Conclusion

Without any doubt human civilization further evolves into information society. We are able to establish certain laws and rules of this development while specific details and further progress remain enigma to all of us. Intelligent systems and agents through the Internet and partially through PCs form the new software generation, the intelligent systems generation. While this generation still lacks true human intelligence and consciousness, the primitive network intelligence emerges consisting of intelligent assistants capable of autonomous and social activities (Munindar 1997; Mylopoulos 1997).

In Slovenia, one of the countries wishing to join European Union, information society is perceived as a global phenomenon and as a major technological field which can bring us fortune or stagnation. The essential question is whether the existing or at least the forthcoming generation of political and business leaders will fully embrace the information-age rules of the game.

Acknowledgement:

Financial support for the EMA project was provided by an international project INCO-Copernicus 960154, Cooperative Research in Information Infrastructure, CRII; by the Ministry of science and technology in Slovenia; and by ESS. We would like to thank the CEO of the Employment Service of Slovenia, Mr. J. Glazer.

7 References

[1] L. Bielawski, R. Lewland, Intelligent Systems Design; Integrating Expert systems, Hypermedia and Database Technologies, Wiley, 1991.

[2] N. S. Borenstein, "Whose Net is it Anyway", Communications of the ACM, April 1998, pp. 19-21.

[3] M. Bradshaw (ed.), Software Agents, AAAI Press/The MIT Press, 1997.

[4] I. Bratko, S. Muggleton, "Applications of Inductive Logic Programming", Communications of the ACM, Vol. 38, No. 11, 1995, pp. 65-70.

[5] O. Etzioni, D.S. Weld, "Intelligent Agents on the Internet: Fact, Fiction, and Forecast", IEEE EXPERT, Intelligent Systems & Their Applications, Vol. 10, No. 4, 1995, pp. 44-49.

[6] O. Etzioni, "The WWW: Quagmire or Gold Mine?", Communications of the ACM, Vol. 39, No. 11, 1996, pp. 65-68.

[7] O. Etzioni, "Moving Up the Information Food Chain", AI Magazine, Vol. 18, No. 2, 1997, pp. 11-18.

[8] M. Gams, B. Hribovšek, "Intelligent-Personal-Agent Interface for Operating Systems", Applied Artificial Intelligence, Vol. 10, 1996, pp. 353-383.

[9] M. Gams, M. Drobnič, N. Karba, "Average-Case Improvements when Integrating ML and KA", Applied Intelligence 6, No. 2, 1996, pp. 87-99.

[10] M. Gams, A. Karalič, M. Drobnič, V. Križman, "EMA - An Intelligent Employment Agent", Proc. of the Forth World Congress on Expert Systems, Mexico, 1998, pp. 57-64.

[11] S. Goonatilake, P. Treleaven (eds.), Intelligent Systems for Finance and Business, Wiley, 1995.

[12] S. Hamilton, "Taking Moore's Law into the Next Century", IEEE Computer, Januar 1999, pp. 43-48.

[13] B. Hribovsek: Intelligent Interface for VAX/VMS, M. Sc. Thesis (in Slovene).

[14] N.R. Jennings, M. Wooldridge, "Intelligent Agents and Multi-Agent Systems", Applied Artificial Intelligence, An International Journal, Vol. 9, 1995, pp. 357-369.

[15] N.R. Jennings, M. Wooldridge, Agent Technology, Springer, 1997.

[16] M. Lewis, "Designing for Human-Agent Interaction", AI Magazine, Vol. 19, No. 2, 1998, pp. 67-78.

[17] P. Maes, "Agents that Reduce Work and Information Overload", Communications of the ACM, 37, 1994, pp. 31-40.

[18] P. Maes, R.H. Guttman, A. G. Moukas, "Agents That Buy and Sell", Communications of the ACM, Vol. 42, No. 3, 1999, pp. 81-91.

[19] B. Metcalfe, "What's Wrong with the Internet", IEEE Internet Computing, 1997, pp. 6-8.

[20] M. Minsky, The Society of Mind, Simon and Schuster, New York, 1987.
[21] M. Minsky, "Society of mind: a response to four reviews", Artificial Intelligence 48, 1991, pp. 371-396.
[22] D. Mladenic, "Turning Yahoo into an Automatic Web-Page Classifier", in Proceedings of the 13th European Conference on Artificial Intelligence ECAI'98, 1998, pp. 473-474.
[23] G. E. Moore, "Progress in digital integrated electronics", Technical Digest of 1975 International Electronic Devices Meeting 11, 1975.
[24] J.P. Mueller, The Design of Intelligent Agents, Springer, 1996.
[25] P.S. Munindar, "Agent Communication Languages: Rethink the Principles", Computer, December 1997, pp. 40-47.
[26] J. Mylopoulos, "Cooperative Information Systems", IEEE EXPERT, Intelligent Systems & Their Applications, Vol. 12, No. 5, 1997, pp. 28-30.
[27] N. Negroponte, "A Wired Worldview", EU RTD info, March 1998, pp. 28-30.
[28] D. E. O'Leary, "A Lack of Knowledge at the Top", IEEE Expert, November 1997, p. 2.
[29] C. J. Petrie, A. M. Rutkovski, M. Zacks etc., "Dimensioning the Internet", IEEE Internet Computing, April 1998, pp. 8-9.
[30] T. Sef, A. Dobnikar and M. Gams, "Improvements in Slovene Text-to-Speech Synthesis", Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98), pp. 2027-2030, 1998, Sydney.

A New Perspective in Comparative Analysis of Information Society Indicators

Pavle Sicherl
Law School, University of Ljubljana and SICENTER, Brajnikova 19, 1000 Ljubljana, Slovenia
Tel: +386 61 1501510; fax: +386 61 1501514
Pavle.Sicherl@sicenter.si

Keywords: time distance, S-distance, two-dimensional comparison in time and indicator space

Edited by: Cene Bavec and Matjaz Gams
Received: October 8, 1999 **Revised:** November 23, 1999 **Accepted:** December 12, 1999

The analysis of information society indicators can be enriched by supplying a new view of data that can provide new insight from existing data. The slowdown of growth of Internet hosts per 10000 inhabitants in Slovenia after mid-1997 increased the time lag of Slovenia behind leading Finland from 3 years at the end of 1996 to nearly 5 years by August 1999. Time distance methodology is used as a presentation and communication tool to raise awareness of the problem and its consequences in simple understandable terms and to signal the need for an in-depth analysis and action.

1 Introduction

Problem:
In Slovenia, after a very high rate of growth in the indicator Internet hosts per 10000 inhabitants until mid-1997, such growth slowed down substantially. One can describe the facts in various ways and with various measures.
Objective:
To make the government, other agents and general public aware of these developments and signal the need for immediate action to correct them.
Method:
Time distance will be used as a presentation and communication tool to raise awareness of the problem and its consequences in simple widely understandable terms. Since this method can be a useful addition to existing methods of analysing differences between compared units in many fields, a further illustration is provided for the case when the benchmark for comparison is the average value of the analysed indicator for EU15.

2 Methodology: time distance concept and statistical measure S-distance

The time perspective, which no doubt exists in human perception when comparing different situations, is systematically introduced both as a concept and as a quantifiable measure. Since events are dated in time, in time series comparisons, regressions, models, forecasting and monitoring, the notion of time distance always existed as a "hidden" dimension. In order to systematise and formalise the approach and define an appropriate statistical measure for operational use, amendments to the present state-of-the-art are needed on two levels: conceptual and analytical.

First, a broader theoretical framework is required. The conventional approach does not realise that, in addition to the disparity (difference, distance) in the indicator space at a given point in time, in principle there exist a theoretically equally universal disparity (difference, distance) in time when a certain level of the indicator is attained by the two compared units. Second, a statistical measure S-distance has been defined to suggest a possibility how the broader concept and reference framework can be measured in operational terms. The aim is to provide new insights from existing data due to an added dimension of analysis and thus to complement conventional statistical measures.

Time distance in general means the difference in time when two events occurred. We define a special category of time distance, which is related to the level of the analysed indicator. The suggested statistical measure S-distance measures the distance (proximity) in time between the points in time when the two compared series reach a specified level of the indicator X. The observed distance in time (the number of years, quarters, months, days, minutes, etc.) is used as a dynamic (temporal) measure of disparity between the two series in the same way as the observed difference (absolute or relative) at a given point in time is used as a static measure of disparity [1,2,3].

For a given level of X_L, $X_L = X_i(t_i) = X_j(t_j)$, and the S-distance, the time separating unit (i) and unit (j) for the level X_L, will be written as

$$S_{ij}(X_L) = \Delta T(X_L) = t_i(X_L) - t_j(X_L)$$

where T is determined by X_L. In special cases T can be a function of the level of the indicator X_L, while in general it can be expected to take more values when the same level is attained at more points in time, i.e. it is a vector which can in addition to the level X_L be related to time. Three subscripts are needed to indicate the specific value of S-distance: (1 and 2) between which two units is the time distance measured and (3) for which level of the indicator (in the same way as the time subscript is used to identify the static measures). In the general case also the fourth subscript would be necessary to indicate to which point in time it is related ($T_1,T_2,...,T_n$).

The sign of the time distance comparing two units is important to distinguish whether it is a time lead (-) or time lag (+) (in a statistical sense and not as a functional relationship):

$$S_{ij}(X_L) = -S_{ji}(X_L) .$$

Using the comparison between two units it can be shown that the generic concept of time distance goes together very naturally with the existing concepts of static disparity at a given point in time and the notion of the growth rate over time. Table 1 provides a schematic example for such comparisons for a given indicator. Row one is the most frequently used type of comparative analysis; levels of the indicator at a given point in time are compared. In such comparison two points are used, for each of them we have three elements of information: (i) the respective level of the indicator, (ii) to which unit it belongs, and (iii) at what time it happened. In this case unit as well as time (since it is constant for static comparison) serve as identifiers, while the levels are used to calculate the static difference. Row two compares two levels of the indicator for each unit at two points in time, separately for each unit, which means that one calculation indicated in row two refers to unit 1, and another to unit 2. The simplest example would be growth rate for unit 1 and growth rate for unit 2. Here the unit is the identifier, while the numerical values on levels and time are used in calculating this measure.

These two steps are standard procedures. The first one represents the static type of comparison; the second one measures the dynamic properties of the indicator for each unit separately. Following the same logic, for the novel statistical measure S-distance in row three level is the same, level and unit serve as identifiers, and time is used for calculating time distance. It is remarkable that the notion of time distance, which can be in principle developed from the same information used in steps one and two, has not been developed theoretically and as a standard statistical measure.

	TIME	UNIT	LEVEL	Measure
TIME	same	2	2	static difference
UNIT	2	same	2	change over time
LEVEL	2	2	same	time distance

Table 1. Points of comparison for static difference, change over time and time distance (two units)

While there may be different problems involved in the calculation of these three types of measures, in terms of availability and comparability of data, in principle these three types of measure can be integrated into a formally consistent analytical framework. There are alternative ways of doing this, following from the distinction between backward looking (ex post) and forward looking (ex ante) time distances. They relate to different periods, past and future, the first belongs to the domain of statistical measures based on known facts, the second is important for describing the time distance outcomes of the results of alternative policy scenarios for the future. Looking backwards, ex post or historical time distance indicates how many years ago the more developed unit experienced a specified level of the indicator of the less developed unit at a given point in time [3]. A very important relationship shows that, ceteris paribus, time distance is a decreasing function of the magnitude of the growth rate of the indicator. This conclusion shows that the S-distance as a dynamic (temporal) measure of disparity offers a perspective which may be quite distinct from that provided by static measures.

This new view of the information is using level(s) of the variables as identifiers and time as a focus of comparison and numeraire. This approach and the broad range of its possible applications is much more complex and general, but the time distance is the priority choice because of its intuitive nature, and the importance of the time dimension in semantics of describing various situations in real life and forming our perceptions about them. In this paper only the application to comparison of one indicator between several units will be used. However, the approach has been generalised to complement conventional measures in time series comparisons, regressions, models, forecasting and monitoring, and to analysis of single time series [3] and to variables other than time [4]. In all such applications it can provide from existing data new insights due to an added dimension of analysis.

3 Data and results for Slovenia, EU15 countries and candidate countries

Data on Internet hosts per 10000 inhabitants used relate to the period end of 1993-August 1999 [5,6,7]. At present is the measurement and empirical analysis of information society indicators beset with problems. It is stated that the single most important obstacle to effective data collection is the lack of standardised definitions of information technology and the exclusion of important

costs associated with its use, like personnel and training expenses. A further weakness is the relative absence of systematic information how information technology is actually being used [8]. In addition to these general obstacles there may be also some specific reasons that the slowdown of the increase in Internet hosts per capita in Slovenia in the last two years shown in RIPE data may have been exaggerated [9]. We shall proceed by analysing the available RIPE data, yet there should be an appropriate caution about possible inaccuracy in the available data. Comparative analysis of the differences among countries can be presented in two dimensions. The conventional static differences at a given point in time are in this paper complemented by the time distance dimension. Time distance in Table 3 is for practical reasons calculated for the levels of the indicator for those countries, which are behind Slovenia, and for the level of Slovenia for the countries, which are ahead of Slovenia.

	1993	1994	1995	1996	1997	1998	Aug. 1999
LUX	7.4	12.5	46.0	85.2	113.4	182.3	218.8
DAN	16.1	35.4	96.9	203.3	321.1	571.1	608.3
BEL	7.0	17.3	30.2	64.0	104.8	202.4	307.5
AUT	18.9	34.0	66.3	110.2	134.4	214.0	235.7
DEU	13.7	24.4	58.0	84.4	137.7	177.0	186.3
FRA	9.3	14.4	26.0	40.6	60.7	84.2	106.4
NED	28.6	55.8	110.8	173.4	249.2	395.1	481.9
ITA	2.9	5.0	13.1	25.8	44.2	64.5	96.4
SVE	47.0	84.8	164.1	269.0	394.0	429.9	569.5
UK	19.1	38.7	75.1	122.4	167.3	247.4	272.2
FIN	65.2	133.9	416.7	612.1	945.8	902.6	930.5
IRL	6.5	15.3	37.3	74.2	109.3	155.6	181.7
ESP	3.6	7.0	13.1	28.8	49.9	78.2	94.2
PRT	3.6	5.1	11.9	23.6	42.7	56.3	67.4
GRE	1.7	3.4	7.4	16.0	26.7	47.1	63.5
SLO	3.1	8.2	28.3	69.5	98.2	115.3	116.3
CZE	4.3	10.1	21.1	39.6	55.2	83.6	101.8
SVK	0.7	2.6	5.6	14.8	27.0	41.0	48.3
HUN	3.0	6.6	15.4	29.2	66.7	87.8	106.2
POL	1.3	2.8	6.0	13.7	22.9	32.4	42.9
EST	2.9	7.7	24.1	54.3	108.4	151.2	180.8
ROM	0.0	0.2	0.8	3.5	6.0	9.9	14.1
LIT		0.3	1.2	4.7	10.9	26.0	32.7
LAT	0.2	2.0	5.2	23.1	28.6	54.4	63.7
BG	0.0	0.2	1.3	4.0	8.2	12.2	18.3
EU15	12.2	23.6	50.5	78.6	124.3	171.1	199.3

Table 2: Data on Internet host density per 10000 inhabitants
Source: International Telecommunication Union Database, Geneva 1998 for 1993-1997 [5]; RIPE [6] in RIS [7] for 1998 and 1999.

In Tables 2 and 3 the countries are sorted by the level of GDP per capita (at purchasing power standards) in 1997. Obviously, the Internet hosts per capita are not firmly correlated with GDP per capita. In 1996 Slovenia was occupying a comfortable comparative position in terms of Internet hosts per capita: it was lagging less than 3 years behind Finland as the leading country, and was ahead of several EU countries, i.e. Belgium, France, Italy, Spain, Portugal and Greece. The last four mentioned countries had substantially lower values than Slovenia.

The slowdown of growth rate in this indicator for Slovenia after mid-1997 led to a quick deterioration of the comparative situation of Slovenia. By August 1999 the lag behind Finland increased to nearly 5 years. Namely, in case of indicators with high rates of growth the situation can change very quickly, as distinct from the fields where the rate of change is slow. Figure 1 provides visualization of these changes. Tables 2 and 3, and Figure 1 compare Slovenia with EU15 countries and the nine candidate countries from Central and Eastern Europe.

One could also speculate what would be the situation if the rate of growth for the period 1997–August 1999 would continue until the end of 2000 (this should not be interpreted as projections).

	1994	1995	1996	1997	1998	Aug. 1999
LUX	-0.9	-0.5	-0.4	-0.5	-1.0	-1.5
DAN	#N/A	-1.4	-1.5	-2.0	-2.8	-3.4
BEL	-0.9	-0.2	0.1	-0.2	-0.9	-1.5
AUT	#N/A	-1.4	-0.9	-1.3	-1.8	-2.3
DEU	#N/A	-0.9	-0.6	-0.7	-1.4	-2.0
FRA	#N/A	0.1	0.7	1.2	1.5	1.1
NED	#N/A	#N/A	-1.8	-2.2	-2.9	-3.5
ITA	0.6	0.8	1.1	1.6	2.1	1.7
SVE	#N/A	#N/A	-2.4	-2.8	-3.6	-4.2
UK	#N/A	-1.5	-1.2	-1.5	-2.2	-2.7
FIN	#N/A	#N/A	-2.9	-3.5	-4.3	-4.8
IRL	-0.8	-0.4	-0.1	-0.3	-0.9	-1.4
ESP	0.2	0.8	1.0	1.5	1.7	1.7
PRT	0.6	0.8	1.2	1.7	2.3	2.6
GRE	0.9	1.2	1.6	2.1	2.5	2.7
SLO	0.0	0.0	0.0	0.0	0.0	0.0
CZE	-0.3	0.4	0.7	1.4	1.5	1.4
SVK	#N/A	1.5	1.7	2.1	2.7	3.1
HUN	0.3	0.6	1.0	1.1	1.4	1.1
POL	#N/A	1.4	1.7	2.3	2.9	3.2
EST	0.1	0.2	0.4	-0.2	-0.8	-1.4
ROM	#N/A	#N/A	2.9	3.4	3.9	4.3
LIT	#N/A	#N/A	2.7	2.9	3.1	3.5
LAT	#N/A	1.6	1.3	2.0	2.4	2.7
BG	#N/A	#N/A	2.8	3.0	3.8	4.1
EU15	#N/A	0.8	0.3	0.6	1.2	1.8

Table 3: Time distance between compared countries and Slovenia, S-distance in years: - time lead, + time lag, Slovenia=0
Source: Own calculation based on data in Table 1.

If no action would be taken and such slowdown would continue until the end of 2000, a further deterioration of the relative position of Slovenia for this indicator would take place. Slovenia would within a period of only a few years move from a comfortable position near the EU15 average in 1996 (despite being more than 30 per cent below the average EU15 level of GDP per capita) to a position where the lag behind the forerunner Finland would be already 6 years. The lag behind Sweden, Denmark and Netherlands would be around 5 years, France, Italy, Spain and Greece would surpass or catch up with Slovenia, and only Portugal out of the EU15 countries would be still behind it.

Time distance seems to be an excellent way of presenting the danger of a rapidly deteriorating situation, which everybody can understand, and to signal that an in-depth analysis and corresponding actions are necessary. Some other conventional measures may not provide such warning. E. g., static comparison showed that in 1996 Finland had 8.8 times the number of Internet host per capita in Slovenia, and in 2000 it would be 6.6 times. Time distance adds a qualitatively different conclusion.

Similar consequences can be seen from comparison with selected Central and Eastern European countries. In 1996 Slovenia was with Estonia a clear leader in the region for the indicator Internet hosts per capita. In the meantime Estonia moved ahead, and the gap would widen if the present trends would continue. By August 1999 Slovenia is lagging behind Estonia for more than 1 year.

The quality of time distance measure, being transparent and easy to perceive and understand, can be even more appreciated when a larger set of indicators is analysed, involving more issues and different fields of concern. For instance, in 1997 Italy was 18.3 years ahead of Slovenia for GDP per capita at purchasing power parity, while Slovenia was 1.6 years ahead of Italy for Internet hosts per capita. Some of these indicators can change very quickly, some others, like some demographic variables and some other characteristics of human factor, very slowly. Time distances will be different, smaller for those indicators that are more dynamic by their nature, more conducive to policy measures and given higher priority in decision-making process.

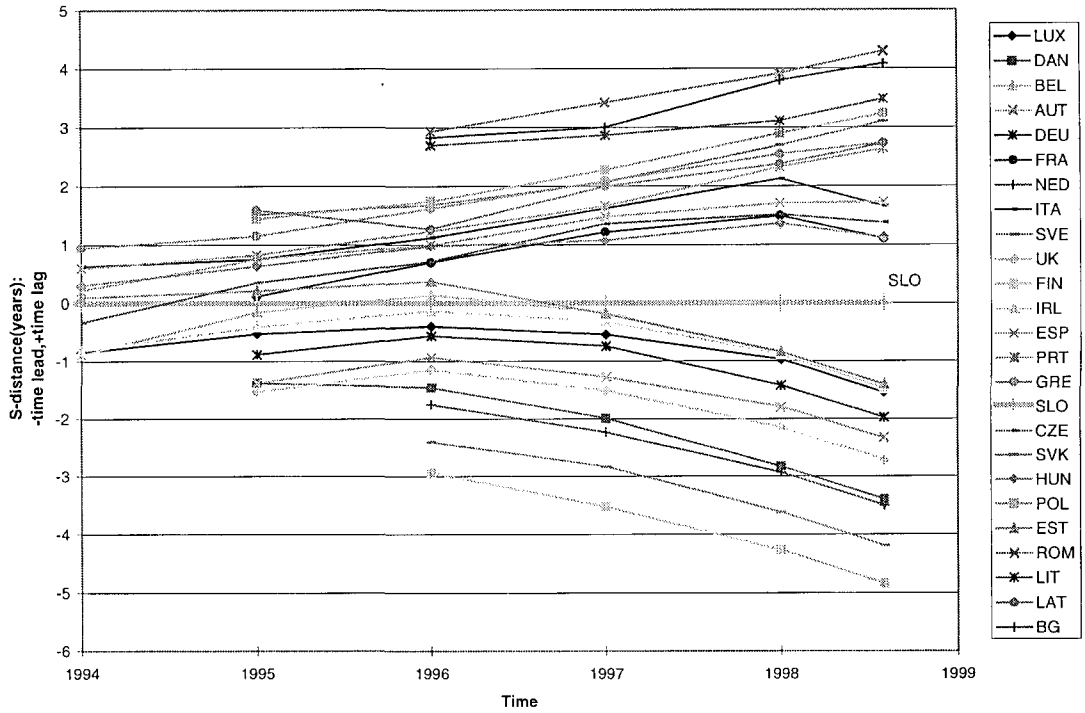


Figure 1. Time distance for Internet host density per 10000 inhabitants, EU and candidate countries, Slovenia=0

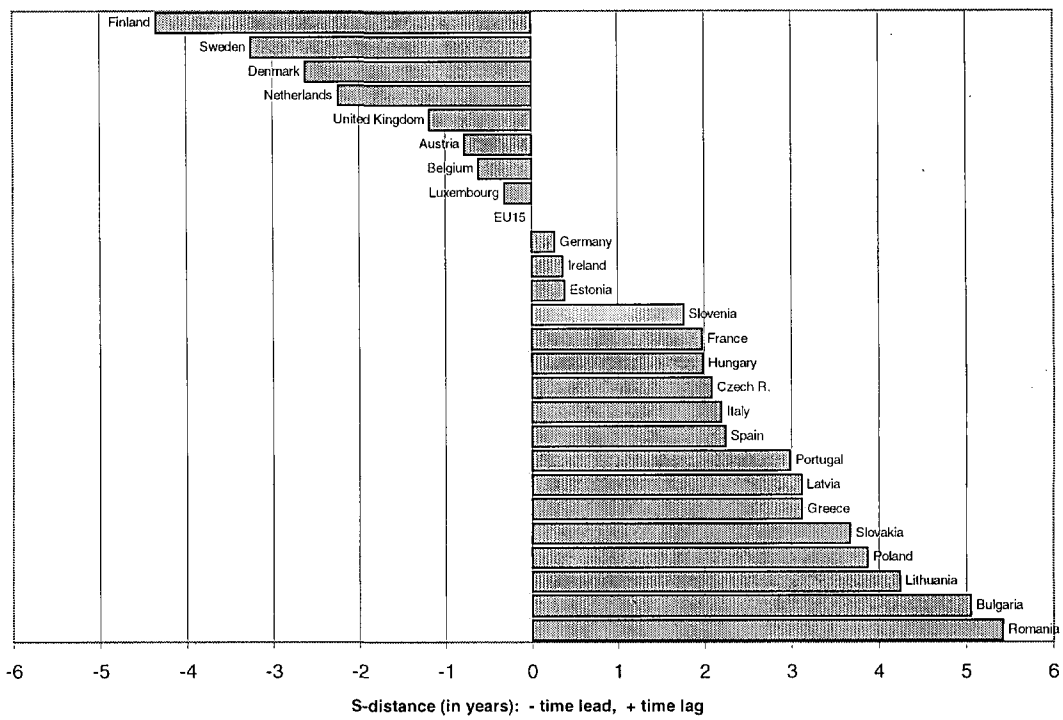


Figure 2. Differences from EU15 average for Internet hosts per capita expressed in time (August 1999)

Figure 2 is an illustration of application of time distance presentation in a similar case of comparative analysis. In this example the average value of Internet hosts per capita for EU15 is the benchmark for comparison. The dispersion of situations in this respect for EU15 countries and Central and Eastern European candidate countries can be presented in various ways, like ratios, percentages, absolute value and absolute differences, etc. Furthermore, various summary measures of dispersion could be calculated.

Absolute values of the indicator are presented in Table 2. A widely used conventional measure would be indeces or percentage differences. For instance, in August 1999 the index for forerunner Finland would be 467, for Portugal and Greece about 33 as the lowest value for EU15 countries, and 9 for Bulgaria and 7 for Romania (EU15=100). Figure 2 presents another complementary view of this set of data. Time distances are calculated in cases of above the average countries for the level of EU15 average, and in cases of below the average countries for the level of indicator in these countries in August 1999. Finland had a lead of about 4.5 years ahead of the EU15 average, Portugal and Greece were laging the EU15 average for about 3 years, and Bulgaria and Romania for more than 4 years. Time distances alow for a distinct new insight that can help to form a richer perception of the situation.

Since time distance is expressed in units of time, which everybody understands from ministers, managers to general public, it possesses one of the ideal characteristics of a presentation and communication instrument. It is expected that the analysis of and discussion about time distances will have considerable influence on how people will form their perception about a situation and on public opinion. For instance, in the EU the consideration of economic and social cohesion is an important goal. A series of presentation of results like Figure 2 for a number of relevant indicators would without any doubt provide a new additional insight to a complex multidimensional problem. Similarly, it would be very useful if the results in Table 3 and in Figures 1 and 2 would be provided for a broad selection of information society indicators.

This offers improved semantics for analysis and policy debate, and can in many cases lead to qualitatively different conclusions from those reached in a static conceptual and analytical framework. By analogy, there is a wide-open possibility to apply this methodology to numerous business problems at the micro, corporate and sector levels. Another important advantage of this approach is that the results and conclusions based on the two-dimensional analysis add new information and new insight, while none of the earlier results are lost or replaced.

4 Conclusions

In empirical research the art of handling and understanding of different views of data is crucial for discovering the relevant patterns. The time distance approach (with associated statistical measure S-distance) is useful at least in two domains: it offers a new view of data that is exceptionally easy to understand and communicate, and it may allow for developing and exploring new hypotheses and perspectives that cannot be adequately dealt without the new concept.

The generic nature of the time distance concept and the S-distance measure leads to the conclusion that the methodology can be usefully applied as an important analytical and presentation tool in numerous applications in a wide variety of substantive fields. Especially in the field of information technology indicators, which is characterised by great speed of change, it would be of great interest to complement rather than replace the conventional measurement of differences between countries or other units with this new perspective of the situation.

5 References

[1] P. Sicherl. *A Novel Methodology for Comparisons in Time and Space.* East European Series No. 45. Vienna Institute for Advanced Studies. Vienna. 1997.

[2] P. Sicherl. Time Distance in Economics and Statistics; Concept, Statistical Measure and Examples. In A. Ferligoj (ed.). *Advances in Methodology, Data Analysis, and Statistics.* Metodoloski zvezki. 14. FDV. Ljubljana. 1998.

[3] P. Sicherl. The Time Dimension of Disparities in the World. *XIIth World Congress of the International Economic Association.* Buenos Aires. August 1999.

[4] P. Sicherl. Measuring disparities in two dimensions: proximity in time and proximity in indicator space. *10th International Conferene on Socio-economics,* Vienna, Austria 13-16 July 1998.

[5] International Telecommunication Union. *Database.* Geneva. 1998.

[6] http://www.ripe.net

[7] http://www.ris.org

[8] National Science Board, *Science & Engineering Indicators – 1998.* Arlington, VA: National Science Foundation, 1998.

[9] V. Vehovar, Spremljanje informacijske druzbe, in P. Sicherl, A. Vahcic (eds.), *Model indikatorjev za podporo odlocanju o razvojni politiki in za spremljanje izvajanja SGRS*, Sicenter, Ljubljana, oktober 1999.

Measuring Information Society: Some Methodological Problems

Vasja Vehovar, Matej Kovačič
Faculty of Social Sciences, Kardeljeva ploščad 5, 1000 Ljubljana, Slovenia
Phone: +386 61 1805 100; Fax: +386 61 1805 102
vasja.vehovar@uni-lj.si, matej.kovacic@uni-lj.si

Keywords: information society, Internet indicators, survey research, electronic commerce

Edited by: Cene Bavec and Matjaz Gams
Received: October 10, 1999 **Revised:** November 30, 1999 **Accepted:** December 20, 1999

The paper addresses methodological problem of measuring information society. Both, technical indicators and attitudinal measurements for Slovenia are discussed in this context. In particular, the results related to the interest for information society services are presented. The comparison between Slovenia and European Union -- despite some methodological problems -- shows that the interest for these services is extremely high in Slovenia. Other figures also confirm that Slovenian households and businesses are generally on European average with respect to the penetration of the basic information technologies. However, certain discrepancies with other sources of data call for more efforts in performing these kind of analysis.

1 Preface

The concept of the information society has already been around for many decades. Nevertheless, its definition is relatively unclear, and the same is true for the corresponding indicators. There do exist some ad-hock measures from as early as sixties, particularly in the area of services, the information professions and the extent of the business information sector. However, it is extremely hard to establish official indicators for the phenomena in such a dynamic field. The Internet, in particular, has brought even more complications in these measurements. Even in United States, the official estimates about the scope of the electronic commerce will be available only in year 2000, after the electronic commerce transactions have already reached hundreds of billions of USD. The first official estimated will thus arrive after several years of extreme variety in the estimates from numerous consulting agencies. In addition, there are considerable discrepancies also in other measurements of information society. The paper presents an overview of most frequent divergences that arise from the interpretation of indicators of information society. The methodological misunderstanding is also an important reason that unnecessarily hinders the understanding of the position of Slovenia.

2 Quantitative measurement

Quantitative indicators of the information society usually refer to numerical figures -- expressed in numbers/percentages of users, or, in the form of financial totals -- which most often relate to the use and penetration of modern technologies, especially the Internet, mobile telephony and electronic commerce. However, we have to be extremely careful when interpreting these data.

We can, for instance, classify the use of electronic payment orders of the Slovenian Agency for Payments as a form of electronic commerce. Many experts claim that this is also a specific form o Electronic Data Interchange (EDI). In this case the amount of electronic commerce in Slovenia can be measured in hundreds of billions of USD. But if we talk about the transactions where searching, ordering, billing and payment procedures are all performed in the electronic forms -- without any paper recording -- it is clear that the amount of electronic commerce is only a fraction of this amount, e.g. only around few millions USD. Therefore, we have to be very precise when stating such observations. It is not surprising that the estimates of leading consulting agencies on electronic commerce have varied at rates 1:10 in the past years, and at present they still vary at the rate 1:2. Often, the source of the problems is not even in the statistical methodology or in the definitions, but in a simple fact that the methodological framework is not properly reported.

Recent international efforts for standardised measurement on electronic commerce, particularly those at OECD (1999) already brought some results and we can perform certain international comparison of the Internet and electronic commerce usage among companies. The available comparisons with Singapore, Scandinavian countries and Australia shows that with respect to PC usage, Internet penetration and Web site

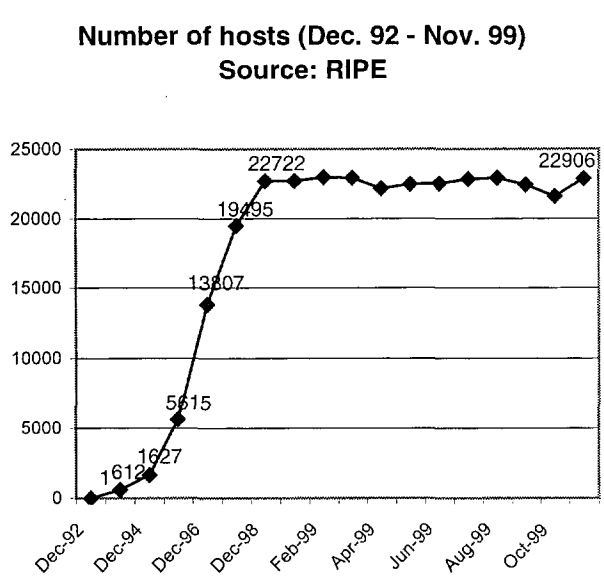
penetration among companies there is, as for 1998, no significant lagging for Slovenia. However, there is a certain time lag in the adoption of electronic commerce applications. Unfortunately, the international comparisons in the area of electronic commerce are much more complicated. Recent experiments in RIS 1999 survey of companies clearly demonstrated the sensitivity in these kinds of measurements. It has been shown that when electronic commerce was defined as any business transaction performed over the computer networks the percentage of companies claiming to use electronic commerce was 10% higher compared to the definition that was restricted only to the transactions that lead to the purchase (RIS, 1999).

A similar problem of definition is the estimate of the number of Internet users. There exist at least five categories of Internet users (Vehovar et al. 1998). This prompts to a need for an exact definition of the term "Internet user". For instance, the estimate of EITO (1999) talks about 60.000 Internet users in Slovenia in 1998, however, the definition of Internet user and description of how the estimate was obtained is not available there. In addition, this estimate differs from all other estimates for Slovenia. Even the number of users with personal e-mail accounts in 1998 is much higher than 60.000. Of course, EITO's estimate could refer to some specific group of intensive users.

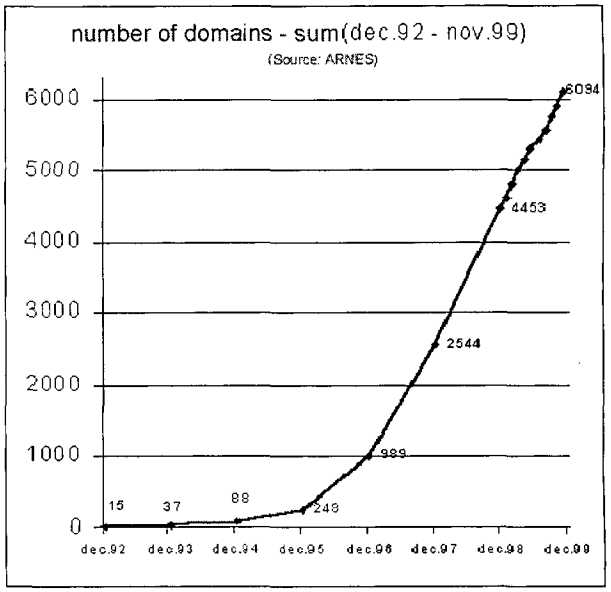
When measuring information society phenomena we are also faced with divergence, which originates from the methodology of data collection. The IDC corporation, for instance, provides estimates based on distribution channels and one of the figures states that 65.000 purchases (shipments) of new personal computers were made in Slovenia in 1998, thereof 17% in households. This suggests that households buy around 10.000 personal computers yearly. Such an estimate also matches the ESIS (1999) figures stating that Slovenian households possess 100.000 personal computers (with a processor 486 or more). However, this does not match the survey estimates that Slovenian households have more than 200.000 personal computers, which is a result of practically all surveys (Statistical office, Mediana, Slovenian public opinion, RIS...). Survey estimates consistently show that the number of personal computers has surpassed 200.000 also in business use, what suggests that Slovenia is highly ranked by number of personal computers per 100 residents. There are more than 25 personal computers per 100 habitants in Slovenia. This is surprisingly high, however, as the usage of information technology is rather complex, the criticisms regarding low technology penetration in Slovenian economy may still hold true (The World Competitiveness Yearbook 1999, IMD Lausanne).

One of the most exposed indicators of the Internet and information society is the statistics on the Internet hosts (Vehovar, 1998). This indicator shows extremely inconvenient trends for Slovenia: the growth of hosts in the last two years has almost entirely stopped while all other countries rapidly progress. However, the number of hosts is a typical example of an indicator that is more complex than a casual observer might think. For instance -- all the hosts which are not included in *.si domain are excepted from Slovenian host cunts. This does not happen so often in larger countries or in countries with more liberal legislation for assigning domains. In Slovenia, non-domestic domains are very frequent, even among the most visited sites and among the largest Internet access providers: siol.net, s.net, s5.net, amis.net. It seems that the large majority of commercial dial-up modems/ hosts is registered under domains *.net. The high usage of dial-up access in Slovenia also presents a problem for itself and contributes to a low host counts, because each host/modem serves many dial-up users. Additional problems can present the multiple IP numbers - e.g. virtual hosts - located on one computer. This is more often the case in countries such as Estonia than in Slovenia. We have to understand that the "host" does not necessarily mean a computer connected to the Internet, but only an IP number. In Slovenia, additional problems are also the computers that are connected to many large local networks with full access to the Internet but without an IP number.

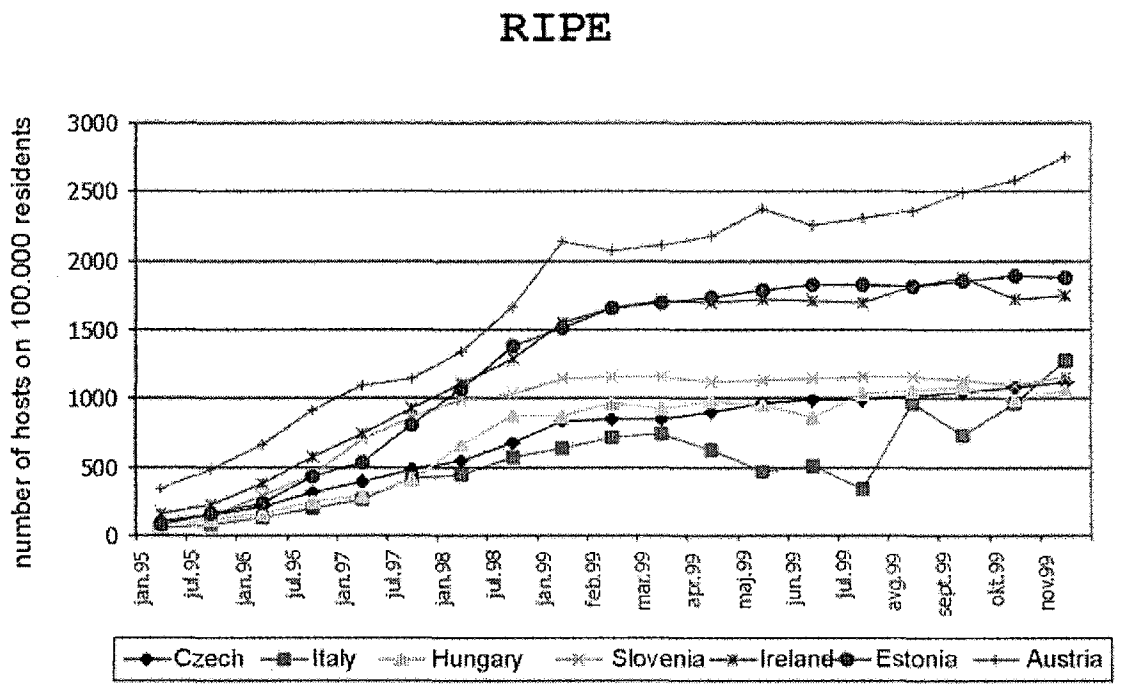
The problem of host counting is getting even more complex because of the technical problems of measurement procedures, which are becoming increasingly more difficult due to fire-walls and other forms of security protections. This forced Network Wizards to change entirely the methodology and broke with the time series. The data about host numbers from RIPE (http://www.ripe.net) and Network Wizard (http://www.nw.com) thus vary considerably. The RIPE host count often shows a clear monthly a recession in the number of hosts for some countries (Italy, for instance) what is not realistic. All the above arguments may explain the situation for Slovenia, where the host counts in the last two years show less than 10% yearly growth (Picture 1, Picture 3), but all other indicators (number of registered domains, number of companies connected to the Internet, number of households with access to the Internet, number of Internet users) demonstrated more than 50% growth (Picture 2).



Picture 1: Number of hosts in subdomain *.si (Source: RIPE)



Picture 2: Growth of registered domains *.si (Source: ARNES)



Picture 3: Hosts per 100.000 residents, January 1995 - November 1999 (Source: RIPE).

Another specific methodological problem can be observed in the areas, where growth is extremely high and the circumstances are changing very fast (in less than a year), as with mobile telephones. This example, however, has an additional dimension with the categories of mobile communications (i.e. GSM, prepaid mobile phones, etc) that are extremely structured and thus not comparable.

3 Measurement of attitudes

Above, we have observed that even the indicators that can be exactly measured on the ratio scale could be very problematic. It is thus reasonable to expect more troubles with the indicators referring to the attitudes towards the information society. However, the users' attitudes to the various aspects of the information society, such as the attitude towards security, privacy, abuse, role of the government, future and intended usage, are extremely important for understanding the context of information society.

Below, we present a typical example of attitudinal measurement. In June 1999, the RIS project and the

Institute for Economic Research (Incopernicus project) co-operated in a telephone survey on household usage of information technologies (n = 1000). Among others, the survey included questions on interest in services of the information society. The questions were based on the Eurobarometer EB 50.1 survey, which was performed among the member states of the European Union in fall 1998. The question was: "We will give you a list of services of the information society. Please tell us whether you are interested in their use or not." The June results for Slovenia were surprisingly high, so they were repeated also in September. However, the estimates were basically the same. In Table 1 we can observe the percentages of respondents interested in corresponding services in June and September for Slovenia, and the averages for EU at the end of 1998 (from EB 50.1).

Similar differences were noticed in other indicators, like the interest for financial management, virtual museum visiting, trip and travel planing information, consumer rights, and employment search (RIS, 1999). In addition, Slovenia has a good standing also in the area of information and communication technology penetration. In general it is on the average of the EU countries. The analysis shows that Slovenia is comparable to Scandinavia and not to Austria, which has certain historical linkages to Slovenia.

When trying to explain this specific position we first think of methodology. It is true, that the Slovenian survey was performed more than half a year after the European one. Certain methodological discrepancy could also arise from the fact that the European survey was conducted face-to-face. However, in Slovenia the telephone coverage has reached 90%, so this can not contribute to a significant overestimation. Alternatively, the overestimation could be related to the general overestimation of the information society phenomena -- for instance, the overestimation of the number of Internet users, the number of PC's in households, etc. However, whenever the external control is possible -- such as with mobile phones or Internet subscribers -- the estimates were correct. The results were also consistent across different surveys performed by Statistical office, RIS project, Slovenian public opinion survey, Mediana, and marketing agencies' surveys.

It is more likely that the proper interpretation of these surprising results goes into a direction of a specific climate that is very opened to the information society issues. Such a conclusion can be of the same importance for the understanding of information society processes as the standard statistical indicators (number of telephone lines, Internet users...).

4 Conclusion

Rapid technological developments bring new problems to the process of measuring social phenomena. The speed of the social change also introduces certain changes in measurement procedures. Of course, we have to be extremely precise about what these measurements actually mean.

We also have to accept the fact that the absence of more profound efforts for measuring and analysing information society phenomena prevents us from determining an exact position for Slovenia. Compared to other socio-economic statistics - like employment or national accounts that are permanently covered by large groups of experts - the research on information society indicators has not yet been instutionalized. Since the dynamics of these phenomena has become a constant, we can expect the establishment of adequate research in a very near future. The process of European unification is merely accelerating this process (i.e. ESIS, 1999). Lots of indicators on information society are also provided by global consulting agencies, which are partially compensating the lack of official statistical data.

We can thus conclude that we are in a temporary information vacuum where we do not have enough information to determine where exactly Slovenia stands in the information society developments. This is also a possible reason for the sensitivity of the data interpretation to a variety of sloppy analyses.

Information service	Slovenia - June '99	Slovenia – September '99	EU average '98
On-line medical diagnosis	54.2%	54.5%	41.9%
Contacts with politicians	18.7%	15.6%	10.9%
Education	55.2%	50.4%	33.9%
Consumer rights service	63.7%	55.6%	33.4%
Conducting public and administrative services	54.5%	58.6%	47.8%

Table 1: Interest for information society services (Source: EB 50.1, RIS)

5 Sources and List of References

[1] EITO (1999), Frankfurt: EITO.

[2] ESIS (1999), http://www.ris.org/id/esis.html.

[3] CMEC (1999), Conference on measurement of electronic commerce, Singapore, 1999, http://www.singstat.gov.sg/EC/echome.html.

[4] Network Wizard (1999), http://www.nw.com.

[5] OECD (1999), http://www.oecd.org/subject/e_commerce/.

[6] RIPE (1999), http://www.ripe.net.

[7] RIS (1999), http://www.ris.org.

[8] The World Competitiveness Yearbook 1999 (1999), IMD Lausanne.

[9] Vehovar, V., Remec, M., Kramberger, T. (1998): Statistika Interneta. Statistični dnevi 98, Radenci, Ljubljana: Statistični urad Republike Slovenije, Statistično društvo Slovenije.

[10] Vehovar, V. et al. (1998): Internet v Sloveniji. Ljubljana: Desk.

[11] Vehovar, V. (1999): Merjenje elektronskega poslovanja s pomočjo vzorčnih anket. Uporab. inform., 7, 2, str. 29-34.

[12] EuroBarometer (1999): The "Measuring Information Society" EuroBarometer, http://www.ispo.cec.be/polls/.

Modelling of an Information Society in Transition - Slovenia's Position in the CE Countries

Marjan Krisper, Tatjana Zrimec
University of Ljubljana
Faculty of computer and information science
Tel: +386 61 1768 390; fax: +386 61 1768 388
e-mail: {marjan.krisper; tatjana.zrimec@fri.uni-lj.si}

Keywords: modelling of socieconomics data, visualisation, information society

Edited by: Cene Bavec and Matjaz Gams
Received: October 4, 1999 **Revised:** December 3, 1999 **Accepted:** December 14, 1999

This paper presents modelling and visualisation of information society development in the six Central European countries associated to the European Union. Modern approach for monitoring and evaluation of the transition processes is presented that enables comparison of the countries' successfulness and analysis of alternative development scenarios. The position of Slovenia, which also experiences gradual transformation to the information society, is shown.

1 Introduction

Slovenia, as well as the other Central European (CE) countries is facing intense transition processes necessary for incorporation to the European Union (EU) and the global information society. These transition processes are especially related to their system development and strategy-based development including the information society capability of these countries.

The change of the political, economic, and legal system of CE counties is the basis for a gradual transition to a modern society. The aim of this transition is to position those countries into the group of open and competitive countries with more effective economies. This will also facilitate their prospective integration within European Union.

It has been recognised the necessity of Slovenia to take an active part in these movements for a gradual transformation to the information society. During the period 1991-1997 a relatively smooth transition of Slovenian S&T institutions into the market-oriented system was made possible (Stanovnik 1997). However, in the last few years, the information society issues have been changing much faster than can be monitored by statistical tools. The penetration rates for PCs, mobile phones, and Internet usage are often growing with the rates larger than 100% annually (Vehovar 1999).

The conventional econometric methods and mathematical models, which have been used for selecting appropriate strategies and testing development alternatives no longer provide comprehensive answers. Consequently, different approaches and techniques are required for modelling various aspects of the information society development.

Here we present an approach, developed under the Copernicus project (INCO 1999) for monitoring and evaluation of the transition processes of our country and other CE countries, which enables data structuring, visualisation, exploration and discovery of regularities. We illustrate the power of the clustering and visualisation techniques applied on statistical data relevant for information society development. The data consists of a selection of statistical indicators for six CE countries associated to the European Union.

2 Sources and methods

In many studies countries are compared and analysed on the basis of selection of statistical indicators that are relevant for various aspects of society development. Indicators that are estimated as relevant for information society are selected from different sources. The International Data Corporation, IDC Company (IDC), is the one of the main sources for the key indicators in the field of Information and Communication Technologies (ICT, EITO 1999), particularly for the countries in transition. For some indictors the EU average is added as the reference level for an international comparison. Detail description of the data can be found in (Bavec 1999). Also OECD (OECD) provides comprehensive and comparative data in this field.

In order to facilitate and make more transparent the evaluation and comparison of countries successfulness we have used different tools and techniques that enable

data structuring, visualisation, exploration and discovery of regularities. The following tools and systems were primarily used during the modelling processes:

a) *An Information Society In Transition Data Analysis system*, (ISIT) developed for intelligent analysis of socio-economic data;

b) *S-PLUS* , a commercial product for Data Mining;

c) *See-5*, a commercial product for Machine Learning.

a) The ***ISIT Data Analysis*** *tool* (Krisper et. al. 1999) for data modelling enables expert system approach to construction of a knowledge base of the socio-economic domain. This approach in contrast to conventional econometrics or statistical methods and mathematical models gives clear insight into the internal structure of the domain. It enables a multi-dimensional treatment of socio-developmental problems from the viewpoint of the information society. The system offers high level of flexibility in modelling of development goals covering various aspects of development. In practice, ISIT system has shown that it is a useful tool, particularly suitable for comparison of successfulness of the countries in transition. Various comparisons facilitate the choice of suitable developmental scenarios.

b) ***S-PLUS*** (S-PLUS 1998) is a powerful tool for exploratory data analysis and statistical modelling. It is known as a tool for manual Data Mining.

c) ***See-5*** (Quinlan 1993) is a tool for machine learning. It enables discovering patterns in the data that are represented as decision trees or production rules.

The modern techniques for data analysis, especially clustering and machine learning are more powerful then the classical statistical tools since they enable:

- domain structuring,
- non-intuitive modelling,
- transparency,
- knowledge elicitation,
- better explanatory power.

In this paper we illustrate the power of those techniques applied on statistical data.

3 Modelling and visualisation

In our experiments, a set of statistical indicators that are relevant for information society development in the six CE countries associated to the European Union have been selected. The EU average is added as the reference level for an international comparison.

A common problem with all selected countries is a lack of official national indicators gathered by internationally compatible methodologies. Presented indicators are a subset of available national data (EITO 1999) that are reasonably comparable in all countries, and particularly with the EU data. Indicators are grouped as follows:

- Political background
- Economic background
- R&D and Technology
- Legal and Organisational background
- Information Technology
- Communications Infrastructure

Graphical representation of the expert developed hierarchical information society indicator model (Zrimec & Krisper 1999, IMD 1997, IMD 1998) is presented in Figure 1.

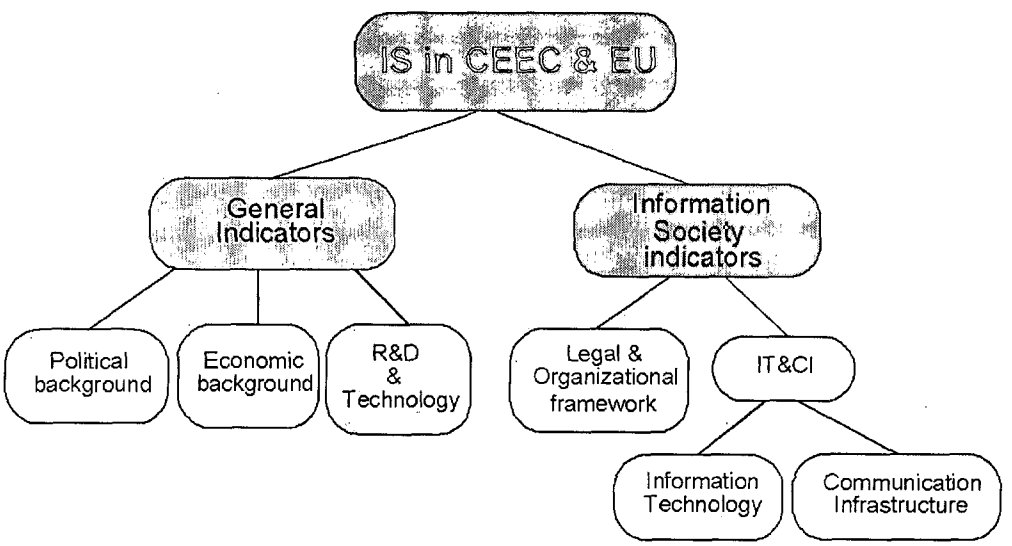


Figure 1: *Information society indicators model*

In order to provide better representation of the observed indicators and the influence of those positioning and grouping the countries, further modelling was performed. For each level of the model, the agglomerative hierarchical clustering method, one of many available S-PLUS methods, was applied. The observed countries were differently grouped together on the basis of self-similarity of the indicators of a particular group. S-PLUS provides a good graphical presentation of the data analysis results. The results of the clustering are represented by cluster trees - dendograms. This graphical presentation, in a form of a tree very clearly indicates the position of each country. The tree representation shows how the countries can be grouped and how well thy are positioned in comparison to EU.

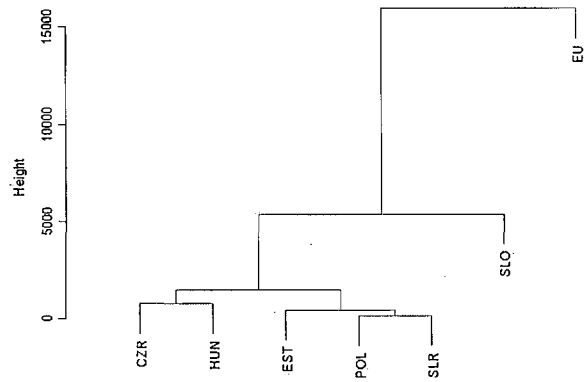


Figure 2: *Results of the clustering using the indicators for Economic Background*

3.1 Political background

A *political* awareness for a transition to the society in CE countries is presented by their membership in major international organisations. The membership indirectly represents their openness and attitude towards the global character of the information society (Krisper et. al. 1999a). These indicators are significant and show two groups of CE countries: Czech Republic, Hungary and Poland are member of all selected organisations, Estonia, Slovakia and Slovenia are not in OECD and NATO.

3.2 Economic background

Economic indicators show significant differences in general development level (clear lead of Slovenia). It is noticeably that economic development is not correlated with the direct foreign investments, which also represent the "openness" of the economy. It is interested that these indicators are correlated with a political "openness" presented in the *political background.* We have used the set of economic indicators and have applied clustering. The results are shown in Figure 2. It can be seen clearly the position of Slovenia and EU (Figure 2).

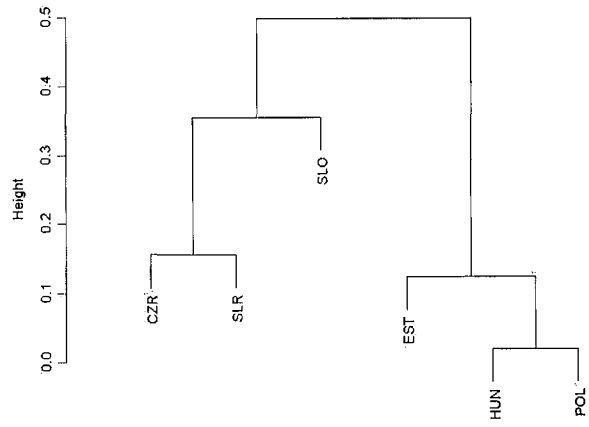


Figure 3: *Clustering of the domain: R&D Technology*

3.3 R&D and Technology

Investments into *R&D* and *Technology* represent a general attitude towards high technologies that could also reflect an absorption level for Information technologies in particular countries. To give better representation for comparison of the CE countries a tree representation, produced by clustering is shown in Figure 3. Because the indicators show a significant lower level than EU average in Figure 3 EU is excluded.

3.4 Legal and Organisational background

Legal and *organisational* framework is not represented by qualitative indicators but rather by qualitative descriptions. All CE countries are gradually implementing the EU "acquies" and face very similar problems concerning organisation and co-ordination of information society activities society (Krisper et. al. 1999a, Bavec 1999). Basic legislation concerning telecommunications is already prepared and will be harmonised with the EU in next three years. Legislation concerning electronic commerce is still in an early phase of preparation. Noticeable is a lack of government administrative structure that could support information society activities. The values of some indicators are given in Table 1.

3.5 Information Technology

Absolute CE countries investments into *Information Technology* are significantly lower than EU average is, but with the same growth rate as in the EU. Differences between individual countries are relatively small. Figure 4 shows a comparison of the GDP/pc (E1) indicator, and the indicator IT21, which represents % of PC connected to the Internet. Slovenia has a leading position in both categories.

More significant are differences in the quality and availability of *Communications Infrastructure.*

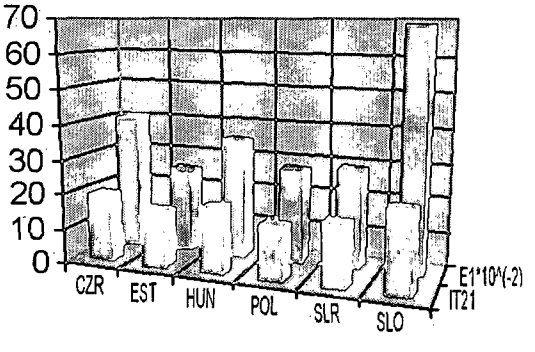
ID	DESCRIPTION	REMARKS	EU	Czech Republic	Estonia	Hungary	Poland	Slovak Republic	Slovenia
				INFORMATION SOCIETY INDICATORS					
	LEGAL AND ORGANIZATIONAL FRAMEWORK								
1	Monopoly on voice telecommunications		1.1.1998	1.1.2001	1.1.2001	1.1.2002	1.1.2003	1.1.2003	1.1.2001
3	Level of telecom. competition	Range 1-5	very high (5)	medium (3)	medium (3)	high (4)	medium (3)	low (2)	low (2)
5	Personal data protection law	Range 1-5	YES (5)	Draft law (2)	Draft law (2)	YES (5)	Draft law (2)	?	YES (5)
6	Digital signature law	Range 1-5	YES (5)	Draft law (2)	NO (1)	Draft law (2)	NO (1)	NO (1)	Draft law (2)
8	Government approved Action Plan on IS	Range 1-3	YES (3)	NO (1)	NO (1)	Governm ent only (2)	NO (1)	NO (1)	Governm ent only (2)
9	National ISPO	Range 1-3	YES	partially (2)	NO (1)	partially (2)	NO	NO (1)	partially (2)
11	Government coordinating body		YES	NO	NO	YES	YES	NO	YES
12	Import duties on IT	Range 1-5	very low (1)	low (2)	medium (3)	high (4)	medium (3)	high (4)	very low (1)

Table 1: *A selection of information society indicators for Legal and Organisational framework, source: Agenda 2000*

3.6 Communications Infrastructure

Recently *Information Technology* and *Communications Infrastructure* are converging. As a result of their convergence a synergetic effect is achieved. Communications Infrastructure is now of the same importance as Information Technology for information society development.

In Figure 5 the results of the clustering on the bases of the Communication Infrastructure indicators are shown. Figure 6 represents a two dimensional representation – portfolio of the relation between Telecom services growth in 1997/98 (IT21) and GDP per capita in EUR for 1997 (E1). The diagram clearly shows the position of the CE countries in comparison to EU. The portfolio representation was generated using the ISIT system (Krisper et. al. 1999) which offers alternative ways of good visualisation.



	CZR	EST	HUN	POL	SLR	SLO
IT21	19.8	16.1	18.7	15.2	17.4	23.1
E1*10^(-2)	40	24.67	34.71	27.37	28.52	69

Figure 4: *E1*10(-2) represends GPD/pc, IT21 represents percentage of PC connected to the internet*

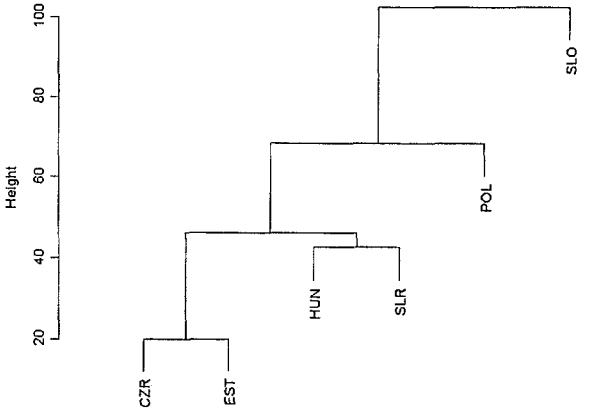


Figure 5: *Clustering using Communication Infrastructure indicators*

4 Conclusion

A common problem with the selected CE countries is a lack of official national indicators gathered by internationally compatible methodologies. We can use only a subset of the national data that is reasonably comparable in all countries, and particularly with the EU data. From the results presented in this paper the following conclusion can be made:

- Indicators on some areas are not correlated with economical power of the countries;
- Absolute investments are significantly lower then EU average;
- Administrative structure is still developing;
- Legal and organisational framework is rapidly improving;
- Level of capitalisation in telecommunication is improving although some monopolies still exits;
- ICT advanced development is beyond legal, organisational and administrative progress.

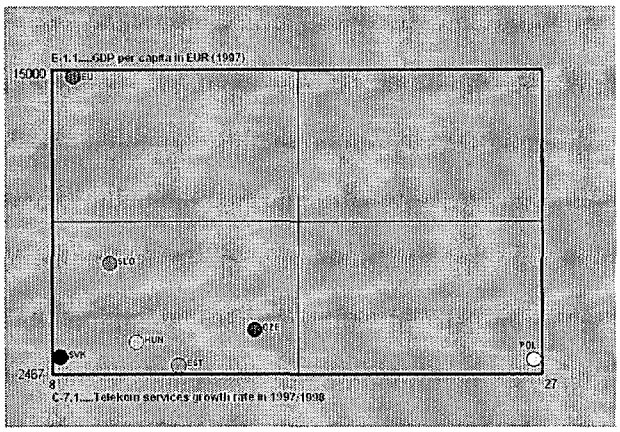


Figure 6: *Portfolio showing GDP/pc and Telecommunication growth rate in 1997/1998*

The proposed approach to data analysis and the developed tools are not only good for comprehensible modelling and presentation of the results, but also for permanent monitoring and studying of Slovenia and other CE countries' progress. Figure 7 shows a six dimensional diagram for CE countries comparison, based on the total values of six indicator's groups. Very clearly can be seen the leading position of Slovenia (SLO-white). This diagram and the diagram in the figure 6 were generated using our ISIT system.

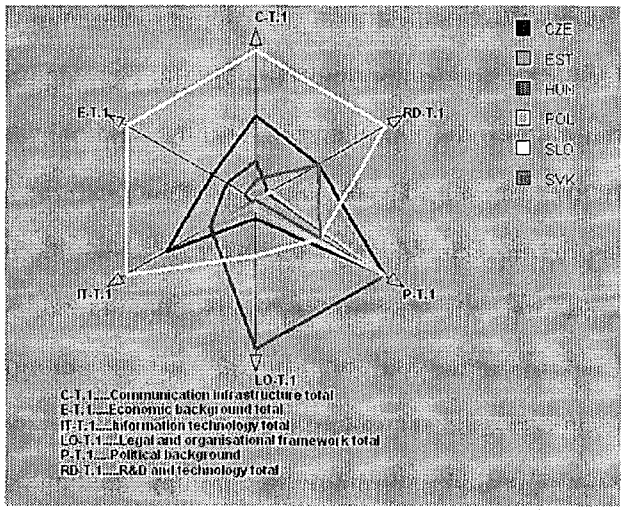


Figure 7: *Six dimensional diagram for CE countries comparison, based on the total values of six indicators groups (note position of SLO)*

5 Acknowledgement

The authors thank Cene Bavec for providing and preparing the data for the experiments. We would also like to thank Matej Grom for his help in using the ISIT system and Zoran Bosnić for the help in preparing the document.

6 References

[1] Stanovnik P. (1997) Some Characteristics of the RTD System in Slovenia with the Perspective to Co-operation within EU Programmes, (*IER, Ljubljana,* p. 12), *IER-CRII*, No. 6.

[2] Vehovar V (1999) Information society measurements - the usage of the Internet in Slovenia, (http://www.ris.org/), *CRII-TR6*, No. 7, p. 5.

[3] INCO (1999) The Role Of Information Society In Transition Processes, *Copernicus project 960154, CRII 1996-99,* (http://lpo.fri-uni-lj.si/crii/).

[4] Bavec C. (1999) Information society indicators in Central and Eastern European Countries: table of the indicators for CEEC and EU, *Report of CRII final workshop, CRII-TR6*, No.2, p.7-9.

[5] IDC, http://www.idc.com

[6] OECD, http://www.oecd.org/dsti/sti/it/index.htm

[7] Krisper M., Zrimec T., Grom M. (1999) ISIT-Information Society in Transition: software package and database with time series, *CRII-TR6*, July, No. 6, p. 16-19.

[8] S-PLUS (1998) Knowledge Discovery without Limitations, *MathSoft International, Statsci Europe,* Oxford.

[9] EITO (1999), EITO'99, http://www.fvit-eurobit.de/DEF-EITO.HTM

[10] Zrimec T., Krisper M. (1999) Intelligent data analysis using Clustering and Machine Learning techniques, *CRII-TR6*, No. 3, p. 1.

[11] IMD (1997) The World Competitiveness Yearbook, *International Institute for Management Development,* Switzerland.

[12] IMD (1998) The World Competitiveness Yearbook, *International Institute for Management Development,* Switzerland.

[13] Quinlan J.R., (1993) C4.5: Programs for Machine Learning, *Morgan Kafman Publishers*, San Mateo, CA.

[14] Krisper M., Zrimec T., Bavec C. (1999) Information Society In Transition Of Central European Countries - Position of Slovenia, *in* C. Bavec & M.Gams, eds, *Proc. Information Society Conference*, Ljubljana, p. 22-25.

Customer Satisfaction of Information System Integration Business in Japan

Kayo H. Iizuka
Graduate School of Systems Management, University of Tsukuba
3-29-1, Otsuka, Bunkyo-ku, Tokyo Japan, iizuka@gssm.otsuka.tsukuba.ac.jp
AND
Mitsuo Wada
Graduate School of Business Administration, Keio University
2-1-1, Hiyoshi-Honcho, Kohoku-ku, Yokohama Japan, HQB00473@nifty.ne.jp

Keywords: customer satisfaction, system integration

Edited by: Cene Bavec and Matjaz Gams
Received: October 2, 1999 **Revised:** December 3, 1999 **Accepted:** December 19, 1999

When we consider about "customer satisfaction" (CS) of system integration (SI) business, "customer" means organization, though many of the cases, when we see the CS in some research papers, "customer" means personal consumer. Understanding the satisfaction structure of organization, being considering organization behavior, must be very important Moreover SI business provides, not only products itself alone but integrated services and products, In this paper, focusing on these matters, showing the structure of CS in SI business, from statistical analysis of customer survey. It shows what is the important factor including technology providing, and project management skill, in order to maximizing CS.

1 Introduction

Almost a decade has past since the word "customer satisfaction (CS)" became one of the most important topics of business.
M.Hanan and P.Karp[1989] insist that, customer satisfaction is one of the most effective things for competitive advantage.

Thinking about CS for system integration business, we can easily imagine a lot of factors to characterize customer satisfaction. First of all, who is the customer? Most of the cases in Japan, purchasing section are IT sections. But of course, IT section people are not the only customer. Next question comes to how the satisfaction structured, though system integration business provided many IT products and their integration services.

If information system integrators can get answer of these questions, they can focus on effective factor for improvement.[1]

2 Literature Review

Two areas of theories identify from literatures, which are needed for designing analysis approaches are: theories about the structure of customer (consumer) satisfaction, and theory of organization behavior.

[1] In this research, "information system vendors" means those provide hardware, software, integration service, and system development services to customer.

2.1 Theories of the Structure of CS

- Degree of Attainment and Range of Desire

Shimaguchi [1986] proposed to define consumer satisfaction, using H. Thourough' s definition of "degree of warfare". "Customer satisfaction can be described by his/her range of desire and power of attainment"

$$F(p,q) = \frac{q}{p}$$

F(p,q): Degree of Consumer Satisfaction
p: range of desire
q:power of attainment

- Function Fufillment

Swan and Com [1976] described consumer satisfaction as function fulfillment.

Essential Function	Surface Function	Satisfied/ Not Satisfied
>=E	>=E	Satisfied
>=E	<E	Not Satisfied
<E	>=E	Disappointed
<E	<E	Disappointed

E:expectation

This theory is based on the idea that essential function effects whether satisfied or disappointed, while surface function effects whether satisfied or not satisfied.

- Performance vs. Expected

Anderson and Dover described satisfaction as below.

$$S = \alpha E + \beta(P - E)$$

S: degree of satisfaction
E: degree of expectation
P:degree of performance
α, β: parameters

2.2 Organization Behavior

- Organization Buying Behavior

J.N. Shess [1977] defines three correlates for organization buying behavior: individual correlates, organizational correlates, and situation correlates. The elements of organizational correlates are lateral and vertical involvement, role for purchasing, demographics, and organization style.

3 Research Design

In order to clarify CS in SI as organization satisfaction in Japan, K. Hirotsu Iizuka made a customer survey in 1992.

The customer survey was targeted on Japanese major and industry leading companies (that we can see their name in Tokyo Exchange Market).[2] Of course they are big users of system integration business. And Survey sheet was mailed to 3 sections to each company. (For IT section, one of the end user's sections, and Business Planning Section).

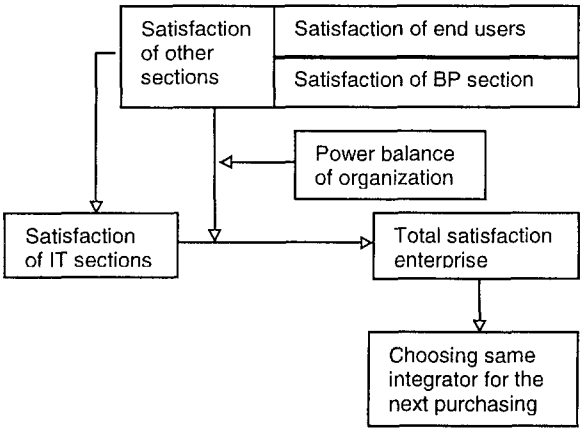


Figure 1: Research Model

In Figure 1, BP (Business Planning) section means a section, which helps Top management's decision.

[2] For the detail of respondent companies, refer Kayo Hirotsu Iizuka, *Customer Satisfaction of SI Business* , Keio University Graduate School of Business Administration, Master's Thesis, 1993

In this research, the hypotheses to be verified are below.

In, SI business

- *Hypothesis1:* Total CS at this time effects next purchasing
- *Hypothesis2:* CS of each one factor of SI business can be explained by "degree of attainment and range of desire".
- *Hypothesis3:* CS of each one factor of SI business can be explained by "performance vs. expected".
- *Hypothesis4: Inside one enterprise,* CS for a certain system integrator is different by section
- *Hypothesis5:* CS of each one element of SI business are different by section
- *Hypothesis6:* CS of one section can be influenced by other section

Hypothsis1 is about structure of CS, and 2 and 3 is for about system integrator, 4 and 5 for structural difference of CS by section.

There used be some customer survey about system integration business in Japan, but it was targeted on system integration skill only, and respondent were only IT section. For the research explains here, the elements for system integration was defined as below.

- *About System Integrator's Skill*
  - System Design/Development Skill
  - Knowledge about Business and Customer's Industry
  - Total System Proposal Skill
  - System Management Skill
  - System Consulting Skill
  - Management Consulting Skill
  - System Maintenance Skill
  - Special Skill Level about Advanced IT
  - Company Ability to Avoid Risks
  - Supporting Services for Trouble
  - Integrate Products of Multi Vendor's Products

(Questions for IT section)
- Satisfaction of end users' section
- Satisfaction of BP section

- *About Product Specification*
  - Ergonomics
  - Specification of the Each Products
  - Cost-Effectiveness

- *About Brand Image for System Integrators*
  - Has good history
  - Trustful
  - Stable
  - Friendly
  - Has future perspective
  - On current trend
  - Has good atmosphere
  - Has great Top Management
  - Offers customer services

Survey respondents were provided degree of expectation, degree of attainment (for the current system provided by system integrator), and degree of satisfaction for the element listed above. And asked to rate each element on a scale of 1 to 5 as below.

- *About expectation*
  5:Very important
  4:Important
  3:Neutral
  2:Not important so much
  1:Not important

- *About attainment*
  5:Very attained
  4:Attaind
  3:Neutral
  2:Not attained very much
  1:Not attained

- *About satisfaction*
  5:Very Satisfied
  4:Satisfied
  3:Nutral
  2:Not satisfied
  1:Disapointed

4 Research Results

Response rate was very high more than expected as you can see in the Table 1. Before mailing the survey sheets to company, explanation by telephone was held in order to let respondents know the purpose of the survey. That may had made sense to respondent, and more over this rate shows so many people were interested in this theme.

Table 1: Research Respondent Characteristics

Section	Total Mailed	Total Received	Response Rate(%)
IT Sections	236	97	41.10%
BP Sections	236	60	25.42%
End Users	236	69	29.24%

We can see power balance of the sections for purchasing from Table 2.

Table 2: How affect does the section to decision of purchasing?

SECTIONS	Affect Percentage (SD)
IT Sections	52% (2.851)
BP Sections	26% (2.514)
End User's Sections	22% (2.252)

4.1 Verification

From the research, you can see verification result in Table 3.

You can see all the hypotheses are verified within 5% level of significance.

4.2 How to Maximizing CS?

"How to maximizing CS of SI business" must be one of the most important strategic questions for system integrators?
From the power balance we can see Table 2, total CS of company level can be defined as algebra a below.
Satisfaction of each section can be described as algebra $b,c,d,$ from multi regression.

$a = 0.52b + 0.26c + 0.22d$
$b = 0.55d + 0.27e + 0.19f - 0.13$
$c = 0.51g + 0.55h + 0.29I - 0.26j - 0.37k + 0.38f + 0.211 - 0.41e - 0.35$
$d = 0.51g + 0.44m + 0.23f + 0.29n - 0.19o - 0.14p - 0.44$

a:Total Customer Satisfaction of Company Level
(Total CS for sections, and elements)
b:Total Customer Satisfaction of IT section
c:Total Customer Satisfaction of BP section
d:Total Customer Satisfaction of End User's Section
e: System Maintenance Skill
f: Knowledge about Business and Customer's Industory
g: System Design/Development Skill
h: Cost-Effectiveness
i: Company Ability to Avoid Risks
j: Skill Level about Advanced IT
k: Trustful Company
l: Integrate Products of Multi Vendor's Products
m: Ergonomics
n: System Management Skill
o: System Consulting Skill
p: Specification of the Each Product

Table 4: The Element Affect to Total CS

Elements of System Integration	Rank
System Maintenance Skill	A
Knowledge about Business and Customer's Industry	A
System Design/Development Skill	A
Cost-Effectiveness	C
Company Ability to Avoid Risks	C
Skill Level about Advanced IT	C
Trustful Company	C
Integrate Products of Multi Vendor's Products	C
Ergonomics	B
System Management Skill	A
System Consulting Skill	C
Specification of the Each Products	C

A: Affect Total CS a lot
B: Affect Total CS
C: Affect Total CS a little

Table 3: Summary of hypothesis verification result

Analysis Objectives	Hypothesis	Verification Method	Level of significance
Importance of CS	Hypothesis	Regression of Degree of satisfaction and will for selecting same system integrator for the next purchasing time	**
Structure of CS for each element of SI business - Degree of Attainment and Range of Desire	Hypothesis	*Regression of* *Degree of satisfaction and q/p* (p: range of desire, q:power of attainment)	***
- Performance vs. Expected	Hypothesis	Multi regression of Degree of satisfaction and E, Degree of satisfaction and (P-E) (E: degree of expectation, P:degree of performance)	***
Difference of Satisfaction (by section)	Hypothesis	Difference of mean degree of satisfaction by section Correlation of satisfaction degree of the three section	**
Structure of Total Satisfaction by section	Hypothesis Hypothesis	Multi regression of Total Satisfaction degree and degree of satisfaction degree for each element	***

Level of significance
*:10%, **:5%, ***:1%

From the simulation for the purpose to maximize a, you can see what element of SI must be strengthened from Table 4.
Some of the element that seems to be very important factor of system integration business marks "C". But it does not mean that factor is not important. Some of the standard factors are usual thing to customer at a certain level.[3]

5 Conclusion

The research showed what is important and effective for the information system integration business in Japan.

In SI business, of course, technology issues are quite important. But the level of customer satisfaction does not in proportion only to level of technology. Though, CS is not the only important factor for IS business, considering why that technology is important and how to use it how to integrate it to total system to maximizing the CS must be indispensable.

This survey is targeted in Japanese market, but some of the factors can be used to system integrators in other countries, and moreover the framework of research is reusable for all the companies whose customers are organization to improve their customer satisfaction.

[3] For the detail of research result, refer Kayo Hirotsu Iizuka, *Customer Satisfaction of SI Business* , Keio University Graduate School of Business Administration, Master's Thesis, 1993

6 References

[1] M.Hanan, and P.Karp, *Customer Satisfaction*, American Management Association, 1989

[2] M.Shimaguchi, *Total Marketing*, Nihon Keizai Shimbumsha, 1986

[3] J. N. Shess, *Industrial Buying Behavior,* North-Holland Publishing. Co.1977

[4] H. Akuto, *Handbook for Social Research*, Nihon Keizai Shimbumsha, 1987

An Infrastructure For Support Of Digital Signatures

Tomaž Klobučar and Borka Jerman-Blažič
Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
Phone: +386 61 1773 900, Fax: +386 61 1232 118
E-mail: klobucar@e5.ijs.si

Keywords: digital signature, public-key infrastructure, technical framework, legal framework

Edited by: Cene Bavec and Matjaž Gams
Received: October 10, 1999 **Revised:** November 5, 1999 **Accepted:** December 14, 1999

In this paper, we present an infrastructure for support of digital signatures in the Information society. Technical aspects are briefly described and a short overview of several existing legal frameworks is given. Certification authorities, certificate policies, signature policies and certification practice statements are identified as important parts of the infrastructure.

1 Introduction

Provision of security is one of the most important issues in the Information society. Several security aspects should be taken care of when communicating on global computer networks. Confidential information must not be made available or disclosed to unauthorised subjects, and users must be able to authenticate other users or check that the source of data is as claimed. In addition, unauthorised data modification must be detectable, methods for the prevention of unathorised use of resources, including the prevention of use of a resource in an unauthorised manner must exist, and falsely denying of participation in certain activities should be prevented. These security requirements are characterised with the following concepts, known as security services: confidentiality, authentication, integrity, access control and non-repudiation. There exist several security mechanisms and structures for security provision, one of them being digital signatures, which are used to provide authentication, data integrity and non-repudiation.

A digital signature means data in electronic form which are attached to, or logically associated with, an electronic message and which serve to ascertain both the originator of the message and the fact that the message has not been modified since it left the originator. Digital signatures are used in various places in the Information society, for example in communication with public institutions (e.g. calls for tenders, exchange of application forms, tax declarations, transmission of legal documents), electronic buying and selling, electronic financial transactions, as well as for personal purposes, such as personal electronic mail, or for identification in the Internet. In order to use these signatures with equivalent legal effect as hand-written signatures in non-electronic documents, certain conditions have to be met. In the next two Sections, technical and legal aspects of digital signatures, and current activities in this field are briefly described. Conclusions are given at the end.

2 Technical framework

A technical framework describes a set of security mechanisms, technologies and technical standards that are used to support digital signatures. Management requirements for supporting those mechanisms are also included. Digital signature methods are based on a public-key technology, which has been widely recognised as a fundamental technology for providing secured Information society. Although the definition of digital signatures is technology neutral, public-key cryptosystems are in practice almost always used for signature creation and validation of signed data.

Two distinct keys are used in public-key cryptosystems - one for encryption and the other for decryption. Anything encrypted with the first key can only be decrypted with the second key. Although both keys are mathematically related, it is computationally unfeasible to derive one key from the other without additional (secret) information. One of the keys (public key) can thus be published, allowing everyone to perform encryption or signature validation with this key, while only a user knowing the corresponding private key can decrypt or sign a message.

There exist several different methods for generating digital signatures, most often being used DSS (Digital Signature Standard) [14] and RSA [17] in combination with one of the one-way hash functions, such as SHA-1 [15], RIPEMD-160 or MD5 [16]. Digital signatures are generally produced in a two step process. Firstly, the message is compressed using a one-way hash function which transforms the information into a string of fixed length, then this so-called message digest is encrypted with the user's private key. As the private key is known only to the originator then no one else could forge its signature. Every slightest change of the document after it being signed produces changes in the digital signature that make the signature not anymore valuable. The recipients verify the signature with the originator's public key.

In order to verify the signature, the recipient must first authenticate the originator's public key. The identity of the public key owner and public key integrity are guaranteed with public-key certificates. A public-key certificate is a digitally-signed data structure which securely binds a public key to the entity's identity. A digital signature is produced with the private key of a trusted entity, called a certification authority (CA), which vouches for the correctness of the information included in the certificate. The de-facto standard public-key certificate format is defined in ITU-T Recommendation X.509 [11]. The first two versions of this document, published in 1988 and 1993, described the version 1 (v1) and version 2 (v2) formats. Since then, several deficiencies have been found and new security requirements identified. As a result, optional extension fields were added into the version 3 format [9, 10], which was standardised in 1997. Besides a public-key and the name of its owner, an X.509 public-key certificate also contains the name of the issuer of the certificate, validity period, serial number and algorithm identifiers. Standard extensions in version 3 certificates define additional key and policy information, subject and issuer attributes, certification path constraints, and information about certificate revocation lists (CRLs). Supported extensions and their semantics for particular use are defined in certificate and CRL profiles. Two important profiles for the use of public-key certificates for digital signatures are "Internet X.509 PKI Certificate and CRL profile" [6] and "Internet X.509 PKI Qualified Certificates" [18]. The first document profiles the X.509 certificates and CRLs for the use in the Internet, while the second describes certificates which are qualified to support digital signatures in a context which is considered to be functional equivalent with hand-written signatures.

2.1 Public-key infrastructure

Certification authorities are an essential part of the infrastructure for the use of digital signatures. A system of certification authorities with supporting registration authorities (RAs) and other agents and servers that provide services that are needed if public-key-based technologies are to be used on a wide scale is called a public-key infrastructure (PKI). The core services of a PKI are registration and identification of users, issuance of certificates, directory services, certificate distribution, archiving services, revocation of certificates, publishing of CRLs, and time-stamping services.

Public keys can be used for different purposes and in different environments, such as military, commercial, research or educational. Since these environments do not have the same security requirements, there will certainly not exist a single public-key infrastructure for the whole Internet. At the moment, there are several, generally isolated and hierarchically structured public-key infrastructures in the world [1]. Most of them are governmental, such as the Government of Canada Public Key Infrastructure, or commercial and consisting only of a few certification authorities, e.g. Verisign or Thawte. A common world PKI for a business-to-business electronic commerce, named Identrus, is also in the process of building-up by several global financial institutions. In Europe, an ICE-TEL public-key infrastructure was established a few years ago with the aim to provide a large-scale public key certification infrastructure in a number of European countries for the use of security services based on public keys. The ICE-TEL (Interworking Public Key Certification Infrastructure for Europe) project was part of the 4th Framework Programme of European Community activities in the field of Research and Technological Development. Tools for the provision of the infrastructure, security toolkits and user tools (e.g. secured e-mail programs) were also developed during the project, which ended in February 1998.

The successor to this project, the ICE-CAR project (Interworking Public Key Certification Infrastructure for Commerce, Administration and Research) [8] from the same 4th EU Framework Programme, was launched in 1999 to foster the development of European-based security technology for the purpose of securing the growing applications of the Internet for administration, electronic commerce, intra-organisation communication, health care applications and research. Solutions to the problem of authenticity, integrity and privacy on the Internet are offered through the tools for secure communication, the provision of the public key infrastructure and the support of users of the PKI. One of ICE-CAR activities is also the transfer of security technologies and services to Central and Eastern European countries through the Security Technology Competence Centre for Central and Eastern European Countries (SETCCE). ICE-CAR is continually improving and enlarging the European PKI set up within the ICE-TEL through integration of public key based security services into applications which make use of the PKI.

A part of this infrastructure, which connects CAs from different European countries, is Slovenian certification authority SI-CA [13]. SI-CA, which has been certified by the ICE-CAR top-level certification authority, certifies individuals and other CAs in Slovenian academic, research, governmental and commercial organisation. Issued certificates for Web clients and servers can be used to secure Web transactions, while public-key certificates for e-mail help users to provide authenticity, integrity and confidentiality of their e-mail messages. Adopted ICE-TEL certificate policy, which has been assigned the unique object identifier 1.3.6.1.4.1.2712.10, specifies that the identity of the certificate applicants must be verified with official documents only, e.g. passports or personal id-cards, and that a personal presence of the applicant is required during verification. This part of the ICE-TEL certificate policy is very important, because it assures verifiers of digital signatures and other certificate users that the owners of the keys have been properly identified and authenticated before the certificate issuance.

2.2 Certificate policies and certification practice statements

Certificate policies are crucial for the operation of global public-key infrastructures since they define where, when and how the public-key certificates and public keys are used. In the X.509 Recommendation, a certificate policy is defined as "a named set of rules that indicates the applicability of a certificate to a particular community and/or class of application with common security requirements." Certificate policies generally include basic information about the policy authority that defined the policy, community and certificate applicability, as well as certificate and CRL profiles, security requirements, requirements for subject identification and authentication, and obligations for CAs, RAs, owners of certificates and relying parties. A certificate policy may, for example, state that the keys for digital signatures must be stored on smart cards and can only be generated by the users themselves. Types of documents which are allowed to be signed can also be specified. Operational procedures of a CA are described in more detail in a CPS that is defined as "a statement of the practices which a certification authority employs in issuing certificates" [3]. Unfortunately, there is still no clear distinction between certificate policies and CPSs. Both will generally contain similar information - only that in policies this information will be less detailed. Each CPS is also specific to one CA, while certificate policies are widely supported by more than one CA. High-level topics which need to be part of certificate policies and certification practice statements are informally described in RFC 2527.

Applied certificate policies are identified in public-key certificates by unique, registered object identifiers. Knowledge about the policies is necessary during digital signature validation in order for users to evaluate the binding of a public key and the originator's identity, and to decide whether the key was meant to be used for a particular purpose or by a particular application. A user may, for example, not completely trust a certificate for the use in a financial transaction knowing that the CA that issued this certificate accepts no liability for its services and does not verify the identity of a certificate applicant with official documents. Certificate policies thus need to be evaluated by the users and compared against their personal requirements or local security policies. Unfortunately, policies are still written in natural language in different forms, which do not allow automatic processing. A format for a formal presentation of certificate policies was proposed in [12] to facilitate their comparison and evaluation. This format:

- Helps users in to decide more easily which policies satisfy their requirements, and which certificates can thus be accepted by their applications, e.g. secure e-mail programs or electronic commerce applications.
- Helps to modify applications to support accepted certificate policies, i.e. recognise them and conform to the semantics of the policies.
- Helps CAs in policy development. By selecting different explicit policy elements from a template they can prepare their policies in a more efficient way.
- Helps CAs to decide which certificate policies from other security domains can be considered equivalent.
- Helps CAs to decide whether the applicant CA's policy is in accordance with their policies.

The proposed specification of a formal certificate policy in ASN.1 is as follows:

```
CertificatePolicy    ::=  SEQUENCE{
 version                   Version DE-
FAULT v1,
 policyId                  OBJECT IDENTI-
FIER,
 policyDescription         DisplayText,
 generalInfo               GeneralInfo,
 issuerCAPolicy            Subpoli-
cyRules,
 subordinateCAPolicy [0]   SubpolicyRules
                             OPTIONAL,
 rAPolicy            [1]   SubpolicyRules
                             OPTIONAL,
 userPolicy          [2]   SubpolicyRules
                             OPTIONAL
}
```

A detailed description of the elements, their order relations and an algorithm for comparison of policies are omitted from this paper. Policies are distinguished on the basis of different levels of security control, different levels of thoroughness of applicant authentication, the complexity of operational procedures, and restrictions on certificate usage and applicability.

Certificate policies are therefore used to specify conditions which have to be met during the use of public-key certificates and public keys for digital signatures. There exist other rules which are specific to digital signatures and are not contained in the certificate policies. These rules are part of signature policies, which define the technical requirements on signature creation and validation. An ASN.1 specification of the signature policy is described in the document, prepared by ETSI TC Security [4]. This formal specification allows an electronic signature to be automatically verified against the signature policy to which it refers.

2.3 Technical standards

High level requirements for legally valid signatures will be defined by legal frameworks, which are briefly described in the next Section. There still needs a lot of work to be done in the area of standardisation before the exact technical rules, that will fulfill those high level requirements, can be specified. A report prepared by the EESSI (European Electronic Signature Standardisation Initiative) Expert Team [5] has identified several missing functional and quality standards, that are needed for the use of digital signatures. One

of the required standards is a standard for trusted systems and products for digital signatures (signature creation and verification products). In addition, there is a need for standards for interoperability, that will for example define a standard syntax and encoding format for electronic signatures or a protocol to interoperate with a time stamping authority, and for standards for secure management of CAs and other service providers. When standards already exist, minimal security requirements need to be defined, e.g. algorithms and key lengths that are strong enough to resist calculation of the private signature key from the public signature key, or from the signature itself. Minimal security requirements generation and protection of private keys can be based on the FIPS PUB 140-1 (Security Requirements for Cryptographic Modules) standard, or one of existing security evaluation criteria, e.g. ITSEC (Information Technology Security Evaluation Criteria), TCSEC (Trusted Computer System Evaluation Criteria), CTCPEC (Canadian Trusted Computer Product Evaluation Criteria) or a Common Criteria (CC). German regulation on digital signatures, for example, requires ITSEC E4 HIGH level for key generation and private key protection in the smart card.

3 Legal framework

Before the legal framework is discussed, we should first explain the difference between digital signatures and electronic signatures. The term electronic signature, which is widely used in several legislations, generally means any data in electronic form which serves as a method of authentication. Message Authentication Codes (MAC) or electronic pens are examples of electronic signatures methods. With this definition, digital signatures can be regarded as electronic signatures which meet additional requirements, i.e. they are capable of providing data integrity and uniquely identifying the subject that created a signature.

Several countries are in a process of updating their legal frameworks in order to regulate and incorporate recognition of signatures in electronic form. First legal frameworks for electronic or digital signatures were established in the U.S. a few years ago. The state of Utah was the first jurisdiction to enact the digital signature legislation in 1995, and many other states followed thereafter. On an international level, UNCITRAL (United Nations Commission on International Trade Law) adopted its Model Law on Electronic Commerce in 1996. Although it is not digital signature legislation, the Law influenced a number of national and international initiatives. In Europe, several countries, such as Germany with its Digital Signature Law and Digital Signature Ordinance, Italy or Austria, have already passed their digital signature legislation. Detailed summaries and comparisons of other existing and draft laws have been prepared by different institutions, for example by the Interdisciplinary Centre for Law and Information Technology at University of Leuven [7].

To facilitate the use of electronic signatures and to contribute to their legal recognition in EU member countries, European Commission published last year a proposal for a Directive on a community framework for electronic signatures. The Directive, which has been adopted by the European Parliament and the Council of the European Union in November 1999, establishes a legal framework for electronic signatures and certain certification services in order to ensure the proper functioning of the internal market. Although it tries to be technology neutral and covers all forms of electronic signatures, not only digital signatures, public key cryptography and certification authorities are mandated for so-called "qualified electronic signatures" (i.e. digital signatures) with legal equivalence to hand-written signatures. Only certificates called qualified certificates, which meet requirements defined in the Directive, can be used for qualified electronic signatures. The Directive also defines basic requirements for certification service providers issuing these certificates, requirements for secure signature-creation devices, and recommendations for secure signature verification. Certificate profile for qualified certificates is being specified in the IETF draft [18], that was already mentioned in Section 2. An importance of the use of digital signatures in electronic commerce to the EU Commission is also reflected in the V. Framework Programme of the European Community for research, technological development and demonstration (RTD). One of the key actions of the Programme is "New methods of work and Electronic Commerce" with the following prioritised areas: identification and authentication, secure electronic financial transactions and digital object transfer.

What about Slovenia? In Slovenia, we are still in the preparatory phase with a Draft law on electronic commerce and electronic signatures. It is expected that a legislation that is in accordance with the EU Directive will be enacted in a few years. However, it should be noted that the absence of legislation does not preclude parties from using digital signatures in bilateral communication. The parties are free to agree among themselves the terms and conditions under which they accept digitally signed data.

4 Conclusion

Digital signatures are one of the most important security mechanisms and structures in a secured Information society. There exist various types of signatures in electronic form, that can be used to provide authentication, data integrity or non-repudiation. However, it seems that at this time only digital signatures that are based on public-key cryptography can have, under certain technical and legal conditions, equivalent legal effect as hand-written signatures. In this paper, we have briefly described a technical framework required for support of digital signatures, that consists of a set of security mechanisms, technologies and technical standards. An emphasis was given on public-key infrastructures, certification authorities and certificate poli-

cies. We have also presented current activities for the standardisation of a legal framework.

References

[1] Anderson R., Crispo B., Lee J.H., Manifavas C., Matyas V. Jr., Petitcolas F.A.P., The Global Internet Trust Register 1999, MIT Press, Cambridge, MA, 1999.

[2] Baum M.S., Ford W., Secure electronic commerce: building the infrastructure for digital signatures and encryption, Prentice-Hall, 1997.

[3] Chokhani S., Ford W., Internet X.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework, RFC 2527, 1999.

[4] European Telecommunications Standards Institute (ETSI), Electronic Signature Format, draft ETSI ES 201 733 v1.1.4, 1999.

[5] European Electronic Signature Standardization Initiative (EESSI), Final Draft of the EESSI Expert Team Report, 1999.

[6] Housley R., Ford W., Polk W., Solo D., Internet X.509 Public Key Infrastructure Certificate and CRL Profile, RFC 2459, 1999.

[7] Interdisciplinary Centre for Law and Information Technology, K.U.University, The legal aspects of digital signatures, http://www.law.kuleuven.ac.be, 1998.

[8] ICE-CAR (Interworking Public-Key Certification Infrastructure for Commerce, Administration and Research) project, http://ice-car.darmstadt.gmd.de, 1999.

[9] ISO/IEC, Draft Amendment DAM 1 to ISO/IEC 9594-8 on Certificate Extensions, 1997.

[10] ISO/IEC, Final Proposed Draft Amendment on Certificate Extensions, 1999.

[11] ITU-T Recommendation X.509 (1993) | ISO/IEC 9594-8:1994, Information technology - Open Systems Interconnection - The Directory: Authentication framework.

[12] Klobučar T., Jerman-Blažič B., A formalisation and evaluation of certificate policies, Computer Communications 22 (1999), 12, pp. 1104-1110.

[13] Klobučar T., Jerman-Blažič B., SI-CA: agencija za certificiranje javnih ključev v Sloveniji, Varnost in zaščita v telekomunikacijskih omrežjih, Peta delavnica o telekomunikacijah, Brdo pri Kranju, 1997, Marko Jagodic, ur., Sašo Tomažič, ur., Ljubljana, Elektrotehniška zveza Slovenije, 1997 (In Slovene).

[14] National Institute of Standards and Technology (NIST), Digital Signature Standard, Federal Information Processing Standards Publication 186-1, 1998.

[15] National Institute of Standards and Technology (NIST), Secure Hash Standard (SHS), Federal Information Processing Standards Publication 180-1, 1995.

[16] Rivest R., The MD5 Message Digest Algorithm, RFC 1321, MIT in RSA Data Security, Inc., 1992.

[17] Rivest R., Shamir A., Adleman L., A method for obtaining digital signatures and public key cryptosystems, Communications of the ACM, Vol. 21, No. 2, pp. 120-126, 1978.

[18] Santesson S., Polk T., Gloeckner P., Internet X.509 Public Key Infrastructure Qualified Certificates, Internet Draft, 1999.

Using An Electronic Book In Distance Education

Eva Jereb, Branislav Šmitek
University of Maribor
Faculty of organisational sciences
Kidričeva 55a, 4000 Kranj, Slovenia
E-mail: eva.jereb@fov.uni-mb.si

Keywords: distance education, electronic book, multimedia in education

Edited by: Cene Bavec and Matjaz Gams
Received: October 9, 1999 **Revised:** December 2, 1999 **Accepted:** December 19, 1999

In scope of project "Phare Multi - Country Programme for Distance Education" in years 1997/98 and 1998/99 we formed more learning material for the need of our new Distance learning centre. The reason for forming this kind of learning material was that we became aware that new forms and methods of distance learning have to assure the bigger independence of students in the whole learning process. For achieving these aims smaller groups of students with high individualisation are appropriate what brings increased level of activities in the whole learning process and enables periodic information and more realistic knowledge control. In this article we present some experience we had by by forming electronic book for subject Office automation and some students opinion about studying with help of electronic book.

1 Introduction

Distance education is a very suitable form of an education for a modern man. It offers the possibility of studying at home or in virtual classroom. There is no time pressure, man can study at the time which is most appropriate for him. During the process of designing and producing electronic book for distance education purposes we have to consider several different aspects, like:

- distance education is a unique learning process with its own rules within the frame of general didactic rules (Keegan 1991, Van der Brande 1993);
- modern technology (TV, computers and communication equipment) make distance education possible (Laurillard, 1993);
- distance education demands more initiative and participation by students in the educational process; this kind of study offers a higher level of student activity (Boud, 1992); feedback information and more control of knowledge [Rowntree, 1992);
- a learning module can be offered to students with suitable material for self-education (books, programmed material, audio and video cassettes, diskettes with computer supported learning program, etc.) (Laurillard, 1993, Rowntree, 1991);
- many students already have audio and video recorders and computers, which provides the basic technology for distance education (Batagelj & Rajkovic, Dhanarajam 1994, Jereb 1992);
- with adequate telecommunication equipment, students can work on a central computer from home or on a terminal in a classroom;

2 Forming of an electronic book

In continuation of the article we will describe basic steps of designing of an electronic book. We have to stress that the described methodology represents only a methodological framework of electronic book design, which helps the author in developing process.

2.1 Analysis of study plan

The data, contained in the study plan for a subject, present the basic starting-point for the organising of study process. The teacher must include in the educational process also the results of his research work. The goals, nature and structure of the study material describe the forms and methods of the study process and also direct and indirect study sources.

2.2 Study plan articulation

We can divide the whole process of study plan articulation in three basic steps:

- articulation of the study plan into study units
- didactical articulation of single study units
- time articulation of single study units

In accordance with previous by mentioned basic steps the appropriate study plan articulation - also in distance education and designing of an electronic book - has to answer these questions:

- how many planned hours can we realise with the electronic book help and how many with other appropriate electronic media or knowledge sources help?
- what kind of media (hardware and software) do students need for studying with electronic book?
- what kind of instruments for self evaluation will be available?

From theoretical point of view the documented study plan articulation gives us all data for the preparation of all needed study sources for direct and indirect realisation of the study subjects.

2.3 Preparing of electronic book

2.3.1 Text editing

In the writing and editing phase it is suitable to take in consideration some results of the previous research.

- the content has to meet the logics of study material;
- text has to be clear and understandable and index of used words is very welcome;
- the size of electronic book, which includes also other forms of information (sound, picture, animation, video), is smaller than the size of classical textbook; if there is a need we can direct students to secondary knowledge sources in WWW environment;
- choosing of appropriate font size and type assist to clearness of study material, it also stresses different ideas, conceptions and rules;
- in web environment we have to use all advantages of HTML;
- the content of logical rounded parts must contain also questions for self assessment and exercises which direct students to solve the essential problems and form their mentality;
- the content has to be organised for computer usage that it can give the material in any kind of time tempo and sequence.

2.3.2 Sound and graphic editing

In the process of sound and graphic editing we have to consider these empirical findings:

- pictures, graphs, tables, spreadsheets and diagrams in electronic book play an important instrument for getting students attention;
- using of didactical formed cartoons is more effective than using of very realistic pictures;
- the pie charts are more suitable for representation of percent values than histograms, and histograms are more suitable than linear graphs;
- the clear and readable legend near any graph is very important;
- in the examples where for transfer of knowledge the audio representation is enough we include the sound files in the electronic book; if we need the audiovisual representation we use sound and animation;
- when we use different kinds of media we have to take in consideration the availability of different hardware and software and the qualification of the student to use them;
- beside recording equipment, hardware and software we need also a lot of knowledge and experiences to produce the multimedia electronic book;
- there is a luck of professional digital audio and video study products on our market;

2.3.3 Instruments for self assessment

The electronic book should contain also suitable tests for self assessment at the beginning, in the middle and at the end of the studying process. As for how the exercises are constructed and how students answer questions we can divide exercises in two groups: exercises where student write down adequate answer and exercises where student choose the correct answer. We can write down several general findings according to preparing self assessment questionnaires:

- exercises where students write down answers are very seldom used in electronic books because of the problem how to tell computer which answer is correct;
- short write down answers are more applicable but we still have problems of quite a big number of the correct answers (capitalisation, abbreviation, etc);
- most suitable type of questions are multi-choice questions;
- the questions have to be in order from easiest to most difficult ones;
- people involved in preparing self assessment questionnaires have to know a lot about the theory of assessment;

3 Advantages of an electronic book

The structure of an electronic book is very similar to a classical textbook. The contents in electronic book is also divided in chapters and subchapters. Subchapters discuss small rounded subject (material) so we can say that subchapters present study modules. The advantages of electronic books are:

- electronic book can be built up step by step (module by module);
- content can be actualised from time to time;
- with the appropriate form we can achieve the structure of programmed book (information, questions and exercise, feed-back information);

- with the appropriate form we can combine levels of the study process (introducing, working with new contents, repetition, exercise, verifying);
- there is only essential information on the screen and the user can use it when he needs it (individualisation of tempo, path and mode of study).

4 The electronic book

The electronic book and knowledge tests were designed with help of GUIDE 3.1 hypermedia software The first version, which was designed for individual study on PC, was also re-formed into HTML form for use on the world wide web.

The experimentally tested electronic book was designed for office technology comprehension and is used by subject Office automation. Working with the electronic book is not difficult. Students can choose their own sequence of learning subjects and simultaneously check the acquired knowledge of each chapter. All questions are also gathered in an extra part in the electronic book so those students can test themselves at the end of the learning process. The only difference between checking the knowledge simultaneously and testing it at the end is that students by checking their knowledge simultaneously get the feedback information about their success. By final testing students get only the number of correct and wrong answers and a percentage score. The difference between an electronic and a classic book is that the electronic book does not only mediate the information but it also tells the student if she or he understood the information. The electronic book tells the student weather hers or his answer was correct or not. Figure 1 shows an example of the electronic book screen used for chapter: Office machines and devices.

5 Student's opinion about using the electronic book

Students' opinions about individual learning with help of the electronic book were gathered with help of adequate opinion scale. The scale was divided into five basic parts representing five views (elements) of studying process:

- motivation
- methodology
- pretentiousness
- presentation and material organisation quality
- learning contest.

Each part includes eight statements so the opinion scale has 40 different statements.

The opinion scale was expressed with 4 categories:

- I fully agree
- I partly agree
- I partly do not agree
- I do not agree at all.

So at the end we had a questionnaire with 40 questions or statements where 20 of them expressed a positive point of view and 20 a negative point of view.

The inquiry was carried out among the students of second and third class of Faculty of Organisational Sciences in year 1994/95, 1995/96 and 1996/97. In these three years data from 133 students were collected. The students' response was satisfactory - we could say above expected. The inquiry was always done at the end of the education process, when lectures, exercises and individual study with help of the electronic book were finished. The inquiry was anonymous.

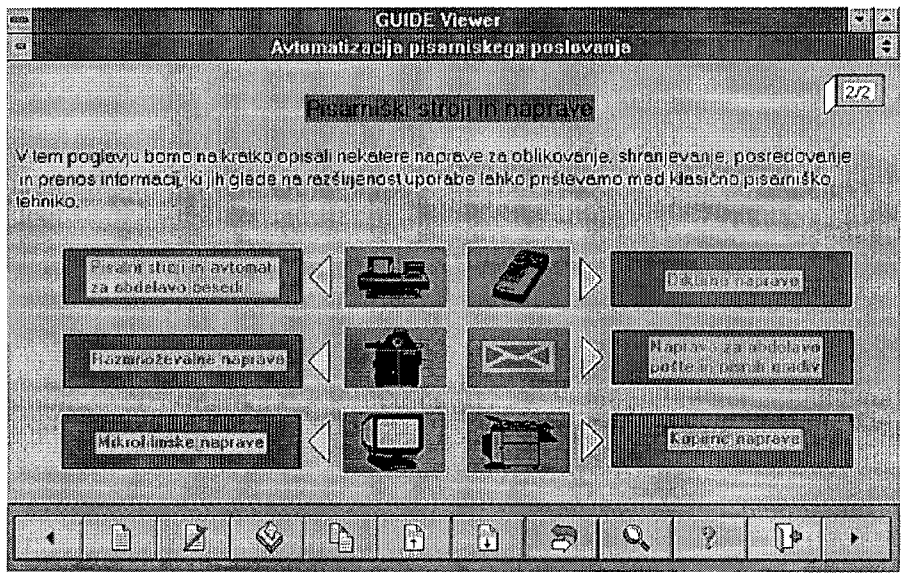


Figure 1: Example of electronic book screen

At the end of our inquiry we reckoned up the arithmetic means of answers for each statement so that we could see weather the students had a positive or a negative relation to the specific statement and so to the specific element of studying process. The results of inquiry are shortly described below:

- Students were satisfied with a new way of study, they meant that the learning quality was grooving by using the electronic book.
- They said that the chapter separation in the electronic book was good and that all parts were well linked.
- Opinions about the learning speed were divided.
- Students' opinion weather concrete examples contribute to better understanding of material or not were also divided.
- Students were satisfied with the contents of the electronic book and they meant that the extent of the learning material is appropriate.
- They showed a positive statement about methodological point of view of the electronic book.
- Students were satisfied with the pretentiousness of the electronic book, with the systematic work, with learning goals and with the acquired knowledge.
- They could not tell weather the learning material mediates only facts or not. These could be the consequence of not knowing the learning material in full. Maybe the answers would have been different if the inquiry had been curried out after the final exam, when the students would have been more acquainted with learning material, and not immediately after the first work with the electronic book.

6 Conclusion

In the article the use of an electronic book as a new method of distance learning is represented. The results of the research, which was carried out among the students of Faculty of Organisational Sciences who were using an electronic book by their study, are shown. Students are satisfied with this kind of study and are looking forward to using electronic books for other subjects as well. The use of an electronic book variegates the study and increases the individual work and motivation of students. The positive experiences of using an electronic book are pointing out that the introducing of distance learning would probably also have a good response among the students.

7 References

[1] Batagelj V., Rajkovič V.: Information Technology Project in Slovenia Schools, Proc. of 1st Euro Education Conference, Aalborg, 9.-15.

[2] Boud D.: The Challenge of Problem Based Learning, Kogan Page, London, 1992.

[3] Collins J.: Computers in Classroom and College, Computer Education, June 1994.

[4] Dhanarajam G. ed.: Economics of Distance Education: Recent Experience, Open Learning Institute, Hongkong, 1994.

[5] Hedberg, J.: Converging Technologies in Education: Interactive Multimedia and Online Learning; The University of Wollongong, New South Wales, Australia, 1996.

[6] Jereb J., Jug J.: Učna sredstva v izobraževanju, Moderna organizacija, Kranj, 1987.

[7] Jereb J.: Računalnik v izobraževanju, Mc&Boss, Kranj, 1991.

[8] Jereb J.: Strokovno izobraževanje in razvoj kadrov, Moderna organizacija, Kranj, 1989.

[9] Jerram P.; Gosney M.: Multimedia Power Tools, Verbum Inc. and Gosney Company, 1995.

[10] Jereb J, Jug J. et.al.: Izobraževanje odraslih, poročilo o raziskovalni nalogi, Fakulteta za organizacijske vede, Kranj, 1992.

[11] Keegan, D.: The Study of Distance Education: Terminology, Definition and Field of Study in Research and Distance Education, Peter Lang, Frankfurt am Main, 1991.

[12] Laurillard, D.: Rethinking University Teaching: A Framework for the Effective Use of Educational Technology, Routledge, London, 1993.

[13] Marentič - Požarnik B.: Prispevek k visokošolski didaktiki, DZS, Ljubljana, 1978.

[14] Rowntree, D.: Teaching through Self Instruction , How to develop Open Learning materials, Kogan Page, London, 1991.

[15] Rowntree, D.: Exploring Opem and Distance Learning, Kogan Page, London, 1992.

[16] Rowntree D.: Preparing Materials for Open, Distance and Flexible Learning, Kogan Page, 1994.

[17] Strmčnik F.: Sodobna šola v luči programiranega pouka, DDU Univerzum, Ljubljana, 1978.

[18] Van der Brande, L.: Flexible and Distance Learning, John Wiley & Sony, Chichester, 1993

[19] Zorman I.: Sestava testov znanja in njihova uporaba v šoli, Zavod za šolstvo, Ljubljana, 1974.

Multi-Attribute Decision Modeling: Industrial Applications of DEX

Marko Bohanec[1], Vladislav Rajkovič [2,1]
[1] Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia
Phone: +386 61 1773 309, Fax: +386 61 1258 058
[2] University of Maribor, Faculty of Organisational Sciences, Kranj, Slovenia
{marko.bohanec, vladislav.rajkovic}@ijs.si

Keywords: decision support, multi-attribute decision making, qualitative decision models

Edited by: Cene Bavec and Matjaz Gams
Received: October 2, 1999 **Revised:** November 20, 1999 **Accepted:** December 12, 1999

DEX is an expert system shell for qualitative multi-attribute decision modeling and support. During the last decade, it has been applied over fifty times in complex real-world decision problems. In this article we advocate for the applicability and great potential of this approach for industrial decision-making. The approach is illustrated by a typical industrial application in land use planning, and supplemented by an overview of some other completed industrial applications. The learned lessons indicate the suitability of the qualitative DEX methodology particularly for "soft", i.e., less structured and less formalized, decision problems. Practical experience also indicates the importance of methods that facilitate the analysis, simulation, and explanation of decisions.

1 Introduction

In complex decision-making processes, it is often necessary to deal with the problem of choice (Simon, 1977). Given a set of *options* (or alternatives), which typically represent some objects or actions, the goal is

(1) to *choose* an option that best satisfies the aims or goals of decision maker, or
(2) to *rank* the options from the best to the worst one.

One of the approaches to such problems, which is well known and commonly employed within Decision Support Systems (Andriole, 1989), is based on *evaluation models* (Figure 1). The idea is to develop a model that evaluates options giving an estimate of their worthiness (*utility*) for the decision-maker. Based on this estimate, the options are ranked and/or the best one is identified. Usually, a decision model is designed in an interaction between the decision maker and decision analyst.

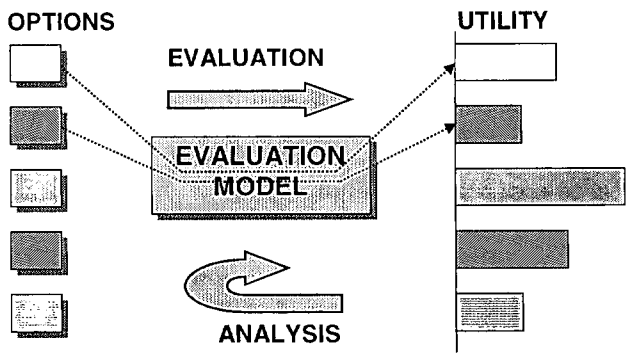


Figure 1: Evaluation-based decision modeling

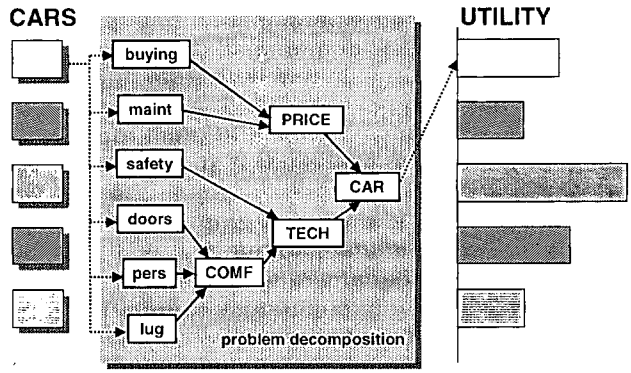


Figure 2: Multi-attribute decision modeling

An important feature of evaluation models is that they can be, in addition to the sole *evaluation* of options, used for various *analyses* and *simulations*, which may contribute to a better justification and explanation of decisions. For example, a *what-if analysis* can provide a better insight into a causal relation between problem parameters and outcomes. Another example is a *sensitivity analysis* that can assess the sensitivity of model with respect to small changes of options.

An evaluation model can be developed in many ways. The approach that prevails in decision practice is based on *multi-attribute decomposition* (Chankong and Haimes, 1983; Saaty, 1993; Buede and Maxwell, 1995): we take a complex decision problem and decompose it into smaller and less complex subproblems. The result of such development is a *decision model* that consists of *attributes*, each of which represents a decision

subproblem. Attributes are organized hierarchically and connected by *utility functions* that evaluate them with respect to their immediate descendants in the hierarchy. Figure 2 illustrates this basic principle of multi-attribute modeling by showing a simple hierarchy of attributes for the evaluation of cars.

Real-life applications of multi-attribute methods, which were conducted at Jožef Stefan Institute in Ljubljana, were all based on DEX (Bohanec and Rajkovič, 1990). This is an expert system shell for multi-attribute decision making that combines the "traditional" multi-attribute decision making with some elements of Expert Systems and Machine Learning. The distinguishing characteristic of DEX is its capability to deal with *qualitative* models. Instead of numerical variables, which typically constitute traditional *quantitative* models, DEX uses qualitative variables; their values are usually represented by words rather than numbers, for example "low", "appropriate", "unacceptable", etc. Furthermore, to represent and evaluate utility functions, DEX uses *if-then decision rules*. In contrast, this is traditionally carried out in a numerical way, using weights or similar indicators of attributes' importance.

An important additional feature of DEX is its capability to deal with inaccurate, uncertain or even missing data about options. In such cases, DEX represents options by distributions of qualitative values, and evaluates them by methods based on probabilistic and/or fuzzy propagation of uncertainty.

During the last decade, DEX was used in more than fifty real-life decision problems. The aim of this article is to advocate for the wide applicability of DEX to complex decision problems that occur in industry. In the next section, we first illustrate the approach by a typical industrial application in land use planning. This is followed by an overview of several other completed industrial applications in performance evaluation of companies, evaluation of products, projects and investments, ecology, and loan allocation. Finally, we summarize the lessons learned in these applications, and propose some future directions for the development of underlying methodology.

2 A Real-World Case

One of the most typical applications of DEX occurred with Goriške opekarne, a company located near the Slovenian city of Nova Gorica. The company is engaged in a very traditional business: production of bricks and tiles. Decades ago, they had built a factory near a suitable clay pit that was then providing raw material for their production. Until 1993, however, the clay pit has become almost completely exhausted, so the company was faced with a critical strategic decision of how to survive and continue with this type of production. Their only option was to find a new appropriate clay-pit location.

An exploratory study revealed three possible candidate locations. Unfortunatelly, none of them was really appropriate as numerous difficult problems were foreseen, ranging from technological, transportational and financial to environmental and socio-psychological. The latter two problems seemed particularly important as the project was inevitably going to affect the environment, leading to a possible rejection of local inhabitants. For these reasons, a group of experts was formed to thorougly analyze the problem and propose alternative solutions (Bohanec, et al., 1993).

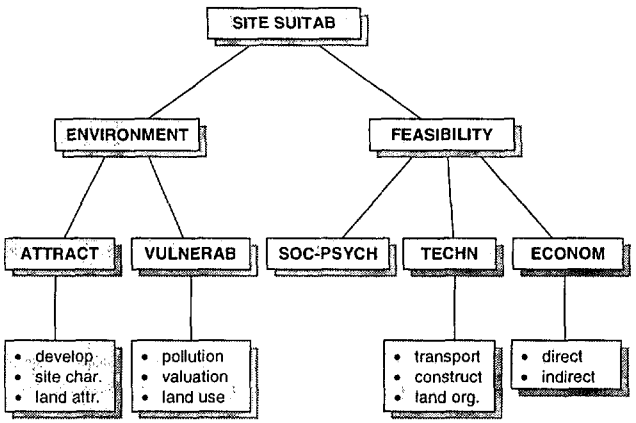


Figure 3: Topmost levels of clay-pit evaluation model

In the first stage, the experts developed the structure of multi-attribute model for the evaluation of clay-pit locations.. Two primary evaluation dimensions were taken into account: Environmental impact and Feasibility of the project. For each of these, the most relevant attributes were identified and organized into a hierarchical structure (Figure 3). Note that only topmost levels of the model are shown in the figure. In total, the model contained 49 attributes: 29 basic (terminal nodes) and 20 aggregate (internal nodes).

Table 1: Decision rules for Site suitability

	ENVIRONMENT	FEASIBILITY	**SITE**
1	*	unacc	unacc
2	unacc	*	unacc
3	less-acc	less-acc	marg-acc
4	≥ acc	less-acc	less-acc
5	less-acc	acc	less-acc
6	acc	acc	acc
7	good	acc	good

The second stage involved the definition of decision rules. Basically, these are simple *if-then* rules that for each of the 20 internal nodes in the model determine its evaluation with respect to its lower-level descendants in the hierarchy. Usually, they are represented in a tabular form. For example, Table 1 shows decision rules that were defined by the experts for the topmost node Site suitability. In the table, an asterisk '*' represents any value, and '≥' means 'better or equal'.

In the third stage of the decision-making process, the options are identified and described by the values of basic attributes. In our case, there were three clay-pit locations, each of which was represented by 29 data items that corresponded to basic attributes of the model. Furthermore, as some of these items, such as Social-psychological feasibility, were inherently inaccurate or difficult to obtain, several variations of the descriptions were formed, anticipating either an "optimistic" or "pessimistic" development of the project. Effectively, this increased the number of considered options to eight (Figure 4) and provided a foundation for subsequent what-if analysis.

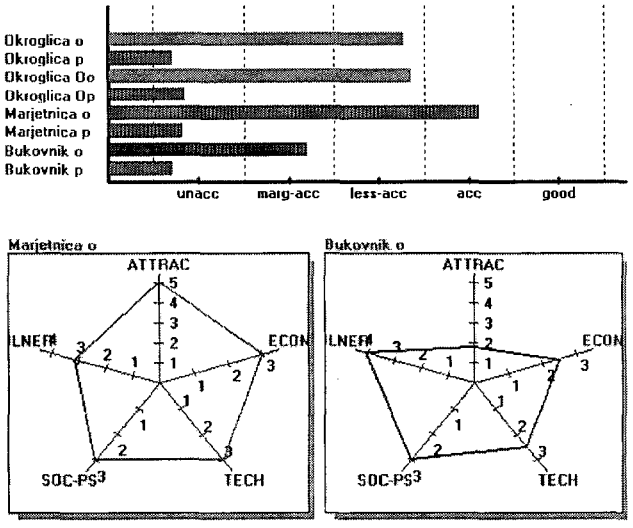


Figure 4: Visualization of clay-pit evaluation results

In the last stage, the model was utilized to evaluate the clay-pit locations. As shown in Figure 4, the best location was the one called Marjetnica, which was evaluated as "acceptable", but only in its "optimistic" instance. On the other hand, all the "pessimistic" instances were unacceptable, indicating the great sensitivity of decision. Therefore, thorough what-if and sensitivity analyses were performed for each location. The most important result was achieved by comparing "optimistic" and "pessimistic" options with respect to basic attributes. The outcome of this comparison was a comprehensive list of possible problems that could occur with each location. On this basis, the expert team not only was able to find the best location, but also to foresee potential pitfalls and suggest how to avoid them.

3 Other Applications

In about ten years time, DEX was used in more than fifty real-life decision problems in various areas. About one half of the problems can be classified as industrial, while the remaining were conducted in the fields such as education or medicine and health care (Bohanec, et al., 1999). Some of the industrial problems were very difficult and involved substantial financial and other risks for decision-making organizations. In what follows we briefly outline five representative application areas, which clearly indicate the wide applicability of DEX for a variety of decision problems. The description of some other early industrial applications can also be found in (Urbančič, et al., 1991).

3.1 Performance Evaluation of Companies

Here, the general task is that a company or agency develops an evaluation model that assesses the performance of some other companies. The aim is, for example, to find a suitable business partner. The work with DEX in this area began in 1987, where a number of such models were developed in collaboration with the International Center for Public Enterprises (Bohanec and Rajkovič, 1990). An example hierarchy of attributes that was used to assess the performance of 54 public enterprises in Pakistan, is shown in Figure 5. This work culminated in 1989 with the development of models that were used in the privatization of Peruvian public enterprises.

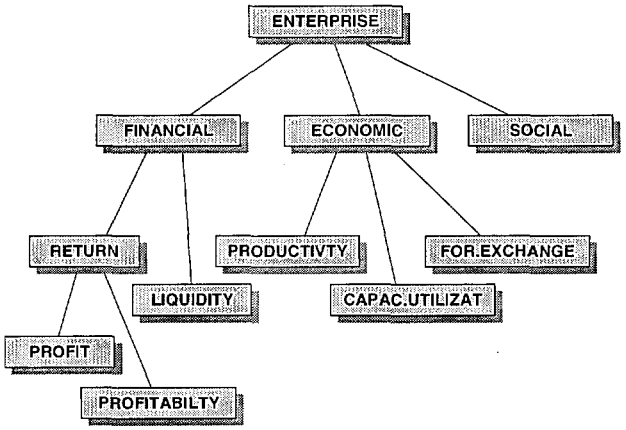


Figure 5: Topmost levels of the model for performance evaluation of public enterprises

3.2 Product portfolio evaluation

The problem is to assess the quality of products made by a company or production unit. This assessment is vital for the formation of strategies. The approach with DEX was based on the so-called portfolio method (Krisper, et al., 1991), which evaluates products using two primary evaluation dimensions: market attractiveness and competitive ability. Several practical cases were analyzed in this way, including the products of some well-known Slovenian companies Fructal, Radenska, SRC, and DZS.

3.3 Evaluation of projects and investments

The evaluation of projects or investment strategies is an industrial application context in which DEX has got the largest number of applications. The most typical investments included various software, hardware and technology, such as data base management systems, production control software, meteorological radar equipment, or a production line. The decision problems were often related to various investment proposals and tenders. An example of such applications, which is

documented quite in detail, is a model for the evaluation of research and development projects (Bohanec, et al., 1995).

3.4 Remediation of dumpsites

This is a recent application in the field of environmental care. In order to alleviate the problem of illegal dumpsites in Slovenia, an expert system was developed that assesses the environmental impact of dumpsites and suggests activities for their remediation (Špendl, 1998). The environmental impact of dumpsites is assessed by a qualitative DEX model (Figure 6), which is embedded in the expert system.

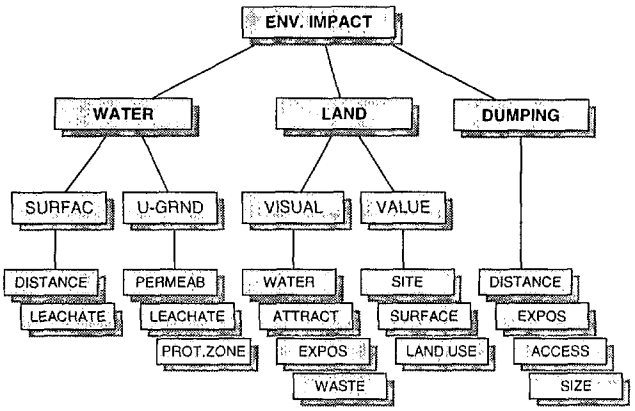


Figure 6: Model for the assessment of dumpsite's environmental impact (topmost levels only)

3.5 Housing loan allocation

This is an example of a repetitive decision-making task being supported by a DEX model. The model is a part of a management decision support system that is used since 1991 by the Housing Fund of the Republic of Slovenia for the allocation of housing loans with favorable terms to citizens (Bohanec, et al., 1996). Until 1999, the Fund has issued 16 floats of loans, i.e., about two per year, and approved almost 20 thousand loans.

The amount requested by applicants in a float typically far exceeds the available funds. Thus, the applicants must be ranked into a priority order. The procedure is required to be fast, reliable, transparent, and fair for all applicants. The request for transparency asks for effective explanations of loan priority order, which have to be provided to both the decision-making committee and a large number (usually, several thousands) of applicants. In the Fund's system, these requirements were fulfilled by a qualitative model that ranks the applications into five priority classes and provides a foundation for various explanations, which are obtained by analyses and simulations of application data and the model itself.

4 Experience

Some important lessons have been learned in the applications of DEX. Here, we present some findings related to the duration of model development processes, difficulty of development stages, and categories of decision problems that seem to be particularly well suited for the application of DEX.

The time needed to develop a DEX model turns out to be extremely problem-dependent: it may take from few hours to several months. Most typically, however, the development requires about two working days for the development of model structure, from one to two days to define decision rules, and from one to several days to collect data about options, to evaluate, and analyze them. Therefore, the process most typically lasts from two to ten working days.

The most difficult stage of the process is its first one, in which the relevant attributes must be identified and appropriately organized into a hierarchical structure. This stage heavily relies on knowledge and experience of decision-makers and experts, and requires a deep understanding of the decision problem. It can still be considered more art than science. The remaining stages have been found much less problematic. Therefore, an appropriate identification of model structure mostly determines the success of the decision-making process.

DEX with its qualitative modeling and ability to handle inaccurate and/or incomplete data about options appears particularly well suited for decision problems that involve qualitative concepts and a great deal of expert judgement. Also, it seems that the usefulness of DEX increases with the increasing difficulty, or "complexity", of the decision problem. So far, the best results were achieved in problems that required large models, consisting of at least 15 attributes, and/or involving a large number of options, i.e., from about 10 to several hundreds of options. On the other hand, DEX turned out to be unsuitable for problems that require exact formal modeling, numerical simulation and/or optimization.

5 Further Work

Currently, there are three limitations of the DEX approach that, we believe, can be greatly improved by appropriate extensions of the methodology. First, the difficult stage of model structure development could be additionally supported by a machine learning method that would develop (or at least suggest) model structure using decision examples taken either from an existing database of past decisions, or provided explicitly by the decision-maker. A considerable progress in this direction has already been made by the development of a learning method called HINT (Zupan, et al., 1999). Given training examples, HINT develops a hierarchical multi-attribute evaluation model that explains and possibly generalizes the examples. The structure of the models developed by HINT is essentially the same as the structure of models developed "manually" using DEX. The HINT's model development is based on function decomposition, an approach that was originally developed for the design of digital circuits.

Another limitation of DEX is that it is strictly limited to qualitative decision models; it cannot use numerical variables nor analytically represented utility functions that are commonly used in traditional quantitative models. This is sometimes advantageous in comparison with other decision modeling systems, which exclusively rely on quantitative models. However, many real-life decision problems require both qualitative and quantitative attributes, so the integration of these two may have a great practical impact: it may increase the flexibility of the method and extend the range of decision problems that can be successfully approached. Methodologically, such integration appears quite difficult and requires more research. In the context of DEX, we consider it a long-term goal.

Last but not least, the major part of DEX software has been developed about ten years ago and currently appears quite outdated. Therefore, an overall redesign and renewal of software is planned for the near future. Currently, we are developing a program called DEXi, an educational subset of DEX to be used by students and teachers in secondary schools and faculties. We plan to follow this by the development of a functionally complete state-of-the-art DEX system.

6 Conclusion

The DEX system effectively integrates two methodologies: multi-attribute decision making and expert systems. To a limited extent, it also includes some elements of machine learning and fuzzy logic. By this, it facilitates a structured and systematic approach to complex decision problems. So far, DEX has been successfully used in over fifty real-life decision problems in industry, medicine, health care and education, which all speak in favor for its wide applicability and flexibility. From the practical viewpoint, the most important characteristics of DEX are:

1. Qualitative (symbolic) decision modeling, which is particularly well suited for "soft" decision problems, i.e., less structured and less formalized problems, which involve a great deal of expert judgement.
2. Focus on the explanation and analysis of options, which lead to better-understood and justified decisions.
3. Active support of the decision-maker in the acquisition of decision rules, which speeds up model development and reduces the number of errors.

The goals of further research and development related to DEX are twofold. First, we wish to improve the support in the difficult stage of model structure development, and propose to use machine learning methods, such as HINT, for that purpose. To further improve the flexibility and general applicability of the approach, we suggest further research towards an integration of qualitative and quantitative decision models.

7 References

[1] S.J. Andriole: Handbook of Decision Support Systems. TAB Books, 1989.

[2] M. Bohanec, V. Rajkovič, V.: DEX: An expert system shell for decision support, *Sistemica* 1, 145–157, 1990.

[3] M. Bohanec, B. Kontić, D. Kos, J. Marušič, S., Polič, J. Rakovec, B. Sedej, et al.: Comparison of clay-pit locations Okroglica, Bukovnik, Marjetnica with respect to environmental protection (in Slovenian). Ljubljana: Jožef Stefan Institute, Report DP-6742, 1993.

[4] M. Bohanec, V. Rajkovič, B. Semolič, A. Pogačnik: Knowledge-based portfolio analysis for project evaluation. *Information & Management* 28, 293–302, 1995.

[5] M. Bohanec, B. Cestnik, V. Rajkovič: A management decision support system for allocating housing loans. *Implementing Systems for Supporting Management Decision* (eds. P. Humphreys, L. Bannon, A. McCosh, Migliarese, J.-C. Pomerol). Chapman & Hall, 1996.

[6] M. Bohanec, B. Zupan, V. Rajkovič: Hierarchical multi-attribute decision models and their application in health care. *Proc. Medical Informatics Europe 99* (eds. P. Kokol, B. Zupan, J. Stare, M. Premik, R. Engelbrecht), Amsterdam: IOS Press, 670–675, 1999.

[7] D.M. Buede, D.T. Maxwell: Rank disagreement: A comparison of multi-criteria methodologies. *Journal of Multi-Criteria Decision Analysis* 4, 1–21, 1995.

[8] V. Chankong, Y.Y. Haimes: *Multiobjective Decision Making: Theory and Methodology*. North-Holland, 1983.

[9] M. Krisper, V. Bukvič, V. Rajkovič, T. Sagadin: Strategic planning with expert system based portfolio analysis. *EXPERSYS-91: Expert system applications* (eds. J. Hasemi, J.G. Gouardères, J.P. Marciano). IITT International, 1991.

[10] T.L. Saaty: *Multicriteria Decision Making: The Analytic Hierarchy Process*. RWS Publications, 1993.

[11] A.H. Simon: *The New Science of Management Decision*. Prentice-Hall, 1977.

[12] R. Špendl: *An expert system for the evaluation of environmental impact and remediation of illegal dumpsites* (in Slovenian). M.Sc. Thesis. University of Ljubljana, Faculty of Information and Computer Science, 1998.

[13] T. Urbančič, I. Kononenko, V. Križman: *Review of Applications by Ljubljana Artificial Intelligence Laboratories*. Ljubljana: Jožef Stefan Institute, Report DP-6218, 1991.

[14] B. Zupan, M. Bohanec, J. Demšar, I. Bratko, I.: Learning by discovering concept hierarchies. *Artificial Intelligence* 109, 211–242, 1999.

Perception-Based Classification

Mihael Ankerst, Christian Elsen, Martin Ester, Hans-Peter Kriegel
Institute for Computer Science, University of Munich
Oettingenstr. 67, D-80538 München, Germany
{ankerst | ester | kriegel}@dbs.informatik.uni-muenchen.de, c.elsen@elsen.net

Keywords: classification, decision tree, data mining, visualization

Edited by: Cene Bavec and Matjaz Gams
Received: October 2, 1999 **Revised:** December 2, 1999 **Accepted:** December 19, 1999

Classification is an important problem in the emerging field of data mining. Given a training database of records, each tagged with a class label, the goal of classification is to build a concise model that can be used to predict the class label of future, unlabeled records. A very popular class of classifiers are decision trees because they satisfy the basic requirements of accuracy and understandability. Instead of constructing the decision tree by a sophisticated algorithm, we introduce a fully interactive method based on a multidimensional visualization technique and appropriate interaction capabilities. Thus, domain knowledge of an expert can be profitably included in the tree construction phase. Furthermore, after the interactive construction of a decision tree, the user has a much deeper understanding of the data than just knowing the decision tree generated by an arbitrary algorithm. The interactive approach also overcomes the limitation of most decision trees which are fixed to binary splits for numeric attributes and which do not allow to backtrack in the tree construction phase. Our performance evaluation with several well-known datasets demonstrates that even users with no a priori knowledge of the data construct a decision tree with an accuracy similar to the tree generated by state of the art algorithms. Additionally, visual interactive classification significantly reduces the tree size and improves the understandibility of the resulting decision tree.

1 Introduction

The success of computerized data management has resulted in the accumulation of huge amounts of data in several organizations. There is a growing perception that analyses of these large databases can turn this "passive" data into useful information. The term *Data Mining* refers to the discovery of non-trivial, previously unknown, and potentially useful patterns embedded in databases.

Classification is one of the major tasks of data mining. The goal of *classification* is to assign a new object to a class from a given set of classes based on the attribute values of this object. Different methods [12] have been proposed for the task of classification, for instance *decision tree classifiers* which have become very popular. Decision tree classifiers are primarily aimed at attributes with a *categorical* domain, that is a small set of discrete values. *Numeric* attributes, however, play a dominant role in application domains such as astronomy, earth sciences and molecular biology where the attribute values are obtained by automatic equipment such as radio telescopes, earth observation satellites and X-ray cristallographs. [6] discusses an approach that splits numeric attributes into multiple intervals rather than just two intervals. The well-known algorithms, however, perform a binary split of the form for a numeric attribute a and a real number v. The SPRINT decision tree classifier [3] processes numeric attributes as follows. There are $n - 1$ possible splits for n distinct values of a. The gini index is calculated at each of these n - 1 points and the attribute value yielding the minimum gini index is chosen as the split point. CLOUDS [4] draws a sample from the set of all attribute values and evaluates the gini index only for this sample thus improving the efficiency.

A commercial system for interactive decision tree construction is SPSS CHAID [15] which - in contrast to our approach - does not visualize the training data but only the decision tree. Furthermore, the interaction happens only before the tree construction yielding user defined values for global parameters such as maximum tree depth or minimum support for a node of the decision tree.

Visual representation of data as a basis for the human-computer interface has evolved rapidly in recent years. [8] gives a comprehensive overview over existing visualization techniques for large amounts of multidimensional data. Recently, several techniques of visual data mining have been introduced. [5] presents the technique of *Independence Diagrams* for visualizing dependencies between two attributes. The brightness of a cell in the two-dimensional grid is set proportional to the density of corresponding data objects. This is one of the few techniques which does not visualize the discovered knowledge but the underlying data. However, the proposed technique is limited to two attributes. [10] presents a decision table classifier and a mechanism to

visualize the resulting *decision tables*. It is argued that the visualization is appropriate for business users not familiar with machine learning concepts. In contrast to well-known decision tree classifiers, our novel interactive approach enables arbitrary split points for numeric attributes, the use of domain knowledge in the tree construction phase and backtracking.

In this paper, we introduce a novel interactive decision tree classifier based on a multidimensional visualization of the training data. Our approach allows to integrate the domain knowledge of an expert in the tree construction phase and it overcomes the limitation of binary splits for numeric attributes. The rest of this paper is organized as follows. In section 2 we introduce our technique for visualizing the training data. The support for interactively constructing a decision tree - which we have implemented in the Perception-Based Classification (PBC) system - is discussed in section 3. Section 4 reports the results of an extensive experimental evaluation on several well-known datasets. Section 5 summarizes this paper and outlines several issues for future research.

2 Visualizing the training data

In our approach, we visualize the training data in order to support interactive decision tree construction. We introduce a novel method for visualizing multi-dimensional data with a class label such that their degree of impurity with respect to class membership can be easily perceived by a user. Our pixel-oriented method maps the classes to colors in an appropriate way. The basic idea of pixel-oriented visualization techniques [8] is to map each attribute value v_i of each data object to one colored pixel and to represent the values belonging to different attributes in separate subwindows. The proposed techniques [9] differ in the arrangement of pixels within a subwindow.

Circle Segments [2] is a recent pixel-oriented technique which was introduced for a more intuitive visualization of high-dimensional data. The Circle Segments technique maps d-dimensional objects to a circle which is partitioned into d segments representing one attribute each. Figure 1 illustrates the partitioning of the circle as well as the arrangement. Within each segment, the arrangement starts in the middle of the circle and continues to the outer border of the corresponding segment in a line-by-line fashion. These lines upon which the pixels are arranged are orthogonal to the segment halving lines. An extension of this technique has been applied in the context of cluster analysis [1].

While most approaches of visual data mining visualize the discovered knowledge, our approach is to visualize the training data in order to support interactive decision tree construction.

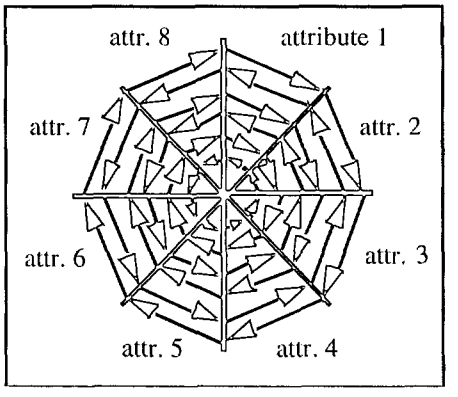


Figure 1.Illustration of the Circle Segments technique for 8-dimensional data objects

We introduce a novel method for visualizing multi-dimensional data with a class label such that their degree of impurity with respect to class membership can be easily perceived by a user. Our method performs pixel-oriented visualization and maps the classes to colors in an appropriate way.

Let D be a set of data objects consisting of d attributes $A_1, \ldots, A_d$ and having a unique *class label* from the set of *Classes* $= \{c_1, c_2, \ldots c_k\}$. For each attribute A_i, let a total order $\leq$ be defined, for example the *$\leq$-order* for numeric attributes or the lexicographic order for string attributes.

To map each attribute value of D to a unique pixel, we follow the idea of the Circle Segments technique, i.e. we represent all values of one attribute in a segment of a circle with the proposed arrangement inside a segment. We do not use, however, the overall distance from a query to determine the pixel position of an attribute value. Instead, we sort each attribute separately and use the induced order for the arrangement in the corresponding circle segment. The color of a pixel is determined by the class label of the object to which the attribute value belongs. In the following, we introduce our technique for mapping classes to colors.

Let *Colors* be the set of all different colors which can be represented in a given color model such as the RGB model, denoted as *Colors* $= \{col_1, col_2, \ldots col_m\}$, $m \leq k$. We are looking for an injective function *visualize*: *Classes* $\rightarrow$ *Colors* which, roughly speaking, should map "similar" classes to "similar" colors and "dissimilar" classes to "dissimilar" colors. We use distance functions to define a formal notion of similarity for both the classes and the colors: the smaller the distance, the larger the similarity and vice versa. For example, when we have no additional information about the semantics of the classes c_i, we use the following distance function for classes:

$$dist_{categorical}(c_i, c_j) = \begin{cases} 0 & if\ i=j \\ 1 & else \end{cases}$$

There are, however, many cases where we know more about the semantics of the classes. For example, there may be a class hierarchy defined by a predecessor

function *pred* for each class. Then we may use the distance function $dist_h$ defined as follows:

$$dist_h(c_i, c_j) = \begin{cases} 0 \text{ if } i=j \\ 1+dist_h(pred(c_i), pred(c_j)) \text{ else} \end{cases}$$

The indices of the classes c_i are chosen such that classes with a low distance receive neighboring indices implying a total order of *Classes* (which may not be uniquely defined). We define the total class distance $dist_{\text{total-class}}$ as follows:

$$dist_{total-class} = \sum_{i=1}^{k-1} dist(c_i, c_{i+1})$$

To define a distance function and a total order for the *Colors*, we need a suitable color scale with the following properties:

- preservation of the order of the attribute values
  In the color scale, each color col_i should be perceived as "preceeding" any color col_{i+1}.
- uniformity of the perceived distances
  For any pairs (col_i, col_{i+1}) and (col_j, col_{j+1}) the perceived "distance" between col_i and col_{i+1} should be the same as the perceived "distance" between col_j and col_{j+1}.

The function $map: \{1,..,k\} \rightarrow \{1,..,m\}$ maps class *indices* to color *indices* as follows:

$$map(i) = \begin{cases} 1 \text{ if } i=1 \\ map(i-1)+\left\lceil \frac{dist(c_{i-1}, c_i)}{total-class-dist} \times (m-1) \right\rceil \text{else} \end{cases}$$

Note that $(m\text{-}1)$ is the maximum difference between the indices of two elements from *Colors* and $\lceil x \rceil$ denotes the smallest integer i with $i \geq x$. Finally, we define the function $visualize: Classes \rightarrow Colors$ mapping classes to colors as follows:

$visualize(c_i) = col_{map(i)}$

Several color scales satisfying these requirements have been proposed [11]. These color scales are appropriate when a total or partial order is defined for the classes. For the purpose of comparability of the results, the experiments reported in this paper have been performed on several datasets where no semantics about the classes is known. If no order of the classes is given, we do not need the first requirement to preserve the order of the attribute values. Furthermore, the second requirement is weakened such that each pair of colors col_i and col_j is perceived as being different, i.e.

$$dist(col_i, col_j) = \begin{cases} 0 \text{ if } i=j \\ 1 \text{ else} \end{cases}$$

The amount of training data that can be visualized at one time is approximately determined by the product of the number of attributes and the number of data objects. For example, 2.000 data objects with 50 attributes can be represented in a 374x374 window and 10.000 objects with 20 attributes fit into a 516x516 window.

We have developed a color scale for class labels based on the *HSI color model* [7] , a variation of the HSV model. The HSI model represents each color by a triple (hue, saturation, intensity). In our experiments, we observed the most distinctly perceived colors for the following parameter settings: For *col 1* we set hue = 2.5 and intensity = saturation = 1.0, for *col m* we set hue = 0.5 and intensity = saturation = 1.0, and all other colors were obtained by partitioning the hue scale into equidistant intervals.

Our approach of visualizing the training data also considers attributes having a low number of distinct values. In that case, there are many objects sharing the same attribute value and their relative order is not uniquely defined. Depending on the chosen order, we might create homogeneous (with respect to the class label) areas within the same attribute value. To avoid the creation of artificial homogeneous areas, we use the technique of *shuffling*: for a set of data objects sharing the same attribute value the required order for the arrangement is determined randomly, i.e. their class labels are distributed randomly.

3 Perception-based classification

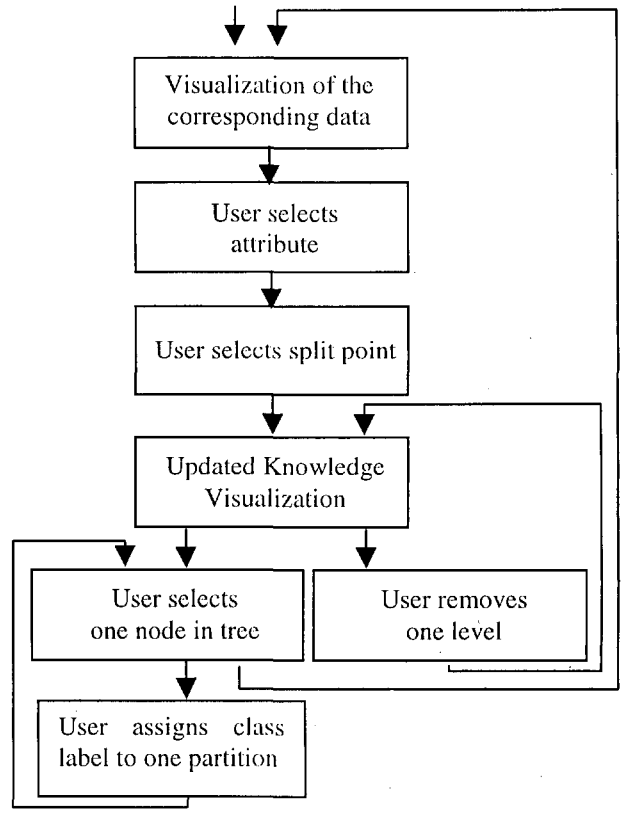


Figure 2. A model for interactive classification

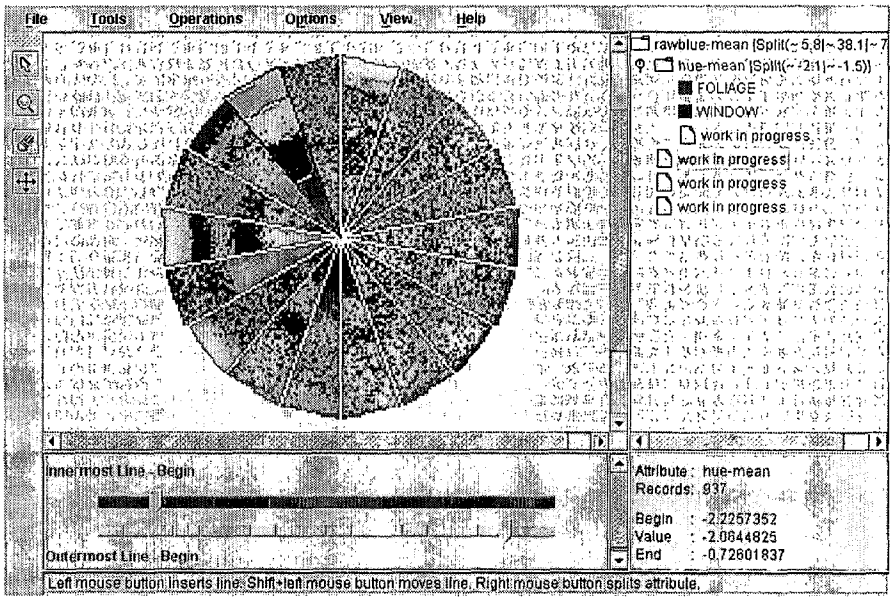


Figure 3. A Screen Shot of the PBC system

The described visualization of the data is the basis of our approach of interactive classification. Figure 2 depicts our model for interactive decision tree construction.

Initially, the complete training set is visualized in the *Data Interaction Window* together with an empty decision tree in the *Knowledge Interaction Window*. The user selects a splitting attribute and an arbitrary number of split points. Then the current decision tree in the Knowledge Interaction Window is expanded. If the user does not want to remove a level of the decision tree, he selects a node of the decision tree. Either he assigns a class label to this node (which yields a leaf node) or he requests the visualization of the training data corresponding to this node. As depicted in figure 3, the latter case leads to a new visualization of every attribute except the ones used for splitting criteria on the same path from the root. Thus the user returns to the start of the interaction loop. The interaction is finished when a class has been assigned to each leaf of the decision tree.

Interactive Selection of Split Points

The interactive selection of split points consists of two steps: (1) selecting splitlines and (2) selecting a split point on each of the selected splitlines.
First, by clicking on any pixel in the chosen segment, the user selects a *splitline* which is one of the lines (orthogonal to the segment halving line) upon which the pixels are arranged. Then by the system this splitline is replaced with an animated line on which alternatively black and white strips move along. Since the colors black and white are not used for the mapping of the classes, the brushed splitline is well perceptible. In a separate area, the pixels of the selected splitline are redrawn in a magnified fashion which enables the user to set the exact split point. Note that the separation of two different colors is not the only criteria for determining the exact split point. If not all attribute values on the splitline are distinct, the same attribute values may belong to objects of different classes. In this case, setting a split point between two differently colored pixels would not be reasonable. Hence we provide feedback to the user in both the basis data visualization and the separate splitline area, such that the attribute value of the pixel at the position of the mouse pointer appear in a subwindow. Figure 3 illustrates the visualization support for the selection of a splitline and an exact split point.

Splitting strategy

Our interactive approach overcomes the limitations of binary splits in attributes with a continuous domain. This additional flexibility rises the question about an appropriate splitting strategy. In our experiments, we observed the best results in terms of accuracy and tree size if the choice of the splitting attribute is based on the strategy described below. The strategy has four options and the first of them which is applicable in the current visualization should be chosen. We will use the term *partition* for a coherent region of attribute values in the splitting attribute that the user intends to separate by split points.

1) *Best Pure Partitions (BPP)*. First choose the segment with the largest pure partitions. A partition is called *pure* if the user decides to label this partition with the most frequent class. This decision leads to leaf nodes in the decision tree, thus reducing the size of data which is not classified.

2) *Largest Cluster Partitioning (LCP)*. If no pure partition is perceptible, the segment with the largest cluster clearly dominant in one color should be chosen. In contrast to a pure partition, such a cluster will not be labeled by the most frequent class.

3) *Best Complete Partitioning (BCP).* If a choice upon BPP or LCP fails, the segment should be chosen that contains the most pixels that can be divided into partitions where each has one clearly dominant color.

4) *Different Distribution Partitioning (DDP).* If none of the above options applies, choose the segment where different distributions can be best separated through partitioning.

After an attribute is chosen the split points have to be set. If the choice follows BPP or LCP, additional split points in the remaining partition should be set if it leads to a separation of clusters or of different distributions. Thus, more inherent information of the splitting attribute is used for deriving the decision tree. Note that the splitting attribute will not reappear in lower nodes of the same path.

4 Experimental evaluation

In comparison to algorithmic decision tree classifiers, the process of interactive classification reveals additional insights into the data. To illustrate this advantage, in this section we discuss an example of two consecutive steps in the tree construction phase. Furthermore, we compare our classifier with popular algorithmic classifiers in terms of accuracy and tree size.

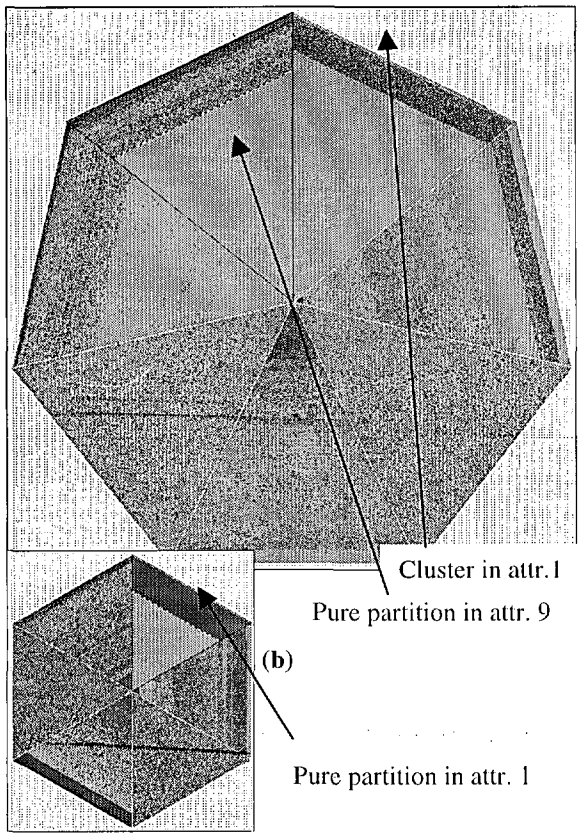


Figure 4. Visualization of the Shuttle data before (a) and after a split (b)

Attributes 1 and 9 are obvious candidates for splitting. According to 'Best Pure Partitions', attribute 9 should be chosen because in contrast to the larger cluster in the segment of attribute 1, the split leads to a pure partition. Note that the non-homogeneity of the cluster in attribute 1 can only be perceived in the color representation. The pure partition can be assigned to the class of its only color. The visualization of the remaining partition has to be examined in a further step. This is shown in figure 4(b) representing the data objects visualized in figure 4(a) except for all objects belonging to the pure partition in attribute 9. Attribute 9 is not visualized any more because it was already used as a splitting attribute on this path of the decision tree. One effect of our visual approach becomes very clear in this example the removal of some training objects from the segment of the splitting attribute may yield the removal of objects from another segment which make a partition of this segment impure. For example, the cluster in attribute 1 (figure 4(a)) becomes a pure partition after the split (figure 4(b)).

We used the accuracy and the tree size (total number of nodes) as quantitative measures to compare PBC with well-known algorithmic approaches. We used the tree size besides accuracy since small trees are easier to understand and we consider understandability of the discovered knowledge to be a major goal. For the comparison, we used three datasets from the Statlog database [13] for which the accuracy and the tree size of many algorithms is known [4] . The Satimage, Segment and Shuttle datasets were chosen because all of their attributes are numeric. We performed the experiments as suggested in the dataset descriptions. As comparison partners we chose the popular decision tree classifiers CART and C4 from the IND package [14] as well as the recently proposed SPRINT [3] and CLOUDS [4] classifiers. The results of CLOUDS were produced with the SSE/DM method.

Accuracy	CART	C4	SPRINT	CLOUDS	**PBC**
Satimage	85.3	85.2	86.3	85.9	83.5
Segment	94.9	95.9	94.6	94.7	94.8
Shuttle	99.9	99.9	99.9	99.9	99.9

Tree Size	CART	C4	SPRINT	CLOUDS	**PBC**
Satimage	90	563	159	135	60
Segment	52	102	18.6	55.2	39.5
Shuttle	27	57	29	41	14.6

Table 1, Table 2: Accuracy and tree size of PBC and algorithmic approaches

Table 1 depicts the accuracy of PBC and the algorithmic approaches, table 2 their tree sizes. Our performance evaluation demonstrates that the approach of interactive visual classification yields an accuracy similar to the accuracy obtained by well-known algorithms. PBC

significantly reduces the tree size and thus obtains decision trees which are much better understandable.

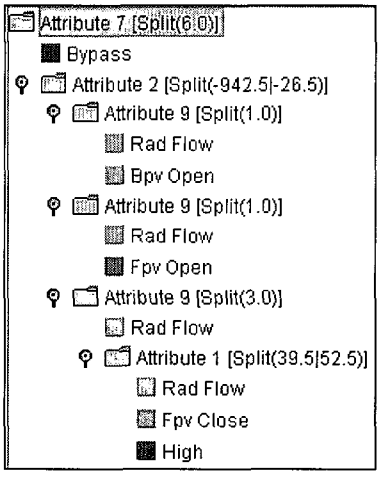


Figure 5. A decision tree for the Shuttle dataset

To illustrate this advantage, figure 5 shows a decision tree for the Shuttle dataset constructed with the PCB system. Attribute 7 represents the root of the tree with one split point at 6.0. The following two nodes are the left (attribute 7 < 6.0) and right son of this root. The left son of the root is already assigned to a class (Bypass). The colored square besides the class label depicts the color representing the class. We observe that the nodes with the splitting attributes 1 and 2 both have two split points yielding a 3-ary decision tree that cannot be generated by the algorithmic approaches.

5 Conclusion

In this paper, we introduced a fully interactive method for decision tree construction based on a multidimensional visualization technique and appropriate interaction capabilities. Thus knowledge can be transfered in both directions. On one hand, domain knowledge of an expert can be profitably included in the tree construction phase. On the other hand, after going through the interactive construction of a decision tree, the user has a much deeper understanding of the data than just knowing the decision tree generated by an arbitrary algorithm. Our approach has several additional advantages compared to algorithmic approaches. First, the user may set an arbitrary number of split points which can reduce the tree size in comparison to binary decision trees that are generated by most state of the art algorithms. Furthermore, in contrast to the greedy search performed by algorithmic approaches, the user can backtrack to any node of the tree when a subtree turns out to be suboptimal. We conducted an experimental evaluation on several popular datasets. We found that even users with no a priori knowledge of the training data construct a decision tree that has a similar accuracy and a significantly smaller tree size compared to algorithmic approaches.

In our future work, we will improve the scalability with respect to the maximum amount of data that can be processed. Furthermore, we plan to extend our PBC system by features of algorithmic approaches and we want to explore methods of integrating PBC with a database management system.

6 References

[1] Ankerst, M., Breunig M, Kriegel H.-P. and Sander J.:"OPTICS: Ordering Points To Identify the Clustering Structure", in *Proc. ACM SIGMOD* '99, Int. Conf. on Management of Data, Philadelphia, PA, 1999.

[2] Ankerst M., Keim D. A. and Kriegel H.-P.: "Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets", *Proc. Visualization '96*, Hot Topic Session, San Francisco, CA, 1996.

[3] Agrawal R., Mehta M. and Shafer J.C.: "SPRINT: A Scalable Parallel Classifier for Data Mining", in *Proc. VLDB '96*, 22nd Intl. Conf. on Very Large Databases, Bombay, India, 1996, pp. 544-555.

[4] Alsabti K., Ranka S. and Singh V.: "CLOUDS: A Decision Tree Classifier for Large Datasets", in *Proc. KDD '98*, 4th Intl. Conf. on Knowledge Discovery and Data Mining, New York City, 1998, pp. 2-8.

[5] Berchtold S., Jagadish H.V. and Ross K.A.: "Independence Diagrams: A Technique for Visual Data Mining", in *Proc. KDD '98*, 4th Intl. Conf. on Knowledge Discovery and Data Mining, New York City, 1998, pp. 139-143..

[6] Fayyad U.M. and Irani K.: "Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning", in *Proc. IJCAI '93*, Int. Joint Conf. on Artificial Intelligence, 1993.

[7] Keim D.A.: "Visual Support for Query Specification and Data Mining", PhD Thesis, University of Munich, Germany, 1994.

[8] Keim D. A.: "Visual Database Exploration Techniques", Proc. Tutorial Int. Conf. on Knowledge Discovery & Data Mining, Newport Beach, CA, 1997. http://www.informatik.uni-halle.de/~keim/PS/KDD97.pdf

[9] Keim D. A., Kriegel H.-P. and Ankerst M.: "Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data", in *Proc. Visualization '95*, Atlanta, GA, 1995, pp. 279-286.

[10] Kohavi R. and Sommerfield D.: "Targeting Business Users with Decision Table Classifiers", in *Proc. KDD '98*, 4th Intl. Conf. on Knowledge Discovery and Data Mining, New York City, 1998, pp. 249-253.

[11] Levkowitz H.: "Perceptual Steps Along Pseudo-Color Scales", International Journal of Imaging Systems and Technology, 7:97-101, 1996.

[12] Mitchell T.M.: "Machine Learning", McGraw Hill, 1997.

[13] Michie D., Spiegelhalter D.J. and Taylor C.C.: "Machine Learning, Neural and Statistical Classification", Ellis Horwood, 1994.
[14] NASA Ames Research Center: "Introduction to IND Version 2.1", 1992.
[15] http://www.spss.com/.

Generalized Blockmodeling

Vladimir Batagelj
University of Ljubljana, Faculty of Mathematics and Physics
Jadranska 19, 1 000 Ljubljana, Slovenia
E-mail: vladimir.batagelj@uni-lj.si
AND
Anuška Ferligoj
University of Ljubljana, Faculty of Social Sciences
Kardeljeva pl. 5, 1 000 Ljubljana, Slovenia
E-mail: anuska.ferligoj@uni-lj.si
AND
Patrick Doreian
University of Pittsburgh, Department of Sociology
PA 15260, Pittsburgh, USA
E-mail: pitpat+@pitt.edu

Keywords: clustering, blockmodeling, social network, pre-specified blockmodel, local optimization.

Edited by: Cene Bavec and Matjaž Gams
Received: October 3, 1999 **Revised:** November 20, 1999 **Accepted:** December 4, 1999

The goal of blockmodeling is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily. In the paper we present an overview of basic ideas and developments in this area.

1 Basic Notions

1.1 Network

Let $E = \{X_1, X_2, \ldots, X_n\}$ be a finite set of *units*. The units are related by binary *relations* $R_t \subseteq E \times E$, $t = 1, \ldots, r, r \geq 1$ which determine a *network*

$$\mathcal{N} = (E, R_1, R_2, \ldots, R_r)$$

In the following we restrict our discussion to a single relation R described by a corresponding binary matrix $\mathbf{R} = [r_{ij}]_{n \times n}$ where

$$r_{ij} = \begin{cases} 1 & X_i R X_j \\ 0 & \text{otherwise} \end{cases}$$

In some applications r_{ij} can be a nonnegative real number expressing the strength of the relation R between units X_i and X_j.

1.1.1 Example: Student Government

In Table 1 and Figure 1 the Student Government network is presented. It consists of communication interactions among twelve members and advisors of the Student Government at the University in Ljubljana (Hlebec, 1993). The results of the measurement are not real interactions among actors but cognition about communication interactions. Data were collected with face to face interviews, conducted in May 1992.

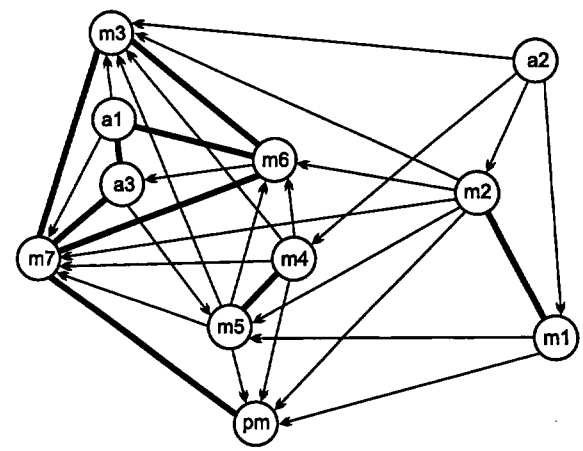


Figure 1: Network graph: Student Government – discussion, recall

Table 1: Student Government matrix

		m 1	p 2	m 3	m 4	m 5	m 6	m 7	m 8	a 9	a 10	a 11
minister 1	1	0	1	1	0	0	1	0	0	0	0	0
p.minister	2	0	0	0	0	0	0	0	1	0	0	0
minister 2	3	1	1	0	1	0	1	1	1	0	0	0
minister 3	4	0	0	0	0	0	0	1	1	0	0	0
minister 4	5	0	1	0	1	0	1	1	1	0	0	0
minister 5	6	0	1	0	1	1	0	1	1	0	0	0
minister 6	7	0	0	0	1	0	0	0	1	1	0	1
minister 7	8	0	1	0	1	0	0	1	0	0	0	1
adviser 1	9	0	0	0	1	0	0	1	1	0	0	1
adviser 2	10	1	0	1	1	1	0	0	0	0	0	0
adviser 3	11	0	0	0	0	0	1	0	1	1	0	0

Communication flow among actors was identified by the following question:

> Of the members and advisors of the Student Government, whom do you (most often) talk with?

The content of the communication flow was limited to the matters of the Student Government. The time frame was also defined: the question was referred to the six months period. One respondent refused to cooperate in the experiment. As he was not considered in the analysis, the network consists of eleven actors.

1.2 Cluster and Clustering

One of the main procedural goals of blockmodeling is to identify, in a given network, *clusters* (classes) of units that share structural characteristics defined in terms of R. The units within a cluster have the same or similar connection patterns to other units. They form a *clustering*

$$\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$$

which is a partition of the set E: $\bigcup_i C_i = E$ and $i \neq j \Rightarrow C_i \cap C_j = \emptyset$. Each partition determines an equivalence relation (and vice versa).

1.3 Block

A clustering $\mathcal{C}$ partitions also the relation R into *blocks*

$$R(C_i, C_j) = R \cap C_i \times C_j$$

Each such block consists of units belonging to clusters C_i and C_j and all arcs leading from cluster C_i to cluster C_j. If $i = j$, a block $R(C_i, C_i)$ is called a *diagonal* block.

1.4 Blockmodel and Blockmodeling

The goal of blockmodeling is to reduce a large, potentially incoherent network to a smaller comprehensible structure that can be interpreted more readily. Blockmodeling, as an empirical procedure, is based on the idea that units in a network can be grouped according to the extent to which they

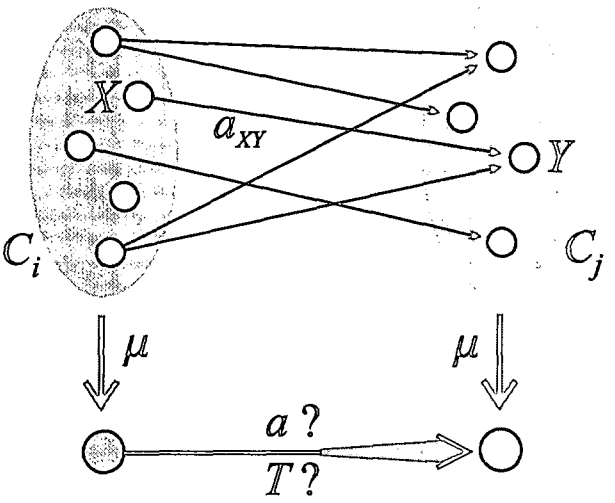


Figure 2: Blockmodeling scheme.

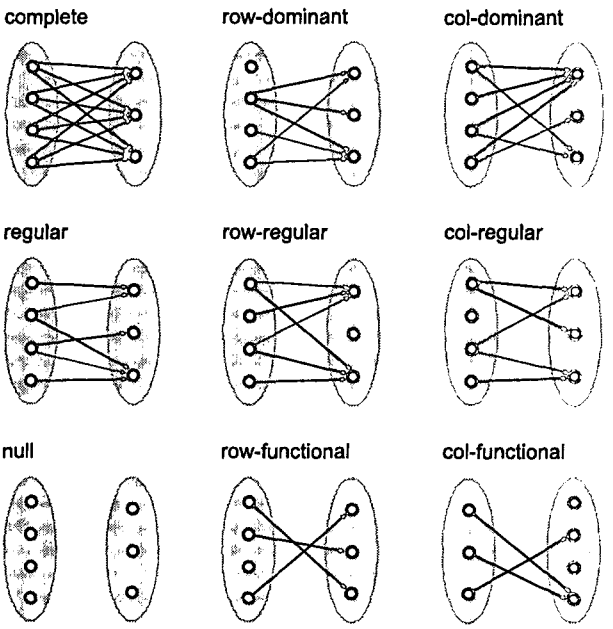


Figure 3: Types of connection between two sets; the left set is the ego-set.

are equivalent, according to some *meaningful* definition of equivalence.

A *blockmodel* consists of structures obtained by identifying all units from the same cluster of the clustering $\mathcal{C}$. For an exact definition of a blockmodel (see Figure 2) we have to be precise also about which blocks produce an arc in the *reduced graph* and which do not, and of what *type*. Some types of connections are presented in Figure 3. A block is *symmetric* if

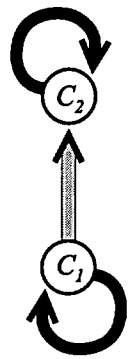
$$\forall (X, Y) \in C_i \times C_j : (X\, R\, Y \Leftrightarrow Y\, R\, X)$$

Note that for nondiagonal blocks this condition involves a pair of blocks $R(C_i, C_j)$ and $R(C_j, C_i)$.

Table 2: Block types and matrices.

1	1	1	1	1	1	0	0
1	1	1	1	0	1	0	1
1	1	1	1	0	0	1	0
1	1	1	1	1	0	0	0
0	0	0	0	0	1	1	1
0	0	0	0	1	0	1	1
0	0	0	0	1	1	0	1
0	0	0	0	1	1	1	0

	C_1	C_2
C_1	complete	regular
C_2	null	complete



The reduced graph can be presented by relational matrix, called also *image matrix* (see Table 2).

A clustering and the induced blockmodel of the Student Government is presented in Figure 4.

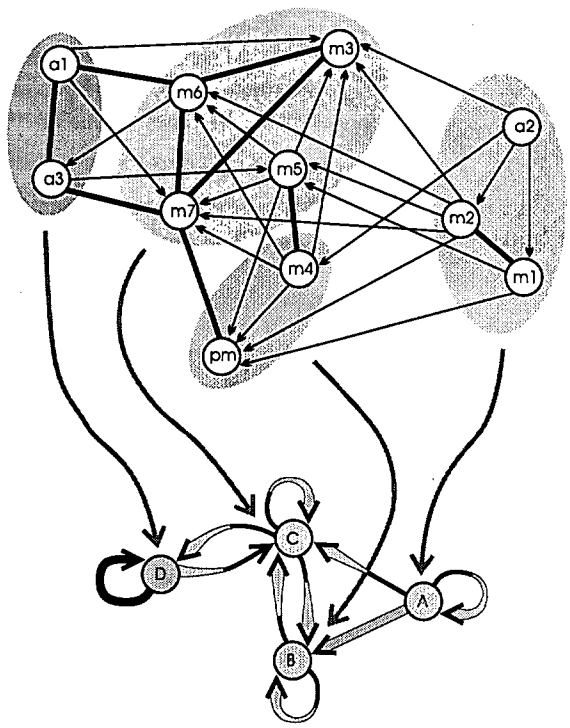


Figure 4: Blockmodeling example.

2 Blockmodeling - Formalization

Let U be a set of *positions* or images of clusters of units. Let $\mu : E \to U$ denote a mapping which maps each unit to its position. The cluster of units $C(t)$ with the same position $t \in U$ is

$$C(t) = \mu^{-1}(t) = \{X \in E : \mu(X) = t\}$$

Therefore

$$\mathcal{C}(\mu) = \{C(t) : t \in U\}$$

is a partition (clustering) of the set of units E.

A *blockmodel* is an ordered sextuple $\mathcal{M} = (U, K, \mathcal{T}, Q, \pi, \alpha)$ where:

- U is a set of *positions* (types of units);
- $K \subseteq U \times U$ is a set of *connections*;
- $\mathcal{T}$ is a set of predicates used to describe the types of connections between different clusters in a network. We assume that nul $\in \mathcal{T}$.
- a mapping $\pi : K \to \mathcal{T} \setminus \{\text{nul}\}$ assigns predicates to connections;
- Q is a set of *averaging rules*. A mapping $\alpha : K \to Q$ determines rules for computing values of connections.

A (surjective) mapping $\mu : E \to U$ determines a blockmodel $\mathcal{M}$ of network $\mathcal{N}$ iff it satisfies the conditions:

$$\forall (t, w) \in K : \pi(t, w)(C(t), C(w))$$

and

$$\forall (t, w) \in U \times U \setminus K : \text{nul}(C(t), C(w)).$$

2.1 Equivalences

Let $\approx$ be an equivalence relation over E and $[X] = \{Y \in E : X \approx Y\}$. We say that $\approx$ is *compatible* with $\mathcal{T}$ over a network $\mathcal{N}$ iff

$$\forall X, Y \in E \, \exists T \in \mathcal{T} : T([X], [Y]).$$

It is easy to verify that the notion of compatibility for $\mathcal{T} = \{\text{nul}, \text{reg}\}$ reduces to the usual definition of regular equivalence (White and Reitz 1983). Similarly, compatibility for $\mathcal{T} = \{\text{nul}, \text{com}\}$ reduces to structural equivalence (Lorrain and White 1971).

For a compatible equivalence $\approx$ the mapping μ: $X \mapsto [X]$ determines a blockmodel with $U = E/\approx$.

3 Optimization

3.1 Criterion Function

The problem of establishing a partition of units in a network in terms of a selected type of equivalence is a special case of **clustering problem** that can be formulated as an optimization problem: determine the clustering $\mathcal{C}^*$ for which

$$P(\mathcal{C}^*) = \min_{\mathcal{C} \in \Phi} P(\mathcal{C})$$

Table 3: Characterizations of Types of Blocks.

null	nul	all 0*
complete	com	all 1*
row-regular	rre	each row is 1-covered
col-regular	cre	each column is 1 -covered
row-dominant	rdo	∃ all 1 row*
col-dominant	cdo	∃ all 1 column*
regular	reg	1-covered rows and 1-covered columns
non-null	one	∃ at least one 1

* except may be diagonal

where $\mathcal{C}$ is a clustering of a given *set of units* E, Φ is the set of all feasible clusterings and $P : \Phi \to \mathbb{R}$ the *criterion function*.

One of the possible ways of constructing a criterion function that directly reflects the considered equivalence is to measure the fit of a clustering to an ideal one with perfect relations within each cluster and between clusters according to the considered equivalence.

Given a set of types of connection $\mathcal{T}$ we can introduce the set of *ideal blocks* for a given type $T \in \mathcal{T}$ by

$$\mathcal{B}(C_i, C_j; T) = \{B \subseteq C_i \times C_j : T(B)\}$$

Using Table 3 we can efficiently test whether the block $R(C_i, C_j)$ is of the type T; and define the *deviation* $\delta(C_i, C_j; T)$ of a block $R(C_i, C_j)$ from the nearest ideal block. For example

$$\delta(C_i, C_j; \text{reg}) = |C_i| \cdot (|C_j| - c_j) + |C_j| \cdot (|C_i| - r_i)$$

where c_j is the number of non-zero columns, and r_i is the number of non-zero rows in the block $R(C_i, C_j)$. For details see (Batagelj 1997).

For the proposed types all deviations are *sensitive*

$$\delta(C_i, C_j; T) = 0 \Leftrightarrow T(R(C_i, C_j)).$$

Therefore a block $R(C_i, C_j)$ is of a type T exactly when the corresponding deviation $\delta(C_i, C_j; T)$ is 0. In the deviation δ we can also incorporate values of lines ν.

Based on deviation $\delta(C_i, C_j; T)$ we introduce the *block-error* $\varepsilon(C_i, C_j; T)$ of $R(C_i, C_j)$ for type T. An example of block-error is

$$\varepsilon(C_i, C_j; T) = w(T)\delta(C_i, C_j; T)$$

where $w(T) > 0$ is a weight of type T.

We extend the block-error to the set of feasible types $\mathcal{T}$ by defining

$$\varepsilon(C_i, C_j; \mathcal{T}) = \min_{T \in \mathcal{T}} \varepsilon(C_i, C_j; T)$$

and

$$\pi(\mu(C_i), \mu(C_j)) = \text{argmin}_{T \in \mathcal{T}} \varepsilon(C_i, C_j; T)$$

To make π well-defined, we order (priorities) the set $\mathcal{T}$ and select the first type from $\mathcal{T}$ which minimizes ε. We combine block-errors into a *total error* – blockmodeling *criterion function*

$$P(\mathcal{C}(\mu); \mathcal{T}) = \sum_{(t,w) \in U \times U} \varepsilon(C(t), C(w); \mathcal{T}).$$

For criterion function P we have

$$P(\mathcal{C}(\mu)) = 0 \Leftrightarrow \mu \text{ is an exact blockmodeling}$$

The obtained optimization problem can be solved by local optimization. Once a partitioning μ and types of connection π are determined, we can also compute the values of connections by using *averaging rules*.

3.2 Local Optimization

For solving the blockmodeling problem we use a local optimization procedure (relocation algorithm):

Determine the initial clustering $\mathcal{C}$;
repeat:
  if in the neighborhood of the current clustering $\mathcal{C}$ there exists a clustering $\mathcal{C}'$ such that $P(\mathcal{C}') < P(\mathcal{C})$ **then** move to clustering $\mathcal{C}'$.

The neighborhood in this local optimization procedure is determined by the following two transformations:

- *moving* a unit X_k from cluster C_p to cluster C_q (*transition*);
- *interchanging* units X_u and X_v from different clusters C_p and C_q (*transposition*).

3.3 Benefits from Optimization Approach

- *ordinary / inductive blockmodeling*: Given a network $\mathcal{N}$ and set of types of connection $\mathcal{T}$, determine $\mathcal{M}$, i.e., μ, π and α;
- *evaluation of the quality of a model, comparing different models, analyzing the evolution of a network* (Sampson data, Doreian and Mrvar 1996): Given a network $\mathcal{N}$, a model $\mathcal{M}$, and blockmodeling μ, compute the corresponding criterion function;
- *model fitting / deductive blockmodeling*: Given a network $\mathcal{N}$, set of types $\mathcal{T}$, and a model $\mathcal{M}$, determine μ which minimizes the criterion function (Batagelj, Ferligoj, Doreian, 1998).
- we can fit the network to a partial model and analyze the residual afterward;
- we can also introduce different constraints on the model, for example: units X and Y are of the same type; or, types of units X and Y are not connected; ...

4 Pre-Specified Blockmodels

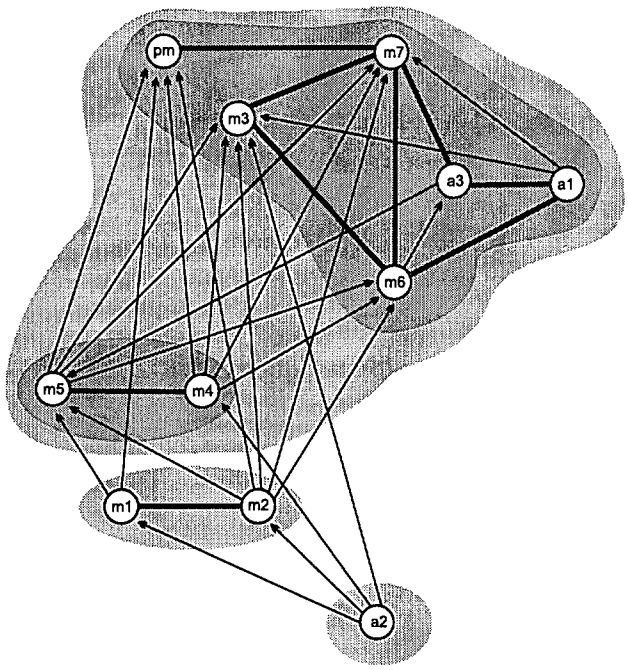


Figure 5: Symmetric acyclic blockmodel of Student Government.

The pre-specified blockmodeling starts with a blockmodel specified, in terms of substance, *prior to an analysis*. Given a network, a set of ideal blocks is selected, a reduced model is formulated, and partitions are established by minimizing the criterion function. The pre-specified blockmodeling is supported by the program MODEL 2 (Batagelj, 1996).

As an example of pre-specified blockmodel we present in Figure 5 a symmetric acyclic blockmodel of Student Government. The obtained clustering in 4 clusters is almost exact – acyclic model with symmetric clusters. The only error is produced by the arc $(a3, m5)$.

5 Final Remarks

The current, local optimization based, programs for generalized blockmodeling can deal only with networks with at most some hundreds of units. What to do with larger networks is an open question. For some specialized problems also procedures for (very) large networks can be developed (Doreian, Batagelj, Ferligoj, 1998).

Another interesting problem is the development of blockmodeling of valued networks.

MODEL 2 and related programs and data can be obtained from

```
http://vlado.fmf.uni-lj.si/
      pub/networks/stran/
```

Acknowledgment: This work was supported by the Ministry of Science and Technology of Slovenia, Project J1-8532.

References

[1] Batagelj, V. (1991): STRAN – STRucture ANalysis. Manual, Ljubljana.

[2] Batagelj, V. (1997): Notes on Blockmodeling. *Social Networks*, **19**, 143-155.
Also in: *Abstracts and Short Versions of Papers*, 3rd European Conference on Social Network Analysis, München, 1993: DJI, 1-9.

[3] Batagelj, V., P. Doreian, and A. Ferligoj (1992): An optimizational approach to regular equivalence. *Social Networks*, 14:121–135.

[4] Batagelj, V., A. Ferligoj, and P. Doreian (1992): Direct and indirect methods for structural equivalence. *Social Networks*, 14:63–90.

[5] Batagelj, V., Ferligoj, A., and Doreian, P. (1998): Fitting Pre-Specified Blockmodels, in *Data Science, Classification, and Related Methods*, Eds., C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. H. Bock, and Y. Baba, Springer-Verlag, Tokyo, p.p. 199-206.

[6] Borgatti, S.P. and M.G. Everett (1989): The class of all regular equivalences: Algebraic structure and computation. *Social Networks*, 11:65–88.

[7] Doreian, P., V. Batagelj, and A. Ferligoj (1994): Partitioning Networks on Generalized Concepts of Equivalence. *Journal of Mathematical Sociology*, 19/1:1–27.

[8] Doreian, P., V. Batagelj, and A. Ferligoj (1998): Symmetric-Acyclic Decompositions of Networks. To appear in *Journal of Classification*.

[9] Doreian, P. and A. Mrvar (1996) A Partitioning Approach to Structural Balance. *Social Networks* 18:149–168.

[10] Faust, K. (1988): Comparison of methods for positional analysis: Structural and general equivalences. *Social Networks*, 10:313–341.

[11] A. Ferligoj, V. Batagelj, and Doreian, P. (1994): On Connecting Network Analysis and Cluster Analysis. In *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (G.H. Fischer, D. Laming Eds.), New York: Springer.

[12] Lorrain, F. and H.C. White (1971): Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1:49–80.

[13] Hlebec, V. (1993): Recall versus recognition: Comparison of two alternative procedures for collecting social network data. Developments in Statistics and Methodology. (A. Ferligoj and A. Kramberger, editors) Metodološki zvezki 9, Ljubljana: FDV, 121-128.

[14] White, D.R. and K.P. Reitz (1983): Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5:193–234.

Adapted Methods For Clustering Large Datasets Of Mixed Units

Simona Korenjak-Černe
IMFM Ljubljana, Dept. of TCS,
Jadranska 19, 1 000 Ljubljana, Slovenia
E-mail: simona.korenjak@fmf.uni-lj.si

Keywords: clustering, large datasets, mixed units, hierarchical clustering, cluster description compatible with merging of clusters, leaders method, adding clustering method

Edited by: Cene Bavec and Matjaž Gams

Received: October 17, 1999 **Revised:** October 30, 1999 **Accepted:** December 11, 1999

The proposed clustering methods are based on the recoding of the original mixed units and their clusters into a uniform representation. The description of a cluster consists for each variable of the frequencies of the variable values over its range partition. The proposed representation can be used also for clustering symbolic data. On the basis of this representation the adapted version of the leaders method and adding clustering method were implemented. We describe both approaches, which were successfully applied on several large datasets.

1 Introduction

Abstraction is the main tool to deal with large amounts of data. The first step is to identify groups of similar units - clusters. In data analysis this is a task of clustering methods. The most popular are hierarchical clustering methods. Because they usually use a similarity/dissimilarity matrix they are appropriate only for clustering datasets of a moderate size (some hundreds of units). On the other hand well known nonhierarchical methods are mostly implemented for datasets of variables measured in the same scale type (such as for example 'k-means method'). Because of these limits we are searching for new clustering methods or at least trying to adapt known methods to be appropriate for clustering **large** datasets of **mixed** units, where variables (properties) of the units are measured in different scales.

Let E be a finite *set of units*. A nonempty subset $C \subseteq E$ is called a *cluster*. A set of clusters $\mathbf{C} = \{C_i\}$ forms a *clustering*. In this paper we shall require that every clustering $\mathbf{C}$ is a partition of E.

The clustering problem can be formulated as an optimization problem:

Determine the clustering $\mathbf{C}^* \in \Phi$, for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where Φ is a set of *feasible clusterings* and $P : \Phi \to \mathbb{R}_0^+$ is a *criterion function*.

In many clustering methods the criterion function measures the deviation of units from *representatives* (*leaders*) of corresponding clusters. In our method we select the criterion function in one of the most frequent form

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \sum_{X \in C} d(X, R_C)$$

where R_C is a representative of cluster C and d is a dissimilarity.

The cluster representatives usually consist of variable-wise summaries of variable values over the cluster. For homogeneous units with only numerical variables their means are usually selected as representatives of clusters. For mixed (nonhomogeneous) units a new description has to be selected.

In this paper we investigate a description satisfying two additional requirements:

1. it should require a fixed space per variable;
2. it should be *compatible* with merging of clusters – knowing the description of two disjoint clusters we can, without additional information, produce the description of their union.

Note that only some of the cluster descriptions are compatible with merging. For example mean (as sum and number of units) for numerical variables and (min, max) intervals for ordinal variables.

2 A description of a cluster

For our adaptation of clustering methods to be appropriate for clustering large datasets of mixed units, we choose a cluster description based on frequencies. For this purpose, the ranges of the variables are partitioned into selected number of classes. Let $\{V_i, i = 1, \ldots, k(V)\}$ be a partition of the range of values of variable V (the number of classes $k(V)$ depends on variable). Then we can define for a cluster C the sets

$$Q(i, C; V) = \{X \in C : V(X) \in V_i\},\ i = 1, \ldots, k(V)$$

where $V(X)$ denotes the value of variable V on unit X.

In the case of an ordinal variable V (numerical scales are a special case of ordinal scales) the partition $\{V_i, i = 1, \ldots, k(V)\}$ usually consists of intervals determined by selected threshold values $t_0 < t_1 < t_2 < t_3 < \cdots < t_{k(V)-1} < t_{k(V)}$, $t_0 = \inf V$, $t_{k(V)} = \sup V$.

For nominal variables we can obtain the partition, for example, by selecting $k(V) - 1$ values $t_1, t_2, t_3, \ldots, t_{k(V)-1}$ from the range of variable V (usually the most frequent values on E) and setting $V_i = \{t_i\}$, $i = 1, \ldots, k(V) - 1$; and putting all the remaining values in class $V_{k(V)}$.

Units are not necessarily represented with single value for each variable, but they can also be represented with frequencies over the classes of variables ranges.

Using classes of ranges we get *frequencies*

$$q(i, C; V) = \operatorname{card} Q(i, C; V)$$

and *relative frequencies*

$$p(i, C; V) = \frac{q(i, C; V)}{\operatorname{card} C}$$

Note that

$$\sum_{i=1}^{k(V)} p(i, C; V) = 1$$

When only a single unit is in the cluster C we get

$$p(i, C; V) = \begin{cases} 1; & \text{if } X \in Q(i, C; V) \\ 0; & \text{otherwise} \end{cases}$$

We can add, for each variable, a new class for a missing value and treat it as a special value, or we can also consider a missing value on V for a unit X by setting $p(i, \{X\}; V) = \frac{1}{k(V)}$, $i = 1, \ldots, k(V)$ (or by some other distribution).

It is easy to see that such a description is compatible with merging, because for two disjoint clusters C_1 and C_2 we have

$$Q(i, C_1 \cup C_2; V) = Q(i, C_1; V) \cup Q(i, C_2; V),$$

$$q(i, C_1 \cup C_2; V) = q(i, C_1; V) + q(i, C_2; V).$$

The threshold values are usually determined in such a way that, for the given set of units E (or the space of units $\mathcal{E}$), it holds that $p(i, E; V) \approx \frac{1}{k(V)}$, $i = 1, \ldots, k(V)$.

As a compatible description of nominal variable over a cluster C also its range $V(C)$ can be used, since we have $V(C_1 \cup C_2) = V(C_1) \cup V(C_2)$.

Example: Recoding of flags dataset

Original data are taken from the address
`ftp://ftp.ics.uci.edu/pub/machine-learning-databases/flags`
(Flags from Collins Gem Guide to Flags, donated by Richard S. Forsyth.)

Let us consider the following three variables:
- *population* (in round millions),
- *mainhue* (predominant color in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)),
- *text* (1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise).

The range of the variable *population* is divided into 5 classes with approximately the same number of units in each of them. The ranges of the others variables are so small that we put for discretization of them each possible value in a separate class:

var= *population*	var= *mainhue*	var= *text*
map	map	map
1 = {0}	1 = {red}	1 = {0}
2 = (0, 4]	2 = {green}	2 = {1}
3 = (4, 18]	3 = {blue}	
4 = (18, 158]	4 = {gold}	
5 = (158, 1100]	5 = {white}	
	6 = {black}	
	7 = {orange}	

ORIGINAL DATA

unit ID	population	mainhue	text
Austria	8	red	0
New-Zealand	2	blue	0
Saudi-Arabia	9	green	1
Switzerland	6	red	0
USA	231	white	0

RECODED DATA

Austria	3	1	1
New-Zealand	2	3	1
Saudi-Arabia	3	2	2
Switzerland	3	1	1
USA	5	5	1

In our case for each variable a unit is represented with index of the appropriate class.

The description of a cluster C_6 (only for considered variables) obtained with the leaders method is

$q(C_6; population)$	8	1	1	0	0		
$q(C_6; mainhue)$	1	0	9	0	0	0	0
$q(C_6; text)$	6	4					

>From this description we can see that in eight countries the population is less than a million, in one country is between 1 and 4 millions and in one country population is between 4 and 18 millions. This cluster is one of the seven clusters obtained with the adapted version of the leaders method with maximal allowed dissimilarity between a unit and its nearest leader 0.5. In one of the countries flags red is a dominant color and all of the remaining units have blue

mainhue. In six units some text is presented and four units inside cluster C_6 have no text in their description. For better understanding, cluster C_6 consists of: Bermuda, Brit. Virg. Isles, Cayman Islands, Falklands Malvi, Fiji, Hong-Kong, Montserrat, St. Helena, Turks Cocos Islands and Tuvalu.

3 Dissimilarity between clusters

Let us return to our approach to clustering problem as an optimization problem. After deciding to use the uniform representation of units and clusters, we have to define a measure of dissimilarity between clusters (a unit is a special case of a cluster with only one element). First the dissimilarity between clusters for individual variable V is defined as

$$d(C_1, C_2; V) = \frac{1}{2} \sum_{i=1}^{k(V)} |p(i, C_1; V) - p(i, C_2; V)|.$$

We shall use the abbreviation

$$d(X, C; V) = d(\{X\}, C; V).$$

In both cases it can be shown that

1. $d(C_1, C_2; V)$ is a semidistance on clusters; i.e.
   (a) $d(C_1, C_2; V) \geq 0$
   (b) $d(C, C; V) = 0$
   (c) $d(C_1, C_2; V) + d(C_2, C_3; V) \geq d(C_1, C_3; V)$
2. $d(C_1, C_2; V) \in [0, 1]$

and for the representation of a single unit also

$$X \in Q(i, E; V) \Rightarrow d(X, C; V) = 1 - p(i, C; V)$$

The semidistances on clusters for individual variable can be combined into a semidistance on clusters for complete descriptions by

$$d(C_1, C_2) = \sum_{j=1}^{m} \alpha_j \, d(C_1, C_2; V_j),$$

where m is the number of variables and α_j are weights ($\alpha_j \geq 0$ and $\sum_{j=1}^{m} \alpha_j = 1$); often $\alpha_j = \frac{1}{m}$. We can use weights to consider dependencies among variables or to tune the dissimilarity to a given learning set in AI applications.

4 Clustering procedures

In the proposed approach the original nonhomogeneous data are first recoded to a uniform representation. For the recoded data efficient clustering procedures can be built by adapting leaders method (Hartigan, 1975) or adding clustering method (Zupan 1982, Jambu and Lebeaux 1983, Batagelj and Mandelj 1993).

4.1 The adapted version of the leaders method

The adapted version of the leaders method is a variant of a dynamic clustering method (Diday 1979, Batagelj 1985). To describe the dynamic clustering method for solving the clustering problem let us denote: Λ a set of *representatives*; $\mathbf{L} \subseteq \Lambda$ a *representation*; Ψ a set of *feasible representations*; $P : \Phi \to \mathbb{R}_0^+$ *criterion function*; $G : \Phi \to \Psi$ *a representation function*; $F : \Psi \to \Phi$ *a clustering function* and suppose that the functions G and F tend to improve (diminish) the value of the criterion function P. Then a simple version of the dynamic clustering method can be described by the scheme:

$\mathbf{L} := \mathbf{L}_0$;
repeat
  $\mathbf{C} := F(\mathbf{L})$
  $\mathbf{L} := G(\mathbf{C})$
until the leaders stabilize

We begin with the initial representation and then repeat to assign each unit to the nearest leader and after that select leaders for each (new) cluster until we reach the minimum of the criterion function or until the leaders don't change any more (local minimum).

Let us assume the following model $\mathbf{C} = \{C_i\}_{i \in I}$, $\mathbf{L} = \{L_i\}_{i \in I}$, $\mathbf{L}(X) = L_i : X \in C_i$ (the nearest leader to the unit X), $L = [L(V_1), \ldots, L(V_m)]$, $L(V) = [s(1, L; V), \ldots, s(k(V), L; V)]$, $\sum_{j=1}^{k(V)} s(j, L; V) = 1$ (the description of a leader has the same form as the description of a cluster) and

$$d(C, L; V) = \frac{1}{2} \sum_{j=1}^{k(V)} |p(j, C; V) - s(j, L; V)|.$$

For selected criterion function

$$P(\mathbf{C}) = \sum_{X \in E} d(X, \mathbf{L}(X)) = \sum_{i \in I} p(C_i, L_i)$$

where

$$p(C, L) = \sum_{X \in C} d(X, L)$$

we define $F(\mathbf{L}) = \{C_i'\}$ with

$$X \in C_i' : i = \min \operatorname{Argmin}_j \{d(X, L_j) : L_j \in \mathbf{L}\}.$$

This means that each unit is assigned to the (first) nearest leader.
We define $G(\mathbf{C}) = \{L_i'\}$ with

$$L_i' = \operatorname*{argmin}_{L \in \Psi} p(C, L).$$

The unique symmetric optimal solution of this optimization problem is

$$s(i, L'; V) = \begin{cases} \frac{1}{t}; & \text{if } j \in M \\ 0; & \text{otherwise} \end{cases}$$

where $M = \{j : q(j, C; V) = \max_i q(i, C; V)\}$ and $t = \text{card}\, M$.

The representative (leader) of a cluster is obtained from the most frequent range(s) of values of variables on this cluster.

Example: Leader of a cluster

For the description of a cluster C_6

$q(C_6; population)$	8	1	1	0	0		
$q(C_6; mainhue)$	1	0	9	0	0	0	0
$q(C_6; text)$	6	4					

the optimal leader L_6 is

$q(C_6; population)$	1	0	0	0	0		
$q(C_6; mainhue)$	0	0	1	0	0	0	0
$q(C_6; text)$	1	0					

The characteristics of the cluster are

population = less than a million	80 %
mainhue in the flag = blue	90 %
text in the flag = no	60 %

For example, 80% of all countries in the cluster C_6 have less than a million inhabitants, 90% of all countries flags have blue mainhue and 60% of the flags in the cluster have no text in their descriptions.

Properties of the leaders method

The main properties of the adapted version of the leaders method are:

1. Selection of the leaders and formation of new clusters diminish the value of the criterion function.
2. The program always stops (converges). The number of iterations is usually less than 10.
3. The program is suitable for clustering (very) large datasets.
4. The leaders descriptions provide us with simple interpretations of clustering results.

4.2 The adapted adding method

The adding clustering method is a hierarchical clustering method in which a new unit is added in a clustering tree. Each vertex corresponds to a cluster. For large datasets usually only the upper part of the hierarchy is maintained, the lower levels subtrees are replaced by 'bags' containing all units from a subtree.

We shall use the same description of a cluster (vertex) and the same definition of a dissimilarity as in the leaders method. Every time we add a unit in a cluster (vertex) the frequencies are recalculated. There are two possible ways how to add a new unit:

a) To maximize the dissimilarity between clusters (sons) of the current vertex or,

b) To minimize the dissimilarity from clusters (sons) of the current vertex.

In the first case (see Figure 1) the dissimilarities between both sons of a current vertex are calculated. Because of greedy approach the case with the biggest dissimilarity is chosen: $\max\{d(C_p \cup \{X\}, C_q), d(C_p, C_q \cup \{X\}), d(C, \{X\})\}$.

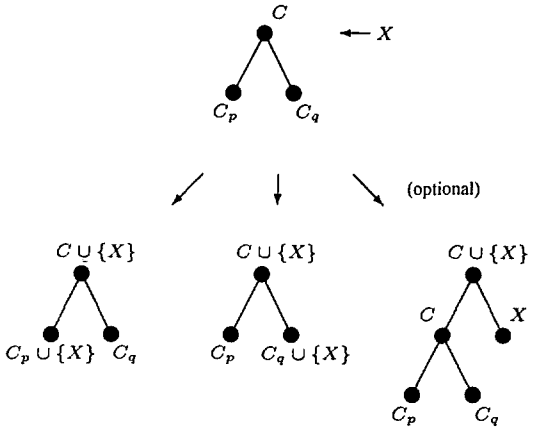


Figure 1: Maximize the dissimilarity between clusters

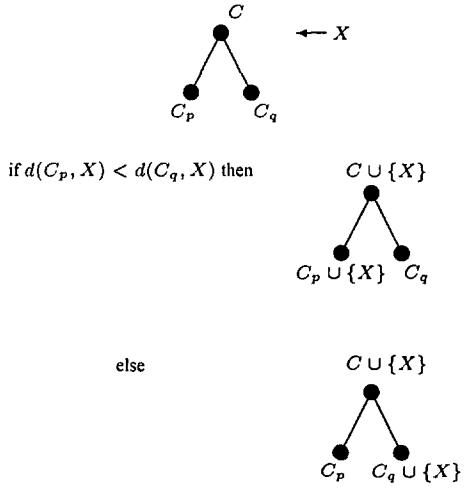


Figure 2: Minimize the dissimilarity from clusters

In the second case (see Figure 2) the dissimilarities from each of the sons of current vertex are calculated and the unit is added to the nearest one: $\min\{d(C_p, \{X\}), d(C_q, \{X\})\}$.

The proposed approaches can also be extended on non-binary trees.

The adding clustering method has some advantages:

1. Presentation of the result with a tree.

2. It can be used for classification.

3. Speed up - if the tree has many (hundreds of) leaves which represent the leaders, it is more efficient adding unit into the tree with this method than to calculate the dissimilarities to each of the leaders.

A drawback of the adding method is that the result strongly depends on the ordering of the input sequence of units. A possible way to avoid this problem is to select a 'good' initial tree. We are suggesting to built the initial tree with some agglomerative hierarchical clustering method on leaders obtained with the leaders method. The other possibility is to include balancing of the tree in the process of adding new unit. Both possibilities are still under the development.

5 Conclusion

We successfully applied the proposed approach on the dataset of types of cars (1 349 units, 26 variables), on the ISSP data (45 784 units, 21 variables) and also on some large datasets from AI collection

`http://www.ics.uci.edu/~mlearn/MLRepository.html`

The first version of the program ClaMix (based on the adapted version of the leaders method) and some of the results are available at

`http://www.educa.fmf.uni-lj.si/datana/`

Acknowledgment: This work was supported by the Ministry of Science and Technology of Slovenia, Project J1-8532.

References

[1] Batagelj, V. (1985) Notes on the dynamic clusters method. *Proceedings of the IV conference on applied mathematics*, Split, May 28-30, 1984. University of Split, Split, p. 139-146.

[2] Batagelj, V. & Bren, M. (1995) Comparing Resemblance Measures. *Journal of Classification*, 12, 1, p. 73-90.

[3] Batagelj, V. & Mandelj, M. (1993) Adding Clustering Algorithm Based on L-W-J Formula. Paper presented at: *IFCS 93*, Paris, 31.aug-4.sep 1993.

[4] Brucker, P. (1978) On the complexity of clustering problems. *Lecture Notes in Economics and Mathematical Systems 175*, in: *Optimization and Operations Research, Proceedings*, Bonn. Henn,R., Korte,B., Oettli,W. (Eds.), Springer-Verlag, Berlin 1978.

[5] Diday, E. (1979) *Optimisation en classification automatique*, Tome 1.,2. INRIA, Rocquencourt, (in French).

[6] Diday, E. (1997) Extracting Information from Extensive datasets by Symbolic Data Analysis. *Indo-French Workshop on Symbolic Data Analysis and its Applications*, Paris, 23-24. September 1997, Paris IX, Dauphine, p. 3-12.

[7] Hartigan, J.A. (1975) *Clustering Algorithms*. Wiley, New York.

[8] Jambu, M. & Lebeaux, M.O. (1983) *Cluster Analysis and Data Analysis*. North-Holland Publishing Company.

[9] Korenjak-Černe, S. & Batagelj, V. (1998). Clustering large datasets of mixed units. *Advances in Data Science and Classification*. Rizzi, A., Vichi, M. and Bock, H.-H. (Eds.), Springer, Berlin, 1998, p. 43-48.

[10] Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

[11] Zupan, J. (1982) *Clustering of Large Data Sets*. Research Studies Press, John Wiley & Sons LTD.

[12] Flags from Collins Gem Guide to Flags. Collins Publishers (1986). Donated by Richard S. Forsyth.
`ftp://ftp.ics.uci.edu/pub/machine-learning-databases/flags`

Equation Discovery System And Neural Networks For Short-Term Dc Voltage Prediction

Irena Nančovska, Anton Jeglič and Dušan Fefer
Faculty of Electrical Engineering,
Tržaška 25, Ljubljana, Slovenia
Phone: +386 61 1768 216, Fax: +386 61 1768 214
E-mail: {Irena.Nancovska,Anton.Jeglic,Dusan.Fefer}@fe.uni-lj.si
AND
Ljupčo Todorovski,
Jozef Stefan Institute,
Jamova 39, Ljubljana, Slovenia
Phone: +386 61 1773 307, Fax: +386 61 1258 058
E-mail: Ljupco.Todorovski@ijs.si

Keywords: neural networks, equation discovery, machine learning

Edited by: Cene Bavec and Matjaž Gams
Received: October 12, 1999 **Revised:** November 25, 1999 **Accepted:** December 15, 1999

The aim of the paper is to compare the predictive abilities of the novel method for time series prediction that is based on equation discovery with neural networks. Both methods are used for short-term (one-step ahead) prediction and have the ability to learn from examples. With purpose to validate the predictive models, they are applied to several data sets. The successful predictive models could be used for voltage monitoring in a high precision solid-state DC voltage reference source (DCVRS) without presence of a high level standard, and further for voltage correction as a segment in the software controlled voltage reference elements (VRE).

1 Introduction

Measured time series could be described as mixtures of dynamic, deterministic part which drives the process and observational noise which is added in the measurement process, but does not influence the future behavior of the system. Many up-to-date scientific researches on predicting the future behavior of system are based on modelling of the deterministic part. Examples ranges from the irregularity in the annual number of sunspots to the changes of currency exchange rates. To make a forecast if the underlying deterministic equations of the observed system are not known, one must find out both the rules governing system dynamics and the present state of the system (Gershenfeld & Weigend 1992). Mainstream statistical techniques for predicting include variations of the auto-regressive technique that Yule invented in 1927. The technique uses weighted sum of previous observations of the series to predict the next value. However, there are a number of cases for which this paradigm is inadequate because of the non-linearity of the underlying model (Gershenfeld & Weigend 1992). In the paper we present two different paradigms for forecasting: neural networks and equation discovery. Neural networks used for prediction are characterized as black-box models whereas models obtained with equation discovery systems are transparent (white-box).

Neural networks represent an emerging technology with some important characteristics: universal approximation (input-output mapping), ability to learn from and adapt to their environment and the ability to evoke weak assumption about the underlying physical system which generates the input data (Haykin 1998). In the paper we use three types of neural networks. The first one is a supervised multilayer feedforward network, which is trained with back-propagation learning algorithm (Haykin 1998, Nielsen 1990, Pham 1995). The second type emphasizes the role of time as an essential dimension of learning. It is a natural extension of the first type, replacing the ordinary synaptic weights with finite-duration impulse response (FIR) filters (Haykin 1998, Gershenfeld & Weigend 1992). The third type of network a recurrent structure with a hidden neurons which introduce time in the network processing by virtue of the built-in feedback loop (Alippi 1996, Haykin 1998).

Equation discovery systems explore the hypothesis space of all equations that can be constructed given a set of arithmetical operators, functions and variables, searching for an equation that fits the input data best. In the paper, we present an equation discovery system LAGRAMGE that uses context free grammars for restricting the hypothesis space of equations. The hypothesis space of LAGRAMGE is a set of equations, such that the expressions on their right hand sides can be derived from a given context free grammar. For the purpose of time series prediction, we use

difference equations, that predicts the present value of the time series. Three different grammars for linear, quadratic and piecewise linear equations are used.

In order to compare the predictive abilities of two described paradigms we performed experiments in two synthetic and three real world time series prediction problems. The domains used in the experiments present models with different amounts of non-linear dynamics (determinism) and noise (randomness). The predictive models obtained in the experiment with reference voltage domain can be used in to improve the metrological characteristics of a DCVRS in two different manners: voltage monitoring and voltage correction. For the purpose of voltage monitoring, predictive models could be used during the inter-calibration period without presence of a high level standard while the predictors are obtained during the calibration period by using a high precision instrument. Further, the models could be used for voltage correction, as a segment in a software controlled VRE. By implementation of a control loop for voltage correction, based on the obtained predictors, the sensitivity of the reference source could be reduced, which contributes to enhancement of the robustness of the system and thereby the stability of the reference voltage (Nancovska 1997).

The paper is organized as follows. First two sections describe the techniques used for time series predicting. In Section 2 a brief description of used neural networks is given and Section 3 gives overview of the equation discovery system LAGRAMGE. The results of applying both techniques on five time series data sets are presented in Section 4. Finally, Section 5 concludes with a summary of the results and directions for further work.

2 Neural Networks

The time series $x(1), x(2), x(3) \ldots$, which describes the system is given. From the series we generate vectors $\mathbf{x(n)} = [x(n-1), x(n-2), \ldots, x(n-p)]^T$, which describe the last p values of the phenomenon until time $n-1$. We are trying to find a map $F(\mathbf{x(n)}) = \widehat{x}(n)$ such that the predicted value $\widehat{x}(n)$ in time n is the most similar to the original signal value $x(n)$ in time n. For accomplishment of F we use three different types of neural networks (Gershenfeld & Weigend 1992, Narendra 1990, Pham 1995).

2.1 Multilayer perceptron (MP)

We use a general multilayer feed-forward network (Lippmann 1987) whose learning algorithm is generalized δ-rule or back-propagation (BP). The user interface provides regulation of the following parameters: number of layers, number of neurons in each layer, learning rate η and momentum term α.

Parameters η and α could be changed during the training. Neurons in input layer act as buffers for distributing the input signals $\mathbf{x(n)}$ to neurons in the hidden layer (Haykin 1998, Lippmann 1987, Nielsen 1990, Pham 1995). MP is usually used as pattern recognition tool, but from a systems theoretic point of view it can be also used for approximation of non-linear maps (Narendra 1990).

2.2 FIR multilayer perceptron (FIR-MP)

In order to allow time to be represented by the effect it has on signal processing or to make the network to be dynamic, time delays are introduced into the synaptic structure of the network and their values are adjusted during the learning phase (Haykin 1998). In fact each synapse is represented by a finite-duration impulse response (FIR) filter (Figure 1).

FIR-MP network is a vector generalization of the MP and its learning algorithm is a vector generalization of the standard BP algorithm, called temporal BP (TBP). The basic form of TBP is non-causal because the computation of weights requires knowledge of future values of weights' changes δ-s and weights w-s. It could be made causal by adding a finite number of delay operators on the feedback connections so that only present and past values of δ-s and w-s are used. We hypothesize that by introducing tapped delay feedbacks the NN performance on problems involving time dependencies could be improved. In (Lin 1996) FIR-MP is compared to the NARX recurrent network by its computational power. NARX is computationally as strong as fully connected recurrent network thus is Turing machine equivalent (Siegelmann 1995, Siegelmann & Sontag 1995).

2.3 Recurrent network in real time (RN)

The net (Haykin 1998, Pham 1995) consists of connected input-output layers and processing layer. RN has ability to connect the external time-varying input with its previous output by using delay operator.

The learning algorithm used is real time recurrent learning (RTLL) (Haykin 1998), which is gradient-descent learning algorithm and minimizes the error function by changing the weights of all visible neurons. This architecture is capable of representation of arbitrary non-linear dynamical system and it is Turing equivalent (Alippi 1996, Siegelmann 1995, Siegelmann & Sontag 1995). However, learning simple behavior can be quite difficult by using gradient descent[1]. RTRL is not guaranteed to follow the negative gradient of the error function. This is a consequence of the feedback connection and it can be improved by slow changing of weights. Although RN has difficulty capturing the global behavior (Lin 1996) it is useful for learning short-term dependencies and thus can be used for short-term predictions.

[1] For example, even it is Turing equivalent, it has been difficult to get it successfully learn finite-state machines from example strings encoded as sequences.

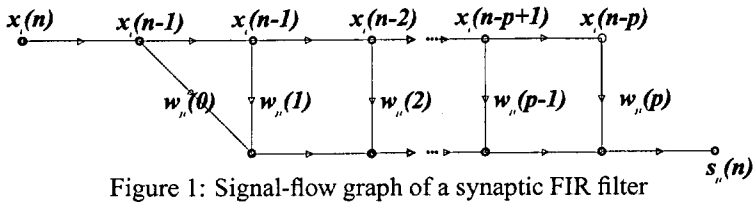


Figure 1: Signal-flow graph of a synaptic FIR filter

3 Equation Discovery

The problem of equation discovery, as addressed by LAGRAMGE, can be defined as follows.
Given are

- a context free grammar $G = (N, T, P, S)$ (see next section) and
- input data $D = (V, v_d, M)$, where
  - $V = \{v_1, v_2, \ldots v_m\}$ is a set of domain variables,
  - $v_d \in V$ is the dependent variable and
  - M is a set of one or more measurements. Each measurement is a table of measured values of the domain variables at successive time points:

time	v_1	v_2	...	v_m
t_0	$v_{1,0}$	$v_{2,0}$	...	$v_{m,0}$
t_1	$v_{1,1}$	$v_{2,1}$	...	$v_{m,1}$
t_2	$v_{1,2}$	$v_{2,2}$	...	$v_{m,2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_N	$v_{1,N}$	$v_{2,N}$	...	$v_{m,N}$

Find an equation for expressing the dependent variable v_d in terms of variables in V. This equation is expected to minimize the discrepancy between the measured and calculated values of the dependent variable. The equation can be:

- differential, i.e. of the form $\partial v_d/\partial t = \dot{v}_d = E$, or
- ordinary, i.e. of the form $v_d = E$,

where E is an expression that can be derived from the context free grammar G.

3.1 Restricting the space of possible equations

The syntax of the expressions on the right hand side of the equation is prescribed with a context free grammar (Hopcroft & Ullman 1979). A context free grammar contains a finite set of variables (also called nonterminals or syntactic categories) each of which represents expressions or phrases in a language (in equation discovery, nonterminals represent sets of expressions that can appear in the equations). The expressions represented by the nonterminals are described in terms of nonterminals and primitive symbols called terminals. The rules relating the nonterminals among themselves and to terminals are called productions.

The original motivation for the development of context free grammars was the description of natural languages. For example, a simple grammar for deriving sentences consists of the productions *sentence* $\rightarrow$ *noun verb*, *noun* $\rightarrow$ *network*, *noun* $\rightarrow$ *equation*, and *verb* $\rightarrow$ *predicts*. Here *sentence*, *noun* and *verb* are nonterminals, while words that actually appears in sentences (i.e. network, predicts) are terminals. The sentences *networkpredicts* and *equationpredicts* can be derived with this grammar.

We denote a context free grammar as a tuple $G = (N, T, P, S)$, where N and T are finite disjoint sets of nonterminals and terminals, respectively. P is a finite set of productions; each production is of the form $A \rightarrow \alpha$, where A is a nonterminal and α is a string of symbols From $N \cup T$. We use the notation $A \rightarrow \alpha_1 \mid \alpha_2 \mid \ldots \mid \alpha_k$ for a set of productions for the nonterminal A: $A \rightarrow \alpha_1, A \rightarrow \alpha_2, \ldots, A \rightarrow \alpha_k$. Finally, S is a special nonterminal called starting symbol.

Grammars used in equation discovery system LAGRAMGE have several symbols with special meanings. The terminal *const* $\in T$ is used to denote a constant parameter in an equation that has to be fitted to the input data. The terminals v_i are used to denote variables from the input domain D. Finally, the nonterminal $v \in N$ denotes any variable from the input domain. Productions connecting this nonterminal symbol to the terminals v_i are attached to v automatically, i.e., $\forall v_i \in V : v \rightarrow v_i \in P$.

The only restriction on the grammar G is that the right sides of the productions in P have to be expressions that are legal in the C programming language. This means that we can use all C built-in operators and functions in the grammar. Additional functions, representing background knowledge about the domain at hand can be used, as long as they are defined in conjunction with the grammar. Note that the derived equations may be non-linear in both the constant parameters and the system variables.

Expressions can be derived by grammar G from the nonterminal symbol S by applying productions from P. We start with the string w consisting of S only. At each step, we replace the leftmost nonterminal symbol A in string w with α, according to some production $A \rightarrow \alpha$ from P. When w consists solely of terminal symbols, the derivation process is over.

3.2 LAGRAMGE - the algorithm

Expressions generated by the context free grammar G contain one or more special terminal symbols *const*. A nonlinear fitting method is applied to determine the values of these parameters. The fitting method minimizes the value of the error function $Error(\mathbf{c})$, i.e. if $\mathbf{c}$ is the vector of constant parameters in expression E, then the result of the fitting algorithm is a vector of parameter values $\mathbf{c}^*$, such that $Error(\mathbf{c}^*) = \min_{\mathbf{c}\in R^{n_c}}\{Error(\mathbf{c})\}$. The error function *Error* is a sum of squared errors function, defined in the following manner:

- for a differential equation of the form $\partial v_d/\partial t = E$:
  $Error(\mathbf{c}) =$
  $\sum_{i=0}^{N}\left[v_{d,i} - \left(v_{d,0} + \int_{t_0}^{t_i} E(\mathbf{c}, v_1, \ldots v_m)\right)\right]^2$, and
- for an ordinary equation of the form $v_d = E$:
  $Error(\mathbf{c}) =$
  $\sum_{i=0}^{N}(v_{d,i} - E(\mathbf{c}, v_{1,i}, \ldots v_{d-1,i}, v_{d+1,i}, \ldots v_{m,i}))^2$,

where N is the size of the measurement table and $v_{j,i}$ the value of the system variable v_j at time t_i. Note that in the case of calculating the error function for differential equations we use the integral of the expression on the right hand side of the equation instead of the derivative of the dependent variable. This is done because the error of algorithms for numerical integration is in general smaller than the error involved of numerical derivation. We use a simple trapezoid formula for numerical integration with the same step size as the time step between successive measurements in the measurement table. The downhill simplex and Levenberg-Marquardt algorithms (Press et al. 1986) can be used to minimize the error function.

Furthermore, the value of a heuristic function for the expression is evaluated. It is equal to the sum of squared errors value *SSE* calculated by the fitting method ($SSE(E) = Error(\mathbf{c}^*)$). An alternative heuristic function *MDL* (minimal description length) can be used, that takes into account the length l of expression E:

$$MDL(E) = SSE(E) + \frac{l}{10 \cdot l_{max}}\sigma_{v_d},$$

where l_{max} is the length of the largest expression generated by the grammar and σ_{v_d} is the standard deviation of the dependent variable v_d. The length is measured as the number of terminals in the expression. The *MDL* heuristic function prefers shorter equations.

A context free grammar can in principle derive an infinite number of expressions (equations). LAGRAMGE thus uses a bound on the complexity (depth) of the derivation used to produce the equation (Todorovski & Džeroski 1997). The LAGRAMGE algorithm exhaustively or heuristically searches for the best equation (according to the selected heuristic function) within the allowed complexity (depth) limits.

3.3 Time series prediction with equation discovery

We reformulate the problem of time series prediction into the equation discovery problem in the following way. Given a time series $x(1), x(2), x(3), \ldots$, we choose a constant p and build matrix M as follows:

time	v_1	v_2	...	v_{p+1}
t_0	$x(1)$	$x(2)$	...	$x(p+1)$
t_1	$x(2)$	$x(3)$	...	$x(p+2)$
t_2	$x(3)$	$x(4)$	...	$x(p+3)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$

Now the input domain for equation discovery problem equivalent to the problem of time series prediction is $D = (\{v_1, v_2, \ldots, v_{p+1}\}, v_{p+1}, M)$. We search for ordinary equation of the form $v_{p+1} = F(v_1, v_2, \ldots v_p)$. The obtained equation can be interpreted as difference equation for predicting the next value of the time series $\widehat{x}(n) = F(x(n-1), x(n-2), \ldots, x(n-p))$.

The form of function F on the right-hand side of the equation is biased with a context free grammar G. We used three different context free grammars for restricting the space of possible equation in time series prediction domains. First grammar is used to produce linear models:

$$E \rightarrow const \mid const * v \mid E + const * v$$

The second grammar generates quadratic multivariate polynomials:

$$\begin{aligned} E &\rightarrow const \mid const * F \mid E + const * F \\ F &\rightarrow v \mid v * v \end{aligned}$$

Finally, the third grammar used in the experiments generates piecewise linear models. The breakpoint is set to 0.5, which is the middle of the interval of normalized values of time series:

```
double If(double v, double e1, dou-
ble e2) {
        return((v < 0.5) ? e1 : e2);
}
```

$$\begin{aligned} IfE &\rightarrow E \mid \texttt{If}(v, E, E) \\ E &\rightarrow const \mid const * v \mid E + const * v \end{aligned}$$

4 Experiments

4.1 Data sets descriptions

We applied the techniques described in the previous two sections to two synthetic and three real world data sets:

Lorenz system Model of the Lorenz attractor is one of the most frequently used examples of the deterministic chaos system. It is described with the following differential equations:

$$\begin{aligned} \dot{x} &= \sigma(y-x) \\ \dot{y} &= x(R-z)-y \\ \dot{z} &= xy-bz \end{aligned}$$

The values of the constant parameters were chosen to be: $\sigma = 16.0, R = 451992, b = 4.0$. For initial state $x(0) = 0.06735, y(0) = 1.8841, z(0) = 15.7734$ the system is well-conditioned. The equations were simulated for 2000 time steps of length $h = 0.001$. For the prediction task we use the time series for variable z, because it clearly reflects non-linear dynamics of the system.

Reference voltage The observed time series are generated by solid-state VRE-s, based on 7V zener diodes LTZ1000 of a group DCVRS. They are produced by measuring the absolute voltage values of VRE, which is controlled by PC-computer. The PC communicates with DCVRS via serial port RS232. Measuring instrument is digital voltmeter HP3458A. The time series present 2000 samples taken in time intervals of 15 minutes during 500 hour measurement.

Fractional Brownian motions or $1/f$ noises FBM is a random function provided by Mandelbrot and Van Ness (FBM). The most important feature of FBM is that its increments $[B_H(t+T) - B_H(t)] = h^{-H}[B_H(t+hT) - B_H(t)]$ are stationary, statistically self-similar and have Gaussian distribution with a standard deviation $C_H T^H$, where C_H is constant. This is usually called T^H *law of FBM*. The parameter H is directly related to the fractal (Hausdorff) dimension D. For generation of the FBM signals we use the method of spectral synthesis (Nancovska 1997).

Lorenz-like chaos in NH_3-FIR lasers Far infrared lasers have been proposed as examples of a physical realization of the Lorenz model, mentioned earlier (Hubner et al. 1992). However, the actual laser systems are more complex then simple coherently coupled three-level systems. The data set was chosen to obtain several of the important quantities pertinent in comparison to the parameters of numerical data sets obtained by the integration of Lorenz equations. We took into consideration first 2000 time points of the time series.

Sunspots The data set is standard benchmark test for various techniques for time series prediction. It contains the observation of the number of annual sunspots for 280 years.

4.2 Experimental setting

The following methodology of the experiments with the time series prediction data sets was used. Each data set was divided in two parts of equal sizes (1000 time points per set, expect for the Sunspot data set which has only 280 time points). The first part was used as input for the learning system in the training phase, and the second one was used for testing the performance of the obtained predictive model. Furthermore, in the training phase 80% of the training set was used directly for learning and the rest is used for error evaluation only (the evaluation applies the estimation of the performance of the predictor learned so far). The length of the input vector p varied between 1 and 6. The criterion for choosing the best predictor was the root mean square error (RMSE).

In the experiments with neural networks, the value of parameter η is gradually decreased from 0.95 to 0.003 to avoid local minima of the error surface. When training the recurrent network η is set to the lower value from the beginning to make the time scale of the weight changes small enough to allow the learning algorithm to follow the negative gradient of error function.

We used beam search strategy in equation discovery system LAGRAMGE with beam width set to 50 and both heuristic functions (SSE and MDL) with downhill simplex method for constant parameters fitting. The depth complexity parameter was set to 10 and three different context free grammars (from the previous section) were used. The best equation was then chosen that minimizes the RMSE on the test training set.

4.3 Results

The results of the experiments for neural networks and equation discovery system LAGRAMGE are given in Table 1 and Table 2, respectively. The architecture of the MP neural network is represented with $x - y - z$, where x, y and z denote numbers of neurons in first, second and third layer, respectively. $1^p y^q 1$ denotes FIR-MP neural network architecture with p and q time operators (taps) between corresponding layers and p equals the length of the input vector. The architecture of the recurrent neural network is represented with $x \leftarrow y - 1$, where x denotes the length of input vector and y the number of feedback connections.

Recurrent neural networks has the best performance for the Reference voltage and FBM data sets. Both data sets represent time series with very fast changing values without long-term trend. The recurrent neural network has worst performance for the time series with trend. In that case the MP and FIR-MP networks better identify the underlying system, as we can see from the results of the experiments for Lorenz and Sunspots data sets. For Lorenz-like chaos data set the best performance is surprisingly achieved with MP network[2]. For all data sets, expect the Lorenz-like, the prediction performance of different types of neural networks are comparable. In the experiments with Lorenz-like chaos MP is significantly better then other two types.

[2]Finding a simple representation for a complex signal might require looking for relationship among input variables. In the case of Lorenz-like chaos input vectors are representative enough to allow the MP to find the "simple" regression model which is good enough for local description of the model.

Data set	Winning NN			RMSE	
	Type	Architecture	p	Training	Testing
Lorenz	FIR-MP	$1^{5}4^{2}1$	5	$9.9 \cdot 10^{-4}$	$1.7 \cdot 10^{-2}$
Ref. voltage	Rec.	$6 \hookleftarrow 4-1$	6	0.0646	0.0749
FBM	Rec.	$5 \hookleftarrow 4-1$	5	0.0895	0.0823
Lorenz-like	MP	$6-3-1$	6	0.0153	0.0260
Sunspots	FIR-MP	$1^{5}4^{4}1$	5	0.09795	-

Table 1: Results of the experiments with neural networks

Data set	Winning equation		RMSE	
	Type	p	Training	Testing
Lorenz	piecewise linear	6	$1.919 \cdot 10^{-6}$	$2.465 \cdot 10^{-6}$
Ref. voltage	piecewise linear	6	0.06375	0.07405
FBM	linear	6	0.08846	0.0827
Lorenz-like	quadratic	3	0.03889	0.05939
Sunspots	quadratic	4	0.103	-

Table 2: Results of the experiments with equation discovery

In the experiments with equation discovery for the Lornez, Reference voltage and FBM data sets the discovered equations are linear. Although the Lorenz data set is obtained with simulating three non-linear differential equations, the interval enclosed in the data set do not expose the non-linearity of the underlying equations (due to the stability of the numerical integration a small time step was chosen). For Lorenz-like and Sunspots data sets quadratic equations were discovered, which was expected because of their non-linearity. The parameter p (number of previous values used for prediction) is significantly smaller in cases where quadratic equations were discovered. As in the experiments with neural networks, for all data sets, expect the Lorenz-like, the prediction performances of different types of equations are comparable. In the experiments with Lorenz-like chaos quadratic equations are significantly better then other two types.

Both methods manifest comparable performance on three data sets. Equation discovery outperforms neural networks for the Lorenz data set, which was expected because of the determinism of the underlying model. Neural networks have better performance on the Lorenz-like and Sunspots data sets where quadratic equations were discovered by LAGRAMGE.

Figure 2 shows the performance of the obtained predictors for different types of neural networks and equations.

5 Discussion

In the paper, we presented the equation discovery system LAGRAMGE that uses context free grammars for restricting the hypothesis space of equations. Background knowledge from the domain of use in the form of function definitions can be used along with common arithmetical operators and functions built in the C programming language. The hypothesis space of LAGRAMGE is a set of equations, such that the expressions on their right hand sides can be derived from a given context free grammar.

In contrast with system identification methods, where the structure of the model has to be provided explicitly by the human expert, LAGRAMGE can use a more sophisticated form of representing the expert's theoretical knowledge about the domain at hand. A context free grammar can be used to specify a whole range of possible equation structures that make sense from the expert's point of view. Therefore, the discovered equations are in comprehensible form and can give domain experts better or even new insight into the measured data. This also distinguishes LAGRAMGE from other system identification methods like neural networks, which can be used for obtaining black-box models, i.e., models with incomprehensible structure.

On the equation discovery side, the presented work is related to equation discovery systems, such as BACON (Langley et al. 1987), EF (Zembowitz & Zytkow 1992), E* (Schaffer 1993), LAGRANGE (Dzeroski & Todorovski 1993) and GOLDHORN (Krizman et al. 1995). However, none of them was applied to the task of time series prediction.

Various architectures of neural networks have already been used for system identification and prediction. Some of them are closely related to the architectures used in this paper (Haykin 1998, Lippmann 1987, Narendra 1990, Pham 1995), and others are different, such as radial basis function (RBF) neural network (Haykin 1998) and Group-Method-of-Data-Handling (Pham 1995). The NARX recurrent neural networks (Alippi 1996, Lin 1996) architecture is suitable for learning long-term dependencies in time series. For the task of short-term prediction, addressed in this paper, learning the local structure is good enough (Narendra

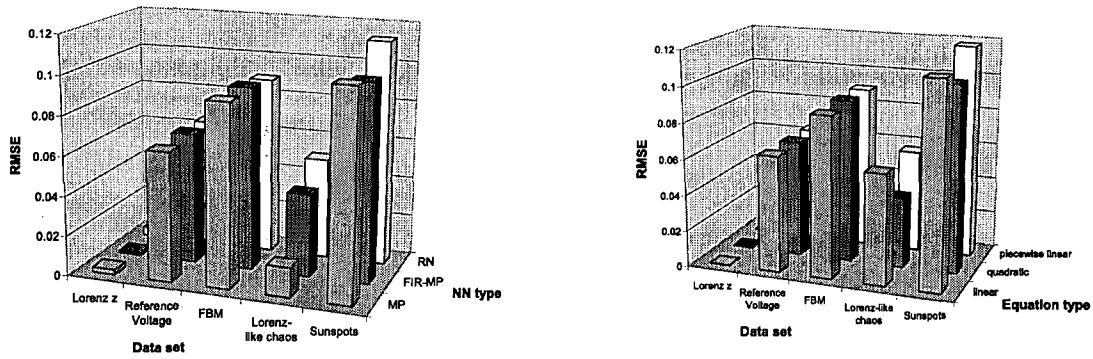


Figure 2: RMSE of the predictors for different types of neural networks and equations

1990). Support vector machine (SVM) for non-linear regression (Haykin 1998), which is approximate implementation of the method of structural risk minimization, could be also used. Finally, SVM may be implemented in the form of a polynomial learning machine, RBF network or MP.

The outcomes of the experiments confirm the potential applicability of both paradigms for time series prediction. For each domain, the performances of the predictors, obtained with both methods, are comparable. The best predictors were obtained for the deterministic Lorenz-z time series. The efficiency of the predictors for real world domains (Reference voltage and Lorenz-like chaos) are comparable. The predictors with worst efficiency were obtained for FBM and Sunspots time series, due to the randomness of the underlying model (FBM) and the small number of measurements available (Sunspots).

The predictive models could be used as a segment of software controlled voltage reference element (VRE), which consists of three main parts: measuring, predictive and control part. Stability of a reference voltage source could be enhanced by implementation of voltage control, which includes a function of correction (feedback loop). This could be done by correction of the current voltage by using a prediction based on measurements made before. It is anticipated for a solid-state voltage reference source to achieve the stability better than 1ppm/1000h (Nancovska 1997).

First step towards the further work will be the comparison of the methods described in the paper with mainstream statistical methods for time series prediction, such as ARIMA or exponential smoothing. It will be of great interest to apply these methods to the variety of data sets from different domains, where some background knowledge from the concrete domain can be used for restricting the equation space. Alternative types of neural networks architectures (NARX recurrent network, SVM for non-linear regression (Haykin 1998)) could be also implemented.

References

[1] Alippi, C., and Piuri, V. (1996) Experimental Neural Network for Prediction and Identification, In *IEEE Transactions on Instrumentation and Measurement*, Vol. 45, No. 2, 1996, pages 670–676.

[2] Džeroski, S., and Todorovski, Lj. (1993) Discovering dynamics In *Proc. Tenth International Conference on Machine Learning*, pages 97–103. Morgan Kaufmann, San Mateo, CA.

[3] Gershenfeld, N. A., and Weigend, A. S. (1992) The future of time series: learning and understanding. In *Time series prediction: Forecasting the future and understanding the past. Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis held in Santa Fe, New Mexico, May 14-17, 1992*, pages 1–70. Addison-Wesley Publishing Company, Reading, MA.

[4] Haykin, S. (1998) *Neural Network - A Comprehensive Foundation*, Second Edition, Macmillan College Publishing Company, Inc.

[5] Hopcroft J. E., and Ullman, J. D. (1979) *Introduction to automata theory, languages, and computation.* Addison-Wesley, Reading, MA.

[6] Hübner, U., and Weiss, C. O., and Abraham, N. B., and Dingyuan T. (1992) Lorenz-like chaos in NH_3-FIR lasers (Data Set A). In *Time series prediction: Forecasting the future and understanding the past. Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis held in Santa Fe, New Mexico, May 14-17, 1992*, pages 1–70. Addison-Wesley Publishing Company, Reading, MA.

[7] Hecht-Nielsen R. (1990) Introduction to backpropagation, In *Neurocomputing*, HNC, Inc. and University of California, San Diego, Addison-Wesley Publishing Company.

[8] Križman, V., and Džeroski, S., and Kompare, B. (1995) Discovering dynamics from measured data. *Electrotechnical Review*, 62: 191–198.

[9] Langley, P., and Simon, H., and Bradshaw, G. (1987) Heuristics for empirical discovery. In Bolc, L., editor, *Computational Models of Learning*. Springer, Berlin.

[10] Lin, T., and Horne, B. G., and Tino, P., and Giles, C. L. (1996) Learning Long-Term Dependences in NARX Recurrent Neural Networks. In *IEEE Transactions on Neural Networks*, Vol. 7, No. 6, November 1996, pages 1329 – 1338.

[11] Lippmann, R. P., (1987) An Introduction to Computing with Neural Nets, In *IEEE ASSP Magazine*, April 1987, pages 4–22.

[12] Mandelbrot, B., and Van Ness, J. W., (1968) Fractional Brownian Noises and Applications, In *SIAM* Rev. 10 (4), 1968, pages 422–436.

[13] Nančovska, I., and Kranjec, P., and Fefer, D., and Jeglič, D. (1998) Case Study of the Predictive Models Used for Improvement of the Stability of the DC Voltage Reference Source, In *IEEE Transactions on Instrumentation and Measurement*, vol.47, no. 6, 1998, pages 1487 - 1491.

[14] Narendra, K. S., and Parthasarathy, K. (1990) Identification and Control of Dynamical Systems Using Neural Networks. In *IEEE Transactions on Neural Networks*, Vol. 1, No. 1., March 1990, pages 4 – 27.

[15] Siegelmann, H. T., and Horne, B. G., and Giles, C. L. (1995) Computational capabilities of recurrent NARX neural network, In Tech. Rep. UMIACS-TR-95-12 nad CS-TR-3408, Inst. of Adv. Comp. Stud., Univ. of Maryland.

[16] Siegelmann, H. T., and Sontag E. D. (1995) On the Computational Power on Neural Networks. In *Journal of Comp. Systems in Science*, Vol. 50, No. 1, pages 132 – 150, 1995.

[17] Pham, D.T., and Liu, X. (1995) *Neural Networks for Identification, Prediction and Control*. Springer-Verlag, London, GB.

[18] Press, W. H., and Flannery, B. P., and Teukolsky, S. A., and Vetterling, W. T. (1986) *Numerical Recipes*. Cambridge University Press, Cambridge, MA.

[19] Schaffer, C. (1993) Bivariate scientific function finding in a sampled, real-data testbed. *Machine Learning*, 12: 167–183.

[20] Todorovski, Lj., and Džeroski, S. (1997) Declarative bias in equation discovery. In *Machine learning. Proceedings of the 14th international conference (ICML'97)*, pages 376–384. Morgan Kaufmann publishers, San Francisco, CA.

[21] Zembowitz, R., and Żytkow, J. (1992) Discovery of equations: experimental evaluation of convergence. In *Proc. Tenth National Conference on Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA.

Adaptive On-line ANN Learning Algorithm and Application to Identification of Non-linear Systems

Daohang Sha and Vladimir B. Bajić
Centre for Engineering Research, Technikon Natal, P.O.Box 953, Durban 4000, South Africa
dhsha@hotmail.com, bajic.v@umfolozi.ntech.ac.za
tel/fax: +27-31-2042560, http://nsys.ntech.ac.za

Keywords: Soft computing, neural networks, gradient descent method, real-time algorithms, Input/Output modelling

Edited by: Cene Bavec and Matjaž Gams
Received: October 13, 1999 **Revised:** November 10, 1999 **Accepted:** December 3, 1999

A new on-line adaptive learning rate algorithm for I/O identification based on two ANNs is proposed. The algorithm is derived from the convergence analysis of the conventional gradient descent method. Simulation experiments are given to illustrate the advantages of the proposed algorithm in its application to an identification problem of some non-linear dynamic systems.

1 Introduction

Feedback linearization can be used for controlling a broad class of non-linear processes. To perform feedback linearization it is necessary that the process allows description by a particular model structure which fits into the affine non-linear form (see [1]-[4]). When the process is unknown we can let two multilayer perceptron (MLP) models approximate the linear relationship between the input and output data of the process. In this paper, we will develop an on-line variable rate training algorithm for this double neural network system.

Determination of the fixed learning rate of the conventional back-propagation (BP) algorithm for feedforeward neural networks (FNNs) [5] has to be made with care. If the learning rate is large, learning may occur quickly, but it may also become unstable. To ensure stable learning, the learning rate must be sufficiently small. However, with a small learning rate, an ANN, for example an MLP, may adapt reliably, but it may take quite a long time. It is thus difficult to select a suitable fixed learning rate for different initial values of the parameters of the ANNs and for different structures that ANNs may have. Moreover, a good fixed learning rate for one system is not necessarily good for another. These are characteristics of the basic ANN learning rule that relies on the gradient descent (GD) method and the chain rule [6]. For the convergence of such algorithms see [7] and [8]. The GD method is known for its slowness and its tendency to become trapped in local minima. To reduce these shortcomings, a number of faster ANN training algorithms have been developed, such as different adaptive learning algorithms (see [9], [10] and [11]), and other modified algorithms (see [12]-[23]). In spite of their improved convergence, these methods are not based on the optimal instantaneous learning rates of the GD approach. One may use second-order nonlinear optimizing methods to accelerate the MLP learning, such as the conjugate gradient algorithm (see [24]) or the Levenberg-Marquardt based method (see [25]). The critical drawbacks of these methods, however, are the ill conditioning of the Hessian matrix in many applications and the computational complexity related to the Hessian. In addition, most of these algorithms are developed only for off-line training of the ANNs.

Recently, the layer-by-layer optimizing algorithm was proposed in which each layer of MLP's is decomposed into both a linear part and a nonlinear part [26]-[33]. The linear part of each layer is solved via the least squares problem formulation. Although these algorithms show fast convergence with much less computational complexity than those of the conjugate gradient or Newton methods, they result in an unavoidable problem caused by target assignments at hidden nodes. When the targets for a hidden layer cannot be linearly separated, it is impossible to reduce the MSE sufficiently at both the hidden layer and the output layer.

Another class of fast learning algorithms is the one based on the extended Kalman filter (EKF) technique for training of a multilayer FNN. It has received considerable attention recently (see [34]-[38]). These algorithm improved the convergence rate considerably and exhibited good performance. However, their numerical stability is not guaranteed. This may degrade learning convergence, increase training time and, generally, can make on-line implementation questionable.

In this paper, we develop an on-line variable rate learning algorithm for double neural network system which can speed up the learning process substantially and can simultaneously provide stability of the learning process. In [39], we proposed a variable rate algorithm for the on-line training of an MLP. Here we extend this solution to on-line training of a double neural network system for I/O modelling of SISO time-invariant non-linear systems. As ANNs are widely used for the identification of non-linear

systems (see [40]-[61]), as well as for prediction of their behavior (see [62]-[64]), we will test this algorithm in the on-line identification and prediction of behavior of three non-linear systems.

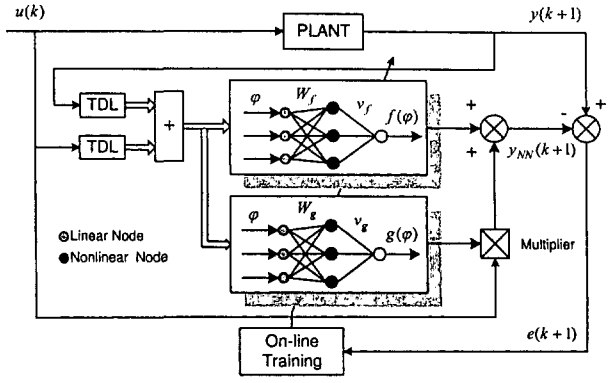


Figure 1: Double neural network system for identification of non-linear plant

2 Modeling Non-linear Plants by ANNs

2.1 Problem Description

Consider a SISO time-invariant non-linear system for which we will attempt to obtain an I/O model in the form of

$$y(k+1) = f\left[\varphi(k)\right] + g\left[\varphi(k)\right] \cdot u(k).$$

Here $\varphi(k) = [-y(k) \ldots - y(k-n+1)\ u(k-1) \ldots u(k-m)]^T$. Integer parameter n may be associated with the system order; m is a non-negative integer. Let the functions f and g of the above model be unknown. We will use two neural networks to model f and g, in order to obtain their approximations $\widehat{f}$ and $\widehat{g}$, respectively. The assumption is that these networks are governed by

$$\begin{aligned} & y_{NN}(k+1|\varphi,\theta) \\ = \ & \widehat{f}_{NN}(\varphi,\theta_f) + \widehat{g}_{NN}(\varphi,\theta_g) \cdot u(k) \\ = \ & v_f^T S(W_f\varphi) + v_g^T S(W_g\varphi) \cdot u(k), \end{aligned} \tag{1}$$

where y_{NN} is the cumulative output of neural networks (see Fig.1), $\theta_k = [W_k, v_k]$, $k = f, g$, are the parameter vectors, W_f, W_g, v_f, v_g, are matrices of weights (W_k) and vectors of biases (v_k) from 'input-to-hidden' layer and 'hidden-to-output' layer, respectively, for ANNs defining $\widehat{f}$ and $\widehat{g}$. The matrix S will be defined later. In what follows a^T denotes the transpose of a matrix or a vector a, while a' stands for a partial derivative of a.

2.2 Model of Neural Networks

Let us assume that for the networks associated with $\widehat{f}$ and $\widehat{g}$, the number of nodes of the input layer is denoted by IN_k, $k = f, g$, and HN_k, $k = f, g$, denote the number of nodes of the hidden layer. Then, starting from (1) one can consider a double three-layers-forward neural network as an identification model for a non-linear plant (Fig.1), where the network model is governed by

$$\begin{aligned} & y_{NN}(k+1) \\ = \ & \sum_{i=1}^{HN_f} s\left(\sum_{j=1}^{IN_f} w_{f,ij}\varphi_{f,j} + w_{f,i0}\right) v_{f,i} \\ & + v_{f,0} + u(k) \\ & \cdot \left\{\sum_{i=1}^{HN_g} s\left(\sum_{j=1}^{IN_g} w_{g,ij}\varphi_{g,j} + w_{g,i0}\right) v_{g,i} + v_{g,0}\right\} \\ = \ & \sum_{i=0}^{HN_f} s\left(\sum_{j=0}^{IN_f} w_{f,ij}\varphi_{f,j}\right) v_{f,i} \\ & + u(k) \sum_{i=0}^{HN_g} s\left(\sum_{j=0}^{IN_g} w_{g,ij}\varphi_{g,j}\right) v_{g,i} \\ = \ & \sum_{i=0}^{HN_f} s\left(net_{f,i}\right) v_{f,i} \\ & + u(k) \sum_{i=0}^{HN_g} s\left(net_{g,i}\right) v_{g,i} \end{aligned} \tag{2}$$

where $net_{k,i} = \sum_{j=0}^{IN_k} w_{k,ij}\varphi_{k,j}$ is the output of i-th hidden node with $\varphi_{k,0} \equiv 1$, for $i = 0, 1, ..., HN_k$, and with $s\left(net_{k,0}\right) \equiv 1$. Here $\varphi_{k,j}$, $j = 1, ..., IN_k$, are the inputs of the neural network, $w_{k,ij}$, $i = 1, ..., HN_k$, $j = 1, ..., IN_k$, are the weights from input layer to hidden layer, $w_{k,i0}$, $i = 1, ..., HN_k$, are the biases of the hidden nodes, $v_{k,i}$, $i = 1, ..., HN_k$, are the weights from the hidden layer to the output layer, $v_{k,0}$ is the bias of the output node, s is the activation function of nodes for the hidden layer. The subscript $k = f, g$. Further, (2) can be rewritten as

$$y_{NN}(k+1) = v_f^T S\left(W_f\varphi_f\right) + v_g^T S\left(W_g\varphi_g\right) u(k)$$

where for $k = f, g$, one has

$$\varphi_k = \left[\ \varphi_{k,0} \quad \varphi_{k,1} \quad \cdots \quad \varphi_{k,IN_k}\ \right]^T \in R^{(IN_k+1)\times 1},$$

$$v_k = \left[\ v_{k,0} \quad v_{k,1} \quad \cdots \quad v_{k,HN_k}\ \right]^T \in R^{(HN_k+1)\times 1},$$

$$W_k = \begin{bmatrix} w_{k,10} & w_{k,11} & \cdots & w_{k,1IN_k} \\ w_{k,20} & w_{k,21} & \cdots & w_{k,2IN_k} \\ \cdots & \cdots & \cdots & \cdots \\ w_{k,HN_k0} & w_{k,HN_k1} & \cdots & w_{k,HN_kIN_k} \end{bmatrix} \in R^{HN_k\times(IN_k+1)},$$

$$\begin{aligned} & S\left(W_k\varphi_k\right) \\ = \ & \left[\ s(net_{k,0}) \quad s(net_{k,1}) \quad \cdots \quad s(net_{k,HN_k})\ \right]^T \\ \in \ & R^{(HN_k+1)\times 1}. \end{aligned}$$

the activation function for non-linear nodes is a symmetric hyperbolic tangent function, i.e. $s(x) = \tanh(\mu_0^{-1}x)$, and its derivative is $s'(x) = \mu_0^{-1}[1 - s^2(x)]$, where μ_0 is the shape factor of the activation function.

3 Derivation of an On-line Learning Algorithm

3.1 GD Method

We define the error function as

$$\begin{aligned} & J(k) = \frac{1}{2}e^2(k) = \frac{1}{2}[y(k) - y_{NN}(k)]^2 \\ = \; & \frac{1}{2}[y(k) - v_f^T S\left(W_f\varphi_f\right) - v_g^T S\left(W_g\varphi_g\right)u(k)]^2 \end{aligned}$$

Applying the GD method to J and using Lemmas 2 and 3 from [39] one obtains increments for the parameters of ANNs as

$$\begin{aligned} \Delta v_f &= -\eta\frac{\partial J(k)}{\partial v_f(k)} = \eta\frac{\partial\left[v_f^T S(W_f\varphi_f)\right]}{\partial v_f(k)}e(k) \\ &= \eta S(W_f\varphi_f)e(k), \end{aligned} \tag{3}$$

$$\begin{aligned} \Delta W_f &= -\eta\frac{\partial J(k)}{\partial W_f(k)} = \eta\frac{\partial\left[v_f^T S(W_f\varphi_f)\right]}{\partial W_f}e(k) \\ &= \eta S'\left(W_f\varphi_f\right)v_f\varphi_f^T e(k), \end{aligned} \tag{4}$$

$$\begin{aligned} \Delta v_g &= -\eta\frac{\partial J(k)}{\partial v_g(k)} = \eta\frac{\partial\left[v_g^T S(W_g\varphi_g)\right]}{\partial v_g(k)}u(k)e(k) \\ &= \eta S(W_g\varphi_g)u(k)e(k), \end{aligned} \tag{5}$$

$$\begin{aligned} \Delta W_g &= -\eta\frac{\partial J(k)}{\partial W_g(k)} \\ &= \eta\frac{\partial\left[v_g^T S(W_g\varphi_g)\right]}{\partial W_g}u(k)e(k) \\ &= \eta S'\left(W_g\varphi_g\right)v_g\varphi_g^T u(k)e(k), \end{aligned} \tag{6}$$

where η is a small positive constant that represents the fixed learning rate.

3.2 Analysis of the GD Algorithm

Consider the error equation

$$\begin{aligned} & e(k+1) - e(k) \\ = \; & [y(k+1) - y_{NN}(k+1)] - [y(k) - y_{NN}(k)]. \end{aligned}$$

Let $\Delta y(k+1) = y(k+1) - y(k)$ and $\Delta y_{NN}(k+1) = y_{NN}(k+1) - y_{NN}(k)$. Let us assume that $|\Delta y(k+1)| \ll |\Delta y_{NN}(k+1)|$, i.e. that the change in the plant output $y(k)$ of the controlled system is sufficiently slower than the change in the output $y_{NN}(k)$ of the neural network. This assumption is realistic for many processes. Then, during the training process of neural network, the error equation is

$$\begin{aligned} e(k+1) - e(k) &= \Delta y(k+1) - \Delta y_{NN}(k+1) \\ &\approx -\Delta y_{NN}(k+1) \\ &= -[\Delta\widehat{f}_{NN} + \Delta\widehat{g}_{NN}\cdot u(k)] \end{aligned}$$

Applying Lemma 4 from [39], the above equation can be rewritten as follows

$$\begin{aligned} & e(k+1) - e(k) \\ \approx \; & -\left[S^T(W_f\varphi_f)\Delta v_f + v_f^T S'(W_f\varphi_f)\Delta W_f\varphi_f\right] \\ & -\left[S^T(W_g\varphi_g)\Delta v_g\right. \\ & \left.+ v_g^T S'(W_g\varphi)\Delta W_g\varphi_g\right]u(k). \end{aligned}$$

After substitution of (3)-(6) into the above equation, and, after some algebraic manipulations, one gets

$$e(k+1) - e(k) \approx -\eta\zeta(k)e(k),$$

i.e.

$$e(k+1) \approx [1 - \eta\zeta(k)]e(k),$$

where

$$\begin{aligned} \zeta(k) = \; & \left[S^T(W_f\varphi_f)S(W_f\varphi_f)\right. \\ & \left.+ v_f^T S'(W_f\varphi_f)S'\left(W_f\varphi_f\right)v_f\varphi_f^T\varphi_f\right] \\ & +\left[S^T(W_g\varphi_g)S(W_g\varphi_g)\right. \\ & \left.+ v_g^T S'(W_g\varphi_g)S'(W_g\varphi_g)v_g\varphi_g^T\varphi_g\right]u^2(k). \end{aligned}$$

Our intention is to make $\lim_{k\to\infty} e(k) \to 0$ as the number of iterations k increases. For this the condition $|1 - \eta\zeta(k)| < 1$ has to be satisfied, i.e. $0 < \eta < 2\zeta^{-1}(k)$. Apparently, the upper bound $2\zeta^{-1}(k)$ of the learning rate η is variable because the value of $\zeta(k)$ depends on the input φ_j and the current values of the neural network parameters v_j and W_j, for $j = f, g$.

3.3 Variable Learning Rate Algorithm

To obtain the variable learning rate algorithm we consider the case when the fastest learning occurs. This will be when $\eta = \zeta^{-1}(k)$, i.e. it will imply $e(k+1) \approx 0$. Substituting $\eta = \zeta^{-1}(k)$ into (3)–(6) one obtains the adaptive learning rate back-propagation increments of ANN parameters, i.e.

$$\begin{aligned} \Delta v_f &= \frac{1}{\zeta(k)}S(W_f\varphi_f)e(k), \\ \Delta W_f &= \frac{1}{\zeta(k)}S'\left(W\varphi_f\right)v_f\varphi_f^T e(k), \\ \Delta v_g &= \frac{1}{\zeta(k)}S(W_g\varphi_g)u(k)e(k), \\ \Delta W_g &= \frac{1}{\zeta(k)}S'\left(W_g\varphi_g\right)v_g\varphi_g^T u(k)e(k). \end{aligned}$$

We use these formulae in the on-line algorithm in the following simulation experiments.

4 Simulation Experiments

Extensive simulation studies were carried out with several examples of nonlinear dynamic systems which were used in [40], [65]. Two typical sets of results are illustrated in the following examples.

Example 1

We will apply the new algorithm to a spring-mass-damper system with a hardening spring governed by

$$y''(t) + y'(t) + y(t) + y^3(t) + F_L = u(t),$$

and we will compare its performance with the fixed learning rate algorithm. Let us use a network with 5 hidden units for approximating f, and a network with 4 hidden units for approximating g. We will use $\varphi(k) = [-y(k) \quad -y(k-1) \quad u(k-1)] = \varphi_f(k) = \varphi_g(k)$. The input $u(t)$ for both the non-linear system and the neural network is taken as a random input band-limited white noise signal. The target for the neural network is the output y of the system. The sampling interval is $T_s = 0.1$ sec. The total simulation time is 300 sec, i.e. we will have 3000 iteration steps. The neural network is trained during the first 2500 iterations in the on-line mode. After training at the time instant of 250 sec, the trained neural network is used to predict the output of the non-linear system during the subsequent 500 iterations. This is done only for the purpose of assessing the quality of the training process. The adaptive learning rate algorithm proposed in this paper and the algorithm with the fixed learning rates of $\eta = 0.001, 0.002, 0.004$, are compared. The time evolution of the training error for these four cases are shown in Fig.2. The training data of the neural network with adaptive learning rate and with the fixed learning rate of $\eta = 0.002$ are depicted in Fig.3 and 4, respectively.

As mentioned previously, the fixed learning rate for on-line training must be chosen with some care. Unlike the situation when training is done off-line, in the on-line situation it is not known which input vectors will be presented to the network. Therefore the learning rate should be fairly conservative so as to assure stable learning. However, if it is too small, the learning will take a long time (as for $\eta = 0.001$). If the learning rate is large, learning occurs quickly, but if it is too large learning can become unstable and errors may even increase (as for $\eta = 0.004$). To reduce these problems the fixed learning rate must be set to a suitable value. But, in general, the different initial values of weights and different processes will require different optimal fixed learning rates. It is thus impossible to find the best one for all of the cases. These problems do not exist with the adaptive learning rate algorithm. It can be observed from the simulation results that the adaptive rate

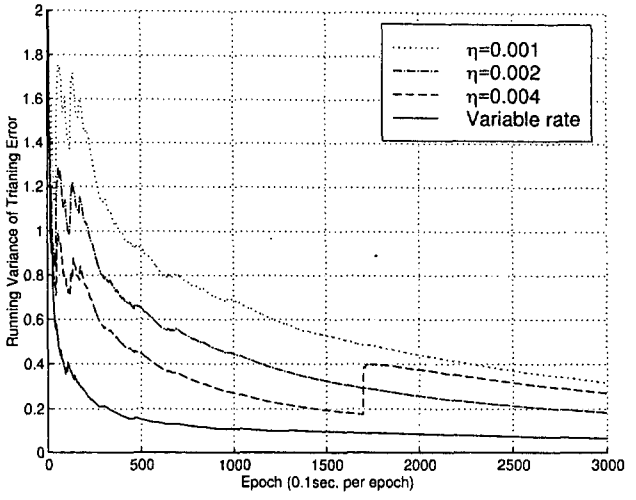


Figure 2: Time evolution of training error

algorithm is far better than the fixed rate algorithm in terms of both the learning speed and the training error.

Example 2

Two plants considered are governed by

$$y(k+1) = 0.3y(k) + 0.6y(k-1) + f[u(k)], \qquad (7)$$

where $y(k)$ and $u(k)$ are the output and input, respectively, at time k, and the function f is assumed unknown to the ANNs. For the purpose of plant simulation, it is taken in the form

$$f(u) = \begin{cases} f_1(u) \\ f_2(u) \end{cases}$$
$$= \begin{cases} 0.6\sin(\pi u) + 0.3\sin(3\pi u) + 0.1\sin(5\pi u), \\ 0.8\sin(2y(k)) + 1.2u(k). \end{cases}$$

The first system (A) is defined by (7) having $f(u) = f_1(u)$, and the other system (B) is defined by (7) with $f(u) = f_2(u)$. The MLPs used for identification and prediction in both cases have the same structure as in Example 1. The input to the plant is a sinusoid defined as

$$u(k) = \sin(\frac{2\pi k}{250}),$$

when $k \leq 500$, and

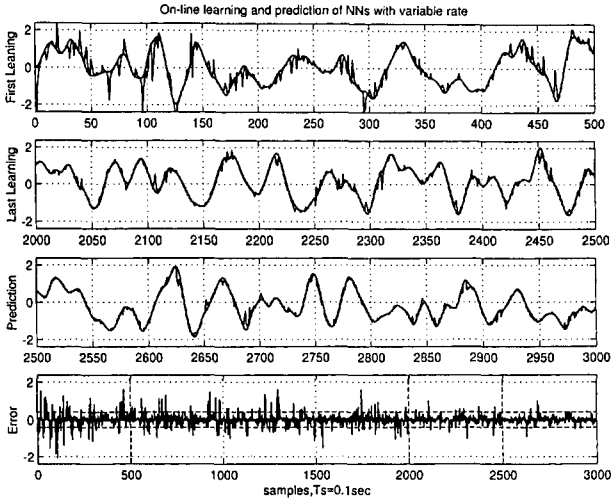


Figure 3: On-line learning and prediction with adaptive learning rate

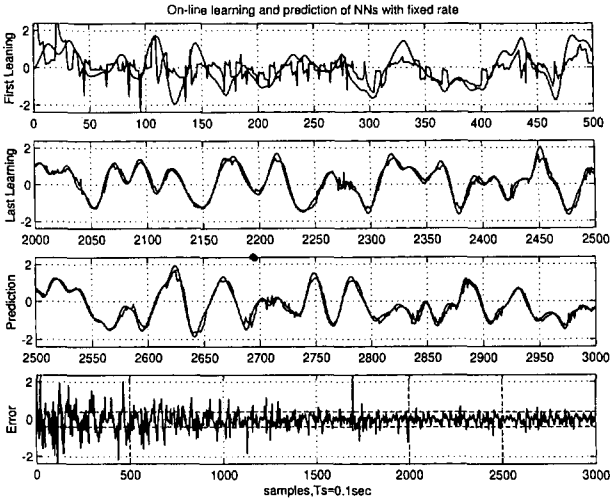


Figure 4: On-line learning and prediction with fixed learning rate $\eta = 0.002$

$$u(k) = 0.5\sin(\frac{2\pi k}{250}) + 0.5\sin(\frac{2\pi k}{25}),$$

when $k > 500$. In the case of system A, the results of on-line identification with fixed and variable learning rates are shown in Fig.5 and Fig.6, and the time evolution of the training error in Fig.7, respectively. Note that we selected by trial and error (experimentally) a fixed learning rate of $\eta = 0.34$ to satisfy both the learning speed and learning stability requirements. The variable rate learning algorithm achieves similar or better results directly, without any requirements for tuning the learning process.

However, for system B, if the fixed learning algorithm is used and the learning rate η is kept the same, i.e. $\eta = 0.34$, then the learning process will become unstable as can be observed from Fig.8 and 10. At the same time, the variable learning rate algorithm retains its good behavior (see Fig.9 and Fig.10). These examples serve to illustrate the convenience of the variable learning rate algorithm and its suitability for real-time identification.

5 Conclusion

Based on the analysis of the convergence of the gradient descent method, a new on-line adaptive rate learning algorithm for I/O modelling by a double ANN system is proposed. Compared to a fixed rate learning algorithm, the adaptive rate learning algorithm can achieve a much better performance in terms of the learning speed and the training error.

References

[1] E. B. Kosmatopoulos and P. A. Ioannou, A Switching Adaptive Controller for Feedback Linearizable Systems, *IEEE Transactions on Automatic Control*, Vol. 44, No. 4, pp. 742-750, 1999.

[2] A. Yesidirek, F. L. Lewis, Feedback linearization using neural network, *Automatica*, Vol. 31 No.11, pp.1659-1664, 1995.

[3] G. A. Rovithakis, M. A. Christodoulou, Neural Adaptive Regulation of Unknown Nonlinear Dynamical Systems, *IEEE Transaction on Systems, Man, And Cybernetics*, Part B: Cybernetics, Vol. 27, No. 5, pp. 810-822, 1997.

[4] K. Nam, Stabilization of Feedback Linearizable Systems Using a Radial Basis Function Network, *IEEE*

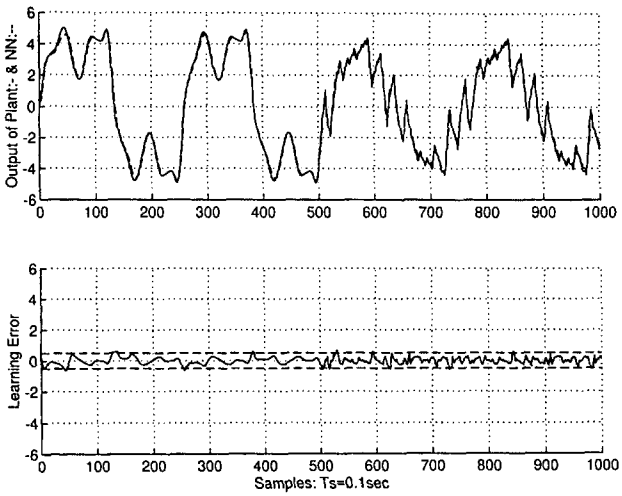


Figure 5: On-line learning with fixed rate $\eta = 0.34$ for $f_1(u)$

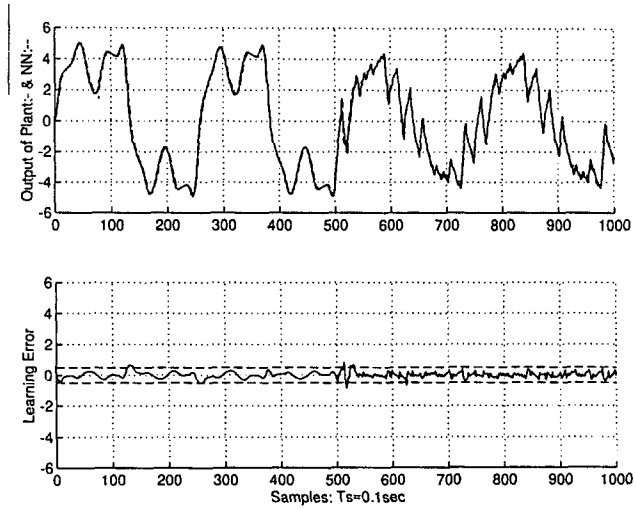


Figure 6: On-line learning with variable rate for $f_1(u)$

Transactions on Automatic Control, Vol. 44, No. 5, pp. 1026-1031, 1999.

[5] D. E. Rumelhart and J. L. McClelland, Parallel Distributed Processing. Cambridge, MA: MIT Press, 1986.

[6] B. Widrow, M. A. Lehr, 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation, *Proceedings of The IEEE*, Vol.78, No.9, Special Issue on Neural Networks, I: Theory & Modeling, pp.1415-1442, September 1990.

[7] C. -M. Kuan, K. Hornik, Convergence of Learning Algorithms with Constant Learning Rates, *IEEE Transactions on Neural Networks,* Vol. 2, No. 5, pp. 484-489, 1991.

[8] Q. Song and J. Xiao, On the Convergence Performance of Multi-layered NN Tracking Controller, *Neural & Parallel Computation*, Vol. 5, No. 3, 1997.

[9] R. A. Jacobs, Increased rates of convergence through learning rate adaption, Neural networks, Vol.1, 295-307, 1988.

[10] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink, and D. L. Alkon, Accelerating the convergence of the back-propagation method, *Biol. Cybern.*, vol. 59, pp. 257-263, 1988.

[11] D. C. Park, M. A. El-Sharkawi, R. J. Marks II, An Adaptively Trained Neural Network, *IEEE Transactions on Neural Networks,* Vol.2, No.3, pp. 334- 345, 1991.

[12] R. Batruni, A Multilayer Neural Network with Piecewise-Linear Structure and Back-Propagation Learning, *IEEE Transactions on Neural Networks,* Vol. 2, No. 3, pp. 395- 403, 1991.

[13] M. Fukumi, S. Omatu, A New Back-Propagation Algorithm with Coupled Neuron, *Transactions on Neural Networks,* Vol. 2, No. 5, pp. 535-489, 1991.

[14] S.-H. Oh, Improving the error backpropagation algorithm with a modified error function, *IEEE Trans. Neural Networks*, vol. 8, pp. 799-803, 1997.

[15] A. van Ooyen and B. Nienhuisl, improving the convergence of the back-propagation algorithm, *Neural Networks*, vol. 5, pp. 465-471, 1992.

[16] P. Saratchandran, Dynamic Programming Approach to Optimal Weight Selection in Multilayer Neural Networks, *Transactions on Neural Networks,* Vol. 2, No. 4, pp. 465-467, 1991.

[17] M. A. Sartori, P. J. Antsaklis, A Simple Method to Derive Bounds on the Size and to Train Multilayer Neural Networks, *Transactions on Neural Networks,* Vol. 2, No. 4, pp. 467-471, 1991.

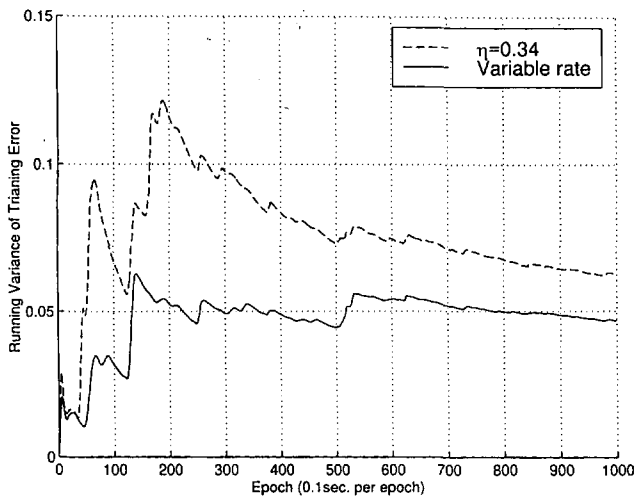


Figure 7: Time evolution of training error for $f_1(u)$

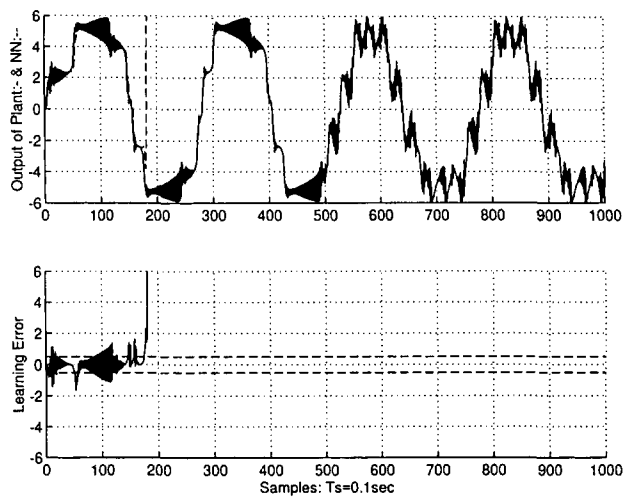


Figure 8: On-line learning with fixed rate $\eta = 0.34$ for $f_2(u)$

[18] S. Shah, F. Palmieri, M. Datum, Optimal filtering algorithms for fast learning in feedforward neural networks, *Neural Networks*, Vol.5, No.5, pp. 779-787, 1992.

[19] C. M. Bishop, Curvature-Driven Smoothing: A Learning Algorithm for Feedforward Networks, *IEEE Transactions on Neural Networks,* Vol. 4, No. 5, pp. 882- 884, 1993.

[20] A. Back, E. Wan, S. Lawrence, A. C. Tsoi, A Unifying View of Some Training Algorithms for Multilayer Perceptrons with FIR Filter Synapses, *Neural Networks for Signal Processing 4*, Edited by J. Vlontzos and J. Hwang and E. Wilson, IEEE Press, pp. 146–154, 1995.

[21] Q. Song, Implementation of Two Dimensional Systolic Algorithms for Multilayered Neural Networks, *JSA Journal of Systems Architecture*, Vol. 44, No. 8, 1998.

[22] E. D. Sontag, A Learning results for continuous-time recurrent neural networks, *Systems & Control Letters*, Vol.34, No.3, pp.151-158, 1998.

[23] S. Cavalieri, O. Mirabella, A novel learning algorithm which improves the partial fault tolerance of multilayer neural networks, *Neural Networks*, vol.12, No.1, pp.91-106, 1999.

[24] R. P. Brent, Fast Training Algorithms for Multilayer Neural Nets, *IEEE Transactions on Neural Networks,* Vol. 2, No. 3, pp. 346- 354, 1991.

[25] M. T. Hagan and M. Menhaj, Training feedforward networks with Marquardt algorithm, *IEEE Transactions on Neural Networks*, Vol.5, No.6, pp.989-993, 1994.

[26] R. Parisi, E. D. Di Claudio. G. Orlandi, and B. D. Rao, A general- ized learning paradigm exploiting the structure of feedforvard neural networks, *IEEE Trans. Neural Networks*, vol. 7, pp. 1450-1459, 1996.

[27] G.-J. Wang and C.-C. Chen, A fast multilayer neural networks training algorithm based on the layer-by-layer optimizing procedures, *IEEE Trans. Neural Networks*, vol. 7, pp. 768-775, 1996.

[28] F. Biegler-Konig and F. Marmann, A learning algorithm for multilay- ered neural networks based on linear least squares problems, *Neural Networks*, vol. 6, pp. 127-131, 1993.

[29] J. Y. F. Yam and W. S. Chow, Extended least squares based algorithm for training feedforward networks, *IEEE Trans. Neural Networks*, vol.8, pp. 806-810, 1997.

[30] S.-H. Oh, S.-Y. Lee, A new Error Function at Hidden Layers for Fast Training of Multilayer Perceptrons, *IEEE Transaction on Neural Networks*, Vol.10, No.4, pp. 960-964, 1999.

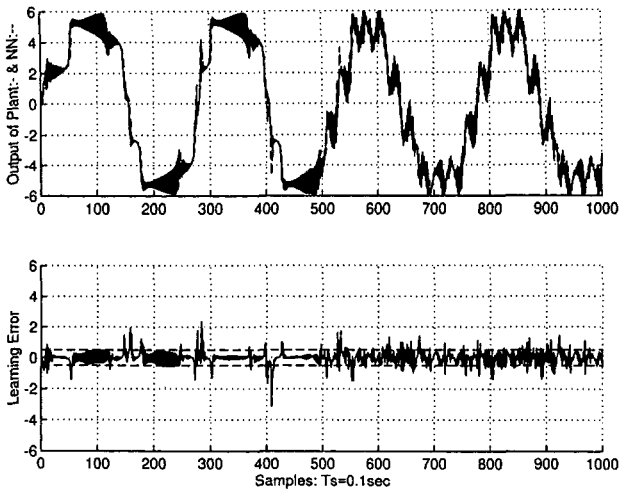


Figure 9: On-line learning with adaptive variable rate for $f_2(u)$

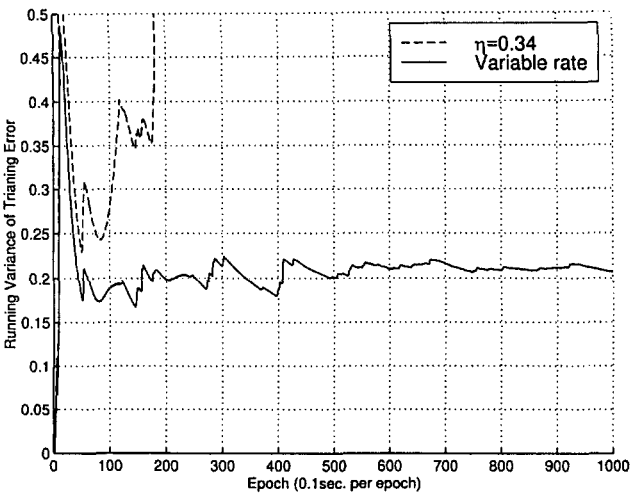


Figure 10: Time evolution of training error for $f_2(u)$

[31] Y. Lee, S.-H. Oh, and M. W. Kim, An analysis of premature saturation in back-propagation learning, *Neural Networks*, vol. 6, pp. 719-728, 1993.

[32] S.-H. Oh and Y. Lee, Effect of nonlinear functions on correlation between weighted sums in multilayer perceptrons, *IEEE Trans. Neural Networks*, vol. 5, pp. 508-510, 1994.

[33] S. Ergezinger and E. Thomsen, An accelerated learning algorithm for multilayer pereceptrons: Optimization layer by layer, *IEEE Trans. Neural Networks*, vol. 6, pp. 3142, 1995.

[34] Y. Zhang, X. R. Li, A Fast U-D Factorization - Based Learning Algorithm with Applications to Nonlinear System Modeling and Identification, *IEEE Transaction on Neural Networks*, Vol.10, No.4, pp. 930-938, 1999.

[35] G. Chen, H. Ogmen, Modified extended Kalman filtering for supervised learning, *Int. J. Syst. Sci.*, Vol.24, No.6, pp.1207-1214,1993.

[36] Y. Iiguni, H. Sakai, H. Tokumaru, A real-time learning algorithm for a multilayered neural network based on the extended Kalman filter, *IEEE Trans. Signal Processing*, Vol.40, No.4, pp. 959-966, 1992.

[37] G. Puskorius, L. A. Feldkamp, Neuralcontrol of nonlinear dynamical systems with Kalman filter trained recurrent networks, *IEEE Trans. Neural Networks*, Vol.5, pp.279-297, 1994.

[38] S. Shah, F. Palmieri, M. Datum, Optimal filtering algorithms for fast learning in feedforward neural networks, *Neural Networks*, Vol.5, pp.779-787, 1992.

[39] D. Sha, V. B. Bajić, On-line Variable Learning Rate BP Algorithm for Multilayer Feedforward Neural Networks, *in Development and practice of artificial intelligence techniques* (V. Bajic and D. Sha, Editors), pp.51-58, IAAMSAD, Durban, South Africa, September 1999.

[40] K. S. Narendra, K.Parthasarathy, Identification and control of dynamical systems using neural networks, *IEEE Transactions on Neural Networks*, Vol. 1, No. 1, pp. 4-27, 1990.

[41] K. S. Narendra, K.Parthasarathy, Gradient methods for optimization of dynamical systems containing neural networks, *IEEE Transactions on Neural Networks*, Vol. 2, pp. 252-263, 1991.

[42] S. Bhama, H. Singh, Single Layer Neural Networks for Linear System Identification Using Gradient Descent Technique, *IEEE Transactions on Neural Networks*, Vol. 4, No. 5, pp. 884-888, 1993.

[43] J. Patra, R. N. Pal, B. N. Chatterji, C. Panda, Identification of Nonlinear Dynamic Systems Using Functional Link Artificial Neural Networks, *IEEE Transactions on Systems, Man, And Cybernetics*, Part B: Cybernetics, Vol. 29, No.2, pp. 254-262, 1999.

[44] S. Chen, S. A. Billings, P. M. Grant, Nonlinear System Identification using Neural Networks, *Int. J. Contr.*, Vol.51, No.6, pp.1191-1214, 1990.

[45] S. Chen, S. A. Billings, P. M. Grant, Recursive hybrid algorithm for nonlinear system identification using radial basis function networks, *Int. J. Contr.*, Vol.55, No.5, pp.1051-1070, 1992.

[46] S. Chen, S. A. Billings, Neural networks for nonlinear dynamic system modeling and identification, *Int. J. Contr.*, Vol.56, No.2, pp.319-346, 1992.

[47] N. Sadegh, A perceptron based neural network for identification and control of nonlinear systems, *IEEE Transactions on Neural Networks,* Vol. 4, pp. 982-988, Nov. 1993.

[48] T. Yamada, T. Yabuta, Dynamic system identification using neural networks, *IEEE Transactions on Systems, Man, And Cybernetics*, Vol. 23, pp. 204-211, Jan./Feb. 1993.

[49] B. Srinivasan, U. R. Prasad, N. J. Rao, Back Propagation Through Adjoints for the identification of Nonlinear Dynamic Systems Using Recurrent Neural Models, *IEEE Transactions on Neural Networks,* Vol. 5, No.2, pp. 213-228, March 1994.

[50] S. V. T. Elanayar, Y. C. Shin, Radial Basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems, *IEEE Transactions on Neural Networks,* Vol. 5, pp. 594-603, July 1994.

[51] A. Parlos, K. T. Chong, A. F. Atiya, Application of the Recurrent Multilayer Perceptron in Modeling Complex Process Dynamics, *IEEE Transactions on Neural Networks,* Vol. 5, No.2, pp. 255-266, March 1994.

[52] P. S. Sastry, G. Santharam, K. P. Unnikrishnan, Memory Neural Networks for Identification and Control of Dynamical Systems, *IEEE Transactions on Neural Networks,* Vol. 5, No.2, pp. 306-319, March 1994.

[53] S. Mukhopadhyay, K. S. Narendra, Disturbance rejection in nonlinear systems using neural networks, *IEEE Transactions on Neural Networks,* Vol. 4, pp. 63-72, 1993.

[54] E. S. Kosmatopoulos, M. M. Polycarpou, M. A. Christodoulou, P. A. Ioannou, High-order neural network structures for identification of dynamical systems, *IEEE Transactions on Neural Networks,* Vol. 6, pp. 422-431, 1995.

[55] A. U. Levin, K. S. Narendra, Recurrent identification using feedforward neural networks, *Int. J. Contr.*, Vol.6, No.3, pp.533-547, 1995.

[56] A. Alessandri and T. Parisini, Nonlinear Modeling of Complex Large-Scale Plants Using Neural Networks and Stochastic Approximation, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, Vol. 27, No. 6, pp.750-757, 1997.

[57] Q. Song, Robust Training Algorithm of Multi-layered Neural Network for Identification of Nonlinear Dynamic Systems, *IEE Proceedings-D, Control Theory and Applications*, Vol. 145, No. 1, 1998.

[58] C. L. Philip Chen, J. Z. Wan, A Rapid Learning and Dynamic Stepwise Updating Algorithm for Flat Neural Networks and the Application to Time-Series Prediction, *IEEE Transactions on Systems, Man, And Cybernetics*, Part B: Cybernetics, Vol. 29, No. 1, pp. 62-72, 1999.

[59] S. Moon, Ali Keyhani, S. Pillutla, Nonlinear Neural -Network Modeling of an Induction Machine, *IEEE Transactions on Control Systems Technology,* Vol. 7, No. 2, pp. 203-211, 1999.

[60] Y. Z. Tsypkin, J. D. Mason, E. D. Avedyan, K. Warwick, I. K. Levin, Neural Networks for Identification of Nonlinear Systems Under Random Piecewise Polynomial Disturbances, *IEEE Transactions on Neural Networks,* Vol. 10, No.2, pp. 303-311, MARCH 1999.

[61] M. Iatrou, T. W. Berger, V. Z. Marmarelis, Modeling of Nonlinear Nonstationary Dynamic Systems with a Novel Class of Artificial Neural Networks, *IEEE Transactions on Neural Networks,* Vol. 10, No.2, pp. 327-339, MARCH 1999.

[62] J. T. Connor, R. D. Martin, L. E. Atlas, Recurrent Neural Networks and Robust Time Series Prediction, *IEEE Transactions on Neural Networks,* Vol. 5, No.2, pp. 240-254, March 1994.

[63] E. S. Chang, S. Chen, B. Mulgrew, Gradient radial basis function networks for nonlinear and nonstationary time series prediction, *IEEE Transactions on Neural Networks,* Vol. 7, pp. 188-194, 1996.

[64] A. G. Kogiantis, T. Papantoni-Kazakos, Operations and Learning in Neural Networks for Robust Prediction, *IEEE Transactions on Systems, Man, And Cybernetics*, Part B: Cybernetics, Vol. 27, No. 3, pp. 402-411, 1997.

[65] J.-S. R. Jang, ANFIS: Adaptive-Network-Based Fuzzy Inference System, *IEEE Transaction on Systems, Man, And Cybernetics*, Vol. 23, No. 3, pp. 665-685, 1993.

A Spanish Interface To LogiMoo: Towards Multilingual Virtual Worlds

Veronica Dahl, Stephen Rochefort and Marius Scurtescu
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada V5A 1S6
{veronica,srochefo,mas}@cs.sfu.ca
AND
Paul Tarau
Department of Computing Science
University of Moncton
Moncton, NB, Canada E1A 3E9
tarau@info.umoncton.ca

Keywords: virtual worlds, Internet applications, natural language processing, Assumption Grammars, LogiMOO

Edited by: Vladimir Fomichov
Received: December 21, 1998 **Revised:** February 8, 1999 **Accepted:** March 12, 1999

LogiMOO is a BinProlog-based Virtual World running under Netscape for distributed group-work over the Internet and user-crafted virtual places, virtual objects and agents. LogiMOO is implemented on top of a multi-threaded blackboard-based logic programming system featuring Linda-style coordination. Embedding in Netscape allows advanced VRML and HTML frame-based navigation and multi-media support, while LogiMOO handles virtual presence and acts as a very high-level universal object broker. In this talk we shall briefly describe Assumption Grammars (the logic grammar tool used in our Spanish interface to LogiMOO) and how they can help solve some crucial computational linguistic problems such as anaphora and coordination. We shall then discuss our translator from Spanish sentences into LogiMoo commands. Finally, we shall discuss what is needed to parameterize a single language processor into specific natural languages, with the ultimate objective of transforming LogiMoo into a multilingual virtual world. In it users from various linguistic backgrounds could communicate using their own language, to be automatically translated into LogiMoo as universal interlingua.

1 Introduction

In a world where the distance from information is constantly shrinking due to the world wide web- that enormous repository of easily accessible knowledge-, one crucial obstacle to the true availability of information remains: language.

Ideally, a user should be able to retrieve documents of interest in his/her own native tongue. Automatic translation of documents is not possible, because even when the domain is restricted (say, to legal documents, or to technical jargon), the problem of translating otherwise free language is too complex to be amenable to automatic solution. Automatic machine translation usually requires downsizing both the language coverage and the application to manageable proportions.

In the specific domain of application of LogiMOO- a virtual world for distributed group-work over the Internet and user-crafted virtual places, virtual objects and agents-, language coverage is naturally restricted, giving rise to a form of controlled English, in which for instance, sentences are subjectless and in imperative form, since LogiMOO is typically used to invoke commands, requests, etc. Thus it would make an ideal candidate for machine translation.

But this very simplicity makes it attractive to explore another avenue which can provide the illusion of an automatic translator while being much simpler: we can abstract from our English parser a simple core grammar which is language independent, and complement it with as many language-dependent modules as languages we want to admit for our front ends. This article describes how this is done for Spanish, how it could easily be done for other languages as well, and discusses possible Spanish specific extensions to the language coverage.

2 Background

2.1 MUDs and MOOs

MUDs and MOOs (Multi User Domains - Object Oriented) have started with virtual presence and interaction. Their direct descendents, Virtual Worlds are a strong unifying metaphor for various forms of net-walk, net-chat and Internet-based virtual presence in general. They start where usual HTML shows its limitations: they do have state and

require some form of virtual presence. "Being there" is the first step towards full virtualization of concrete ontologies, from entertainment and games to schools and businesses.

Some fairly large-scale projects (Intel's Moondo [Int], Sony's Cyber Passage [Son], Black Sun's CyberGate [Bla], Worlds Inc.'s AlphaWorld [Wor]) converge all towards a common interaction metaphor: an avatar represents each participant in a multi-user virtual world. Information exchange reuses our basic intuitions with almost instant *learnability* for free.

The sophistication of their interaction metaphor, with VRML landscapes and realistic 'avatars' (i.e., visual representations on screen of the user) moving in shared multi-user virtual spaces, will require soon high-level agent programming tools, once the initial fascination with 'looking' human is translated into automation of complex behavior. Towards this end, high-level consultation abilities are among the most important additions to various virtual world modeling languages. These should proceed in the user's mother tongue, whose sentences are automatically parsed into a common form which is invisible to all users but is used by LogiMOO to perform its internal operations. Thus sociability can proceed among distant users, each typing in their own language, and receiving feedback in it as well, while internally all proceeds in the invisible common language which triggers the LogiMOO actions requested.

2.2 LogiMOO: a multi-paradigm virtual world

LogiMOO [DBPT96, TDB96, Tar96] is a BinProlog-based Virtual World running under Netscape or Internet Explorer for distributed group-work over the Internet and user-crafted virtual places, virtual objects and agents. LogiMOO is implemented on top of a multi-threaded blackboard-based logic programming system (Multi-BinProlog 5.25) [Tar97] featuring Linda-style coordination[1]. Virtual blackboards [DBT96] allow efficient mirroring of remote sites over TCP/IP links while Solaris 2.x threads ensure high-performance local client-server dynamics. Embedding in Netscape allows advanced VRML or HTML frame-based navigation and multi-media support, while LogiMOO handles virtual presence and acts as a very high-level universal object broker.

The LogiMOO kernel behaves as any other MOO while offering a choice between interactive Prolog syntax and a Controlled Natural Language parser allowing people unfamiliar with Prolog to get along with the basic activities of the MOO: place and object creation, moving from one place to the other, giving/selling virtual objects, talking ('whisper' and 'say'). At login time, a main interactive shell and background notifier (for messages and events targeted to the user) are created. Netscape is used to implement CGI-based BinProlog *remote toplevel* interacting with a remote LogiMOO server (Fig. 1). Objects in LogiMOO are represented as hyper-links (URLs) towards their owners' home pages where their 'native' representation actually resides in various formats (HTML, VRML, GIF, JPEG etc.).

[1] An important difference between Multi-BinProlog and predecessors like [BC91] is direct use of Linda operations, instead of a guard notation.

2.3 Assumption Grammars

Assumption Grammars are logic grammars augmented with a) linear and intuitionistic implications scoped over the current continuation, and b) hidden multiple accumulators, useful in particular to make the input and output strings invisible.

2.3.1 Linear and intuitionistic implications

Implications are additional information which is only available during the continuation, i.e., the remainder of the current proof. If declared to be intuitionisic (noted `assumei`), they can be used (noted `assumed`) an indefinite number of times. In contrast, linear implications (noted `assumel`) can be used at most once, and then they dissapear.

For instance, the Prolog query:

```
?- assumei(p(5)), assumed(p(X)),
     assumed(p(Y)).
```

instantiates both X and Y into 5; whereas the query:

```
?- assumel(p(5)), assumed(p(X)),
     assumed(p(Y)).
```

fails after instantiating X to 5, since p(5) is no longer available.

Both types of implication vanish upon backtracking.

Intuitionistic implications have scoped versions as well: `Clause=>Goal` and `[File]=>Goal` make `Clause` or respectively the set of clauses found in `File`, available only during the proof of `Goal`. Clauses assumed with `=>` are usable an indefinite number of times in the proof, e.g. `a(13) => (a(X),a(Y))` will succeed.

The scoped version of linear implication, `Clause -: Goal` or `[File] -: Goal`, makes `Clause` or respectively the set of clauses found in `File` available only during the proof of Goal. They vanish on backtracking and each clause is usable at most once in the proof, i.e. `a(13) -: (a(X),a(Y))` will fail. Note however, that `a(13) -: a(12) -: a(X)` will succeed with `X=12` as alternative `X=13` as answers, while its non-affine counterpart `a(13) -o a(12) -o a(X)` as implemented in Lolli or Lygon, would fail.

We can see the `assumel/1` and `assumei/1` builtins as linear affine and respectively intuitionistic implication scoped over the current AND-continuation, i.e. having their assumptions available in future computations on the *same* resolution branch.

The following sample Assumption grammar illustrates the use of linear assumptions to handle relativization. It

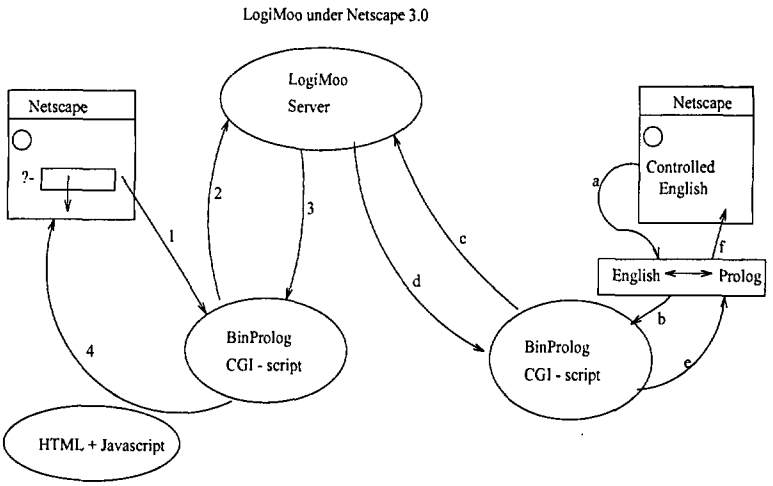


Figure 1: LogiMOO on the Web

also makes use of lambda calculus representations. These are attractive if one views the meaning of syntactic categories as mappings from either variables or other properties into new properties. For example, we can consider the category "name" ("n" for short in the grammar below) as a "meaning device" that takes a variable X and constructs a property from it- such as dilemma(X), or lesson(X). A determiner ("det" in our grammar), on the other hand, can be viewed as a device that takes two properties (roughly representing the subject and the predicate of the sentence) and constructs a new property trying those two up and rendering the meaning of the specific determiner.

The meanings of constituents are, then, compositionally built up by beta-reduction [2] from the representations of their subconstituents. Thus, for instance, the rule that produces a sentence's meaning representation S from the meaning representations of its subject noun phrase (represented NP) and its verb phrase (represented VP) can be stated as:

```
s(S):- np(NP), vp(VP), apply(VP,NP,S).
```

where "apply" implements beta-reduction, and is (strikingly simply) defined in turn as follows ("stands for "lambda", used as an operator in infix notation):

```
apply(X\P,X,P).
```

When representing a noun phrase with a relative clause, we can make use of the general definition for sentence given above to parse the relative clause itself. But since a relative clause is a sentence with an implicit noun phrase recoverable from its antecedent (e.g. "the flower that Florence planted" exhibits a relative clause, "that Florence planted" whose missing direct object is understood to be the antecedent "the flower"), in constructing its semantics we need to save the antecedent in some place from which we can recover it when we come across a missing noun phrase. We can save it as an assumption, made by a noun phrase rule, that the variable representing the noun phrase's head noun will be needed in some missing noun phrase to be found later in the relative clause:

```
np(NP):- det(D), n(X\P), +missing(X),
    nl, relclause(P1),
    apply(D,X\and(P,P1),NP).
```

All we have to do now is to recover this assumption made when we expect a noun phrase that does not materialize:

```
np(P\Q):- -missing(X),
    % retrieve antecedent
    apply(P,X,Q).
```

Its representation X will be used in the call to beta-reduction.

The predicates "used" and "all_consumed", which we have not bothered with here but which appear in our complete listing below, serve to ensure, at appropriate points, that no assumptions made are left unconsumed.

Before showing the complete grammar, here are some sample tests. Our grammar translates into an English-like interlingua, to anticipate the LogiMOO grammar we shall present later, which has this same characteristic. But it would be just as easy, of course, to provide a Spanish-based internal representation instead.

2.3.2 Sample Tests

```
% Every dilemma costs
sentence([todo,dilema,cuesta]).
    every(_x2376,dilemma(_x2376) =>
    costs(_x2376))

% John baffles
sentence([juan,desconcierta]).
```

[2] beta-reduction is the lambda-calculus operation that applies an expression λ(X,P) to an expression Q, obtaining Q', which is equal to Q except that all occurrences of X have been replaced by P.

```
    baffles(john)

% Every lesson that costs baffles
sentence([todo,aprendizaje,que,cuesta,
    desconcierta]).
every(_x2380,and(lesson(_x2380),
    costs(_x2380)) => baffles(_x2380))

% Evey dilemma that John solves costs
sentence([todo,dilema,que,juan,resuelve,
    cuesta]).
every(_x3303,and(man(_x3303),
    saw(mary,_x3303)) => paints(_x3303))

% Johns solves every dilemma that costs
sentence([juan,resuelve,todo,dilema,que,
    cuesta]).
every(_x2407,and(dilemma(_x2407),
    costs(_x2407)) =>
    resolves(john,_x2407))

% Every dilemma that John solves
% solves Prolog
sentence([todo,dilema,que,juan,resuelve,
    resuelve,prolog]).
every(_x2384,and(dilemma(_x2384),
    resolves(john,_x2384)) =>
    resolves(_x2384,prolog))

% Every dilemma that presents a dilemma
% solves a dilemma
sentence([todo,dilema,que,presenta,un,
    dilema, resuelve,un,dilema]).
every(_x2388,and(dilemma(_x2388),exists(
    _x2551,and(dilemma(_x2551),
    presents(_x2388,_x2551)))) =>
    exists(_x2629,and(dilemma(_x2629),
    resolves(_x2388,_x2629))))

% John solves a dilemma that presents
% every dilemma
sentence([juan,resuelve,un,dilema,que,
    presenta,todo,dilema]).
exists(_x2411,and(and(dilemma(_x2411),
    every(_x2592, dilemma(_x2592) =>
    presents(_x2411,_x2592))),
    resolves(john,_x2411)))

% John solves a dilemma that solves
% every dilemma that presents a dilemma
sentence([juan,resuelve,un,dilema,que,
    resuelve,todo,dilema,que,presenta,
    un,dilema]).
exists(_x2419,and(and(dilemma(_x2419),
    every(_x2600,and(dilemma(_x2600),
    exists(_x2780,and(dilemma(_x2780),
    presents(_x2600,_x2780)))) =>
    resolves(_x2419,_x2600))),
    resolves(john,_x2419)))
```

2.3.3 The Complete Grammar

```
% N.B. Words are noted
% with # preceding them

:-op(300,xfy,\).

apply(X\P,X,P).

all_consumed:- \+assumed(missing(_)).

% Grammar:

% Lexicon:

pn(P\Q):- #juan, apply(P,john,Q).
pn(P\Q):- #prolog, apply(P,prolog,Q).

det(P1\P2\every(X,Q1 => Q2)):-
    #todo, apply(P1,X,Q1),
    apply(P2,X,Q2).
det(P1\P2\exists(X,and(Q1,Q2))):-
    #un, apply(P1,X,Q1),
    apply(P2,X,Q2).

n(X\dilemma(X)):- #dilema.
n(X\lesson(X)):- #aprendizaje.

vi(P\Q):-
    #cuesta, apply(P,X\costs(X),Q).
vi(P\Q):-
    #desconcierta,
    apply(P,X\baffles(X),Q).

vt(P1\P2\Q2):-
    #presenta, apply(P2,X\Q1,Q2),
    apply(P1,Y\presents(X,Y),Q1).
vt(P1\P2\Q2):-
    #resuelve, apply(P2,X\Q1,Q2),
    apply(P1,Y\resolves(X,Y),Q1).

% Syntax:

s(S):- np(NP), vp(VP),
    apply(VP,NP,S), all_consumed.

np(NP):- pn(NP).
np(P\Q):-
    -missing(X), % retrieve antecedent
    apply(P,X,Q).

np(NP):- det(D), n(X\P),apply(D,X\P,NP).
np(NP):- det(D), n(X\P), +missing(X),
```

```
                relclause(P1), used(X),
                ap-
ply(D,X\and(P,P1),NP).

used(X):- -missing(X), !, fail.
used(_).

relclause(Rel):- #que, s(Rel).

vp(VP):- vi(VP).
vp(VP):- vt(V), np(NP), apply(V,NP,VP).

test:- sentence(X), dcg_def(X), s(S),
    dcg_val([]), write(S), nl.

% dcg_def gives in X the value of
%   the input stream;
% dcg_val puts the stream's current
%   value in its argument.
```

This grammar is an extension of a grammar developed by Alain Colmerauer which only handled simple noun phrases (i.e., with no relative clauses). Although in our grammar we only treat relative clauses, our methodology for treating them is also applicable to other long-distance dependency phenomena. In [DTL97] we examine the uses of AGs for three crucial such problems in natural language processing: free word order, anaphora and coordination.

An alternative to programming beta-reduction as in the above grammar would be to use a language such as lambda-Prolog. In our opinion this would not provide too much economy with respect to our already very concise code, and would lose the advantage of portability.

3 The Core LogiMOO Grammar

We first use an English lexicon for exemplifying purposes. Next we explain how to make this core grammar language independent, taking Spanish as an example. The resulting Spanish grammar is shown in the Appendix.

3.1 The Lexicon

3.1.1 Noun Definitions.

The rule:

```
proper_name(john-[masc,sing]) :- #john.
```

defines the word `john` as a masculine and singular proper name represented by the constant `john`.

Objects are introduced by noun definitions, e.g.,

```
noun(car-[neut,sing]) :- #car.
```

Virtual places are also introduced by nouns, and are always set to a neutral gender and singular form, e.g.

```
noun(south-[neut,sing]) :- #south.
```

3.1.2 Verb Definitions.

Intransitive verbs are verbs not requiring other objects/persons. correspond to actions performed by an avatar her/himself and involve no other specific avatar, place, or object, e.g.

```
intrans_vb(smile) :- #smile.
```

Transitive verbs requires one extra object/person `Y`. involve one other avatar, object or place in the the virtual world. For instance, the rule:

```
trans_vb(Y,go(Y)) :- #go.
```

corresponds to a user actioning her/his avatar to go someplace in the virtual world. The two arguments identify the object *Y*, and the translation to the unary predicate *go*, which can be used to associate some action to this verb.

Bitransitive verbs. requiring two extra objects/persons. specify an action by an avatar that involves any two objects, avatars, or places. As an example,

```
bitrans_vb(Y,Z,give(Y,Z)) :- #give.
```

specifies the action of giving somebody `Y`, some object `Z`. The third argument is the predicate, *give(Y,Z)*, which can again be used to associate some action to this verb.

Notice that in all verbs, the subject is left implicit. In the application we further describe it will be clear from the context who should be the subject. given that the LogiMOO kernel recognizes it as the avatar who logged in.

3.1.3 Pronouns, Determiners, and Prepositions.

Pronouns are specified in a similar way as nouns:

```
pronoun(_X-[fem,sing]) :- #she.
```

This specification identifies a pronoun, *she*, with a feminine gender and singular form. Agreement information is used to resolve pronoun references into the correct object or person.

Determiners and prepositions are specified as

```
det :- #the.
preposition :- #to.
```

3.2 Syntactic Rules

All sentences are in imperative form, with their subject left implicit. Thus they reduce to verb phrases, which can be of the following forms:

(VP1) An intransitive verb.
(VP2) A transitive verb followed by a noun.
(VP3) A transitive verb followed by a noun phrase.
(VP4) A transitive verb followed by a prepositional phrase.

(VP5) A bitransitive verb followed by two noun phrases.
(VP6) A bitransitive verb followed by a noun phrase and a prepositional phrase.

A prepositional phrase is defined as

(PP1) A preposition followed by a noun phrase.

The noun phrase forms allowed are

(NP1) A proper name.
(NP2) A pronoun (anaphora).
(NP3) A determiner followed by a noun.

(more complex noun phrases will be explained in the next section)

In addition, we identify communication inputs which occur when a user wants their avatar to say, whisper or yell some message, e.g.

say hi how are you.

This form of input is introduced by either:

(F1) The word "whisper" followed by a prepositional phrase followed by a message.
(F2) The word "say" followed by a message.
(F3) The word "yell" followed by a message.

Table 1 shows some sample parses.

4 Adapting the Core Grammar to Spanish

Our technique for splitting the grammar into a language-independent core subset of rules and a language-dependent one is very simple. It comes from the observation that there are two types of rules in the English grammar which are language dependent:

a) rules that create a predicate name which is reminiscent of the noun, verb or adjective from which they spring,

b) rules containing a lexical item (i.e., a symbol preceded by '#').

For rules of type a), we simply translate the predicate name into its Spanish equivalent by means of the Bin-Prolog builtin "means", e.g.

whisper means susurra

For rules of type b), we replace the lexical item by a non-terminal of same name, and relegate its final rewriting into a word to the language-specific lexical module which is called for each language. E.g., for "wizard" we would have:

```
% wizard is now a non-terminal
name --> wizard

%English lexicon:        % Spanish lexicon:

wizard --> #wizard       wizard --> #brujo.
```

Of course, more realistically we will need features such as gender and number in order to produce the right words in each language. For instance, whereas in English we have only one lexical form for the definite noun, whether it is singular, plural, feminine or masculine, in Spanish we have four different lexical items covering all these forms.

The same technique used here for Spanish can be used for implementing at least other romance languages within our language coverage. Thus the core grammar can largely be made language independent with relatively little effort.

5 Conclusion and future work

We have provided another dimension of extensibility to an already extensible English front end to LogiMOO which was described in [DTL97]. This front end was extensible in the sense that "content" words could be added to the grammar definition through a high level description of their syntactic type (e.g. verb requiring one complement, etc.) plus the sequence of LogiMOO commands into which they should translate.

In this article we have explored extensibility into different natural languages, not by using the usual machine translation approach, but by abstracting a core set of language independent rules from our English parser and then adding a language specific lexicon (English, Spanish or other) to complete the grammar definition. A simple change of one lexicon module into another effects the language change invisibly, so that users across the world can type in their interactions in their own language, these are recorded in a "neutral" but invisible form, from which any retrieval continues to respect the language of the caller.

The syntax covered by our controlled natural language subset should not be expanded much more, because it is precisely owing to the controlled nature of our subsets that we are able to get away with such an easy specialization into one language or another of our language-independent core grammar. However within a single language and a single application our techniques can be fruitfully used to cover more ambitious natural language subsets.

An extension that would not result in a larger subset of language but which would increase human-like comprehension would be that of recovering implicit meanings from various forms in language. For instance, lexical definitions for the Spanish singular definite article could include an indication of its presupposition of existence and unicity, so that if the presupposition fails this could be indicated on the fly.

6 Appendix

6.1 The LogiMOO Grammar

```
% NL interface to LogiMOO
% Authors: Veronica Dahl, Paul Tarau
```

NL Input	Translation	LogiMOO Action
look.	look.	Provides the user with a description of the room that their avatar currently occupies.
craft a car.	craft(car).	Creates a virtual object, car, owned by th e avatar.
craft a car. give it to john.	craft(car), give(john, car).	Creates a virtual object, car, and gives i t to john.
take the car that john crafted.	and(take(X), crafted(john, X), is_a(X, car))	Puts a car object crafted by j ohn into the avatar's possession.
whisper to john	whisper(john, 'How are you').	Sends the message 'How are you' to john .

Table 1: Sample Parses

```
:-op(400,xfx,(@)).
:-op(400,fx,(?)).
:-op(400,fx,(++)).

% Dictionary expresed as
% OtherLanguage@English in file
% english.pl ==> trivial def: X@X

@EnglishW:- #OtherW,OtherW@EnglishW.

look_ahead(W):- \+ \+ (@ _,#W).
look_ahead(W1,W2):- \+ \+
    (@ _, #W1, #W2).

% optional 'glue' words: skip if
% present, do not mind if not

{W} :- nonvar(W),!,(@W->true;true).
{W} :- logimoo_err(should_be_nonvar(W)).

?X :- is_assumed(future(X)).

% this is just heuristics: occasionally
%   we get it wrong
% anaphora requires more: for instance,
%   feature matching

% at most 2 uses for an anphora
++X:- assumeal(X),assumel(X).
% we assume both in asserta and
% assertz order

is_avatar(X) :-
    is_a_fact(user(X,_,_)) ; ?avatar(X).
is_crafted(A,C) :-
    is_a_fact(crafted(A,C)) ;
    ?crafted(A,C).
is_place(P) :-
    is_a_fact(place(P)) ; ?place(P).
is_port(P) :-
    is_a_fact(port(_,P,_)) ; ?port(P).
is_in(C,X) :-
    is_a_fact(contains(C,X)) ;
    ?contains(C,X).
does_have(Who,What) :-
    is_a_fact(has(Who,What)) ;
    ?has(Who,What).

% EVALUATES A LIST OF CHARS

eval_nat(Chars):-
    translate_nat(Chars).

translate_nat([S,A,Y|Cs]):-
    member([S,A,Y],["say","Say"]),!,
    that_mes(Cs).
translate_nat(Chars):-
    % ignore case
    !,toLowerChars(Chars,Cs),
    % split in words
    chars_words(Cs,Ws),
    patch_words(Ws,Words),
    % generate and execute commands
    translate(Words).

% split in senteces, then parse each
translate(Words) :-
    !,writeln(['WORDS:',Words]),
    split_nat(Words,Sentences),
    !,
    writeln(['SENTENCES:'|Sentences]),
    !,nl,
    writeln(['==BEGIN CMD RESULTS==']),
    ( text(Sentences)->true
    ; true
    ),
    !,nl,
    writeln(['==END CMD RESULTS==']),
```

```
    nl.

% parse each sentence
text([]).
text([S|Ss]):-
   parse_sent(S,C),
   !,
   evaluate(C),
   !,
   text(Ss).

evaluate(T) :-
    metacall(T),!,
    quietmes(2,'SUCCEEDING'(T)).
evaluate(T) :-
    logimoo_err(
 unexpected_evaluation_failure_in(T)).

% parse a sentence
parse_sent(S,Cs):-
    dcg_def(S),  % open dcg stream
    % Cs contains a list of commands
    sent(Cs),
    dcg_val([]), % close dec stream
    !.

% send errors to 'wizard' over the net
parse_sent(S,_):-
    logimoo_err(
 unable_understand_sentence(S)).

% look for a verb phrase
sent(R) :- vp(R).

% recognizes an avatar,
% using world knowledge
avatar(X) :-
    art, @X, is_avatar(X), !,
    ++future(avatar(X)).
avatar(X) :- anaphora(avatar,X).

% recognizes objects using
% world knowledge
crafted(What):- @the,@What,
    is_crafted(Who,What),
    relative(What),!,
    ++future(crafted(Who,What)).
crafted(What):-
    anaphora(crafted,What).

% uses relative sentences as filters
relative(What) :- @that,!,avatar(Who),
    @Verb,do_relative(Verb,Who,What).
relative(What) :- nonvar(What).

% handles relatives,
% based on world knowledge
do_relative(crafted,Who,What) :-
    is_crafted(Who,What).
do_relative(has,Who,What) :-
    does_have(Who,What).
do_relative(have,Who,What) :-
    does_have(Who,What).
% to add: others,
% based on: is_avatar,is_place,is_in

% generates a com-
mand to craft an object
def_crafted(X):-
    do_craft(X), whoami(I),
    ++future(crafted(I,X)).

% recognizes a place using
% domain knowledge
place(X) :- @X, is_place(X).

% generates a command
% to build a new place
def_place(X) :- @X, ++future(place(X)).

% recognizes a port
direction(X) :-
    @X, is_port(X),
    ++future(direction(X)).

% handles him,her,it
anaphora(avatar,X) :-
    @P, get_avatar(P,X).
anaphora(crafted,X) :-
    @it, -future(crafted(_,X)).

get_avatar(i,X) :- !, whoami(X).
get_avatar(P,X) :-
    member(P,[him,her]),!,
    -future(avatar(X)).

% VERBS

vp(go(Where)) :- @go,!,do_go(Where).
vp(come(Who)) :- @come,!,avatar(Who).
vp(craft(What)) :-
    @craft,!,art,def_crafted(What).
vp(dig(Place)) :-
    @dig,!,art,def_place(Place).
vp(open_port(Port,Place)) :- @open,!,
    art,@port,@Port,{to},art,
    def_place(Place).
vp(close_port(Port)) :-
    @close,!,{the},@port,{to},art,@Port.
vp(take(What)) :-
    @take,!,art,crafted(What).
vp(drop(What)) :-
    @drop,!,art,crafted(What).
vp(show(What)) :-
```

```
    @show,!,art,crafted(What).
vp(give(Whom,What)) :-
    @give,!,obp(Whom,What).
vp(iam(Who)) :-
    @i,@am,!,art,@Who,
    ++future(avatar(Who)).
vp(Cmd) :- @who,!,do_who(Cmd).
vp(what(Verb,Object)) :-
    @what,!,do_what(Verb,Object).
vp(Cmd) :- @where,!,do_where(Cmd).
vp(please(Who,What)) :-
    @please,!,avatar(Who),vp(What).
vp(X) :- @X,nonvar(X),
    member(X,
     [look,list,list0,users,online,
      whoami,whereami,test,ttest,
      listing,save,help,lobby,vanish,
      messages,sstop,sstart]).

art:- @a;@an;@the;true.

obp(Whom,What):-
    @to,avatar(Whom),crafted(What).
obp(Whom,What) :-
    crafted(What),@to,avatar(Whom).

do_where(whereami):- @am,@i;@i,@am.
do_where(where(X)):-
    @(is),art,(avatar(X);crafted(X)).

do_craft(X) :- @Pref, {-}, @Suf,!,
  namecat(Pref,'.',Suf,X).
do_craft(X) :- @X.

do_go(Where) :-
    {to},{the},
    (place(Where);direction(Where)).
do_go(Where) :-
    @there, -future(place(Where)).

do_who(whoami):- @am,@i,!.
do_who(whoami):- @i,@am,!.
do_who(who(Verb,Object)) :-
    @Verb,crafted(Object).

do_what(Verb,X):- @Verb,avatar(X).

% LOW LEVEL TOOLS

% can be used as message sender
logimoo_err(Mes) :-
    errmes('LogiMOO error',Mes).

patch_words(Ws,Words) :-
    append(_,[Last],Ws),
    is_dot(Last),!,Words=Ws.
patch_words(Ws,Words) :-
    is_dot(X),!,append(Ws,[X],Words).

is_dot(X) :-
    member(X,['.', ',' ,'?', '!']),!.

dot :- #X,is_dot(X).

nl_word(_):-dot,!,fail.
nl_word(X) :- #X.

split_nat(Ws,Ss):-
    dcg_def(Ws),
    plus(a_sent,Ss),
    dcg_val([]),!.

a_sent(S):-plus(nl_word,S),dot.

toLowerChar(X,Y) :-
   [A]="A", [LA]="a",
   [Z]="Z",
    X >= A,
    X =< Z,
    !,
    Y is LA+X-A.
toLowerChar(X,X).

toLowerChars(Cs,Ls):-
    map(toLowerChar,Cs,Ls).


toLower(X,LX) :-
    term_chars(X,Cs),
    toLowerChars(Cs,Ls),
    term_chars(Ls,LX).

test:-
   test_data(Cs),
   write_chars("TEST: "),
   write_chars(Cs),nl,
   eval_nat(Cs),nl,
   fail
;  nl.
```

6.2 The Spanish Grammar

```
construya@craft.
puerta@port.

dese@give.
dele@give.

esta@is.
estoy@am.
soy@am.
tiene@has.
cave@dig.
diga@dig.
```

```
abra@open.
vaya@go.
venga@come.
mire@look.
construi@crafted.
construya@craft.

  un@a.
  una@a.

    perro@dog.
    perra@dog.
    gato@cat.
    gata@cat.

      brujo@wizard.
      bruja@wizard.
      yo@i.
      dormitorio@bedroom.
      vestibulo@lobby.
      habitacion@room.

        lo@it.
        el@the.
        del@the.
        la@the.

          a@to.
          al@to.
          de@to.

donde@where.
quien@who.
que@that.

  alsur@south.
  alnorte@north.
  alli@there.

porfavor@please.

X@X.
```

6.3 Sample Tests

```
test_data("Yo soy Paul.").
test_data("Cave una habitacion_huespedes.
 Vaya alli. Cave una cocina.").

test_data("Vaya al vestibulo. Mire.").

test_data("Yo soy el brujo.
 Donde estoy yo?").

test_data("Cave el dormitorio. Vaya alli.
 Cave una cocina, abra una puerta alsur de
 la cocina, vaya alli, abra una puerta
 alnorte del dormitorio. Vaya alli.
 Construya un cuadro. Dese lo al brujo.
 Mire.").

test_data("Yo soy Diana. Construya un
 automovil. Donde esta el automovil?").

test_data("Construya un Gnu. Quien tiene
 lo? Donde esta el Gnu? Donde estoy yo?").

test_data("Dele al brujo el Gnu que yo
 construi. Quien tiene lo?").

/* TRACE:

==BEGIN COMMAND RESULTS==
TEST: Yo soy Paul.
WORDS: [yo,soy,paul,.]
SENTENCES: [yo,soy,paul]

==BEGIN COMMAND RESULTS==
login as: paul with password: none
your home is at http://199.60.3.56/~veronica

SUCCEEDING(iam(paul))

==END COMMAND RESULTS==

TEST: Cave una habitacion_huespedes.
 Vaya alli. Cave una cocina.
WORDS: [cave,una,habitacion_huespedes,.,
 vaya,alli,.,cave,una,cocina,.]
SENTENCES: [cave,una,habitacion_huespedes]
 [vaya,alli] [cave,una,cocina]

==BEGIN COMMAND RESULTS==
SUCCEEDING(dig(habitacion_huespedes))
you are in the  habitacion_huespedes
SUCCEEDING(go(habitacion_huespedes))
SUCCEEDING(dig(cocina))

==END COMMAND RESULTS==

TEST: Vaya al vestibulo. Mire.
WORDS: [vaya,al,vestibulo,.,mire,.]
SENTENCES: [vaya,al,vestibulo] [mire]

==BEGIN COMMAND RESULTS==
you are in the  lobby
SUCCEEDING(go(lobby))
user(veronica,none,'http://...').
user(paul,none,'http://...').
login(paul).
online(veronica).
online(paul).
place(lobby).
place(habitacion_huespedes).
place(cocina).
contains(lobby,veronica).
contains(lobby,paul).
SUCCEEDING(look)

==END COMMAND RESULTS==
```

```
TEST: Yo soy el brujo. Donde estoy yo?
WORDS: [yo,soy,el,brujo,.,donde,estoy,yo,?]
SENTENCES: [yo,soy,el,brujo]
 [donde,estoy,yo]

==BEGIN COMMAND RESULTS==
login as: wizard with password: none
your home is at http://199.60.3.56/~veronica

SUCCEEDING(iam(wizard))
you are in the  lobby
SUCCEEDING(whereami)

==END COMMAND RESULTS==

TEST: Cave el dormitorio. Vaya alli. Cave
 una cocina, abra una puerta alsur de la
 cocina, vaya alli, abra una puerta alnorte
 del dormitorio. Vaya alli. Construya un
 cuadro. Dese lo al brujo. Mire.
WORDS: [cave,el,dormitorio,.,vaya,alli,.,
 cave,una,cocina,(,),abra,una,puerta,alsur,
 de,la,cocina,(,),vaya,alli,(,),abra,una,
 puerta,alnorte,del,dormitorio,.,vaya,alli,
 .,construya,un,cuadro,.,dese,lo,al,brujo,.,
 mire,.]
SENTENCES: [cave,el,dormitorio] [vaya,alli]
 [cave,una,cocina] [abra,una,puerta,alsur,
 de,la,cocina] [vaya,alli] [abra,una,puerta,
 alnorte,del,dormitorio] [vaya,alli]
 [construya,un,cuadro] [dese,lo,al,brujo]
 [mire]

==BEGIN COMMAND RESULTS==
SUCCEEDING(dig(bedroom))
you are in the  bedroom
SUCCEEDING(go(bedroom))
SUCCEEDING(dig(cocina))
SUCCEEDING(open_port(south,cocina))
you are in the  cocina
SUCCEEDING(go(cocina))
SUCCEEDING(open_port(north,bedroom))
you are in the  bedroom
SUCCEEDING(go(bedroom))
SUCCEEDING(craft(cuadro))
logimoo:<wizard># 'wizard:I give you cuadro'
SUCCEEDING(give(wizard,cuadro))
user(veronica,none,'http://...').
user(paul,none,'http://...').
user(wizard,none,'http://...').
login(wizard).
online(veronica).
online(paul).
online(wizard).
place(lobby).
place(habitacion_huespedes).
place(cocina).
place(bedroom).
contains(lobby,veronica).
contains(lobby,paul).
contains(bedroom,wizard).
contains(bedroom,cuadro).
port(bedroom,south,cocina).
port(cocina,north,bedroom).
has(wizard,cuadro).
crafted(wizard,cuadro).
SUCCEEDING(look)

==END COMMAND RESULTS==

TEST: Yo soy Diana. Construya un automovil.
 Donde esta el automovil?
WORDS: [yo,soy,diana,.,construya,un,
 automovil,.,donde,esta,el,automovil,?]
SENTENCES: [yo,soy,diana]
 [construya,un,automovil]
 [donde,esta,el,automovil]

==BEGIN COMMAND RESULTS==
login as: diana with password: none
your home is at http://199.60.3.56/~veronica

SUCCEEDING(iam(diana))
SUCCEEDING(craft(automovil))
automovil is in lobby
SUCCEEDING(where(automovil))

==END COMMAND RESULTS==

TEST: Construya un Gnu. Quien tiene lo?
 Donde esta el Gnu? Donde estoy yo?
WORDS: [construya,un,gnu,.,quien,tiene,
 lo,?,donde,esta,el,gnu,?,donde,estoy,yo,?]
SENTENCES: [construya,un,gnu]
 [quien,tiene,lo] [donde,esta,el,gnu]
 [donde,estoy,yo]

==BEGIN COMMAND RESULTS==
SUCCEEDING(craft(gnu))
diana has gnu
SUCCEEDING(who(has,gnu))
gnu is in lobby
SUCCEEDING(where(gnu))
you are in the  lobby
SUCCEEDING(whereami)

==END COMMAND RESULTS==

TEST: Dele al brujo el Gnu que yo construi.
 Quien tiene lo?
WORDS: [dele,al,brujo,el,gnu,que,yo,
 construi,.,quien,tiene,lo,?]
SENTENCES: [dele,al,brujo,el,gnu,que,yo,
 construi] [quien,tiene,lo]

==BEGIN COMMAND RESULTS==
logimoo:<diana># 'wizard:I give you gnu'
SUCCEEDING(give(wizard,gnu))
wizard has gnu
SUCCEEDING(who(has,gnu))

==END COMMAND RESULTS==
```

```
SUCCEEDING(test)

==END COMMAND RESULTS==
*/
```

Acknowledgement

We thank for support from NSERC (grants OGP0107411 and 611024), and from the FESR of the Université de Moncton. Special thanks go to Daniel Perron for long discussions helping to come out with the initial idea of LogiMOO, to Koen De Bosschere for the implementation of the Multi-BinProlog Linda engine.

References

[BC91] A. Brogi and P. Ciancarini. The Concurrent Language, Shared Prolog. *TOPLAS*, 13(1):99–123, 1991.

[Bla] BlackSun. CyberGate. http://www.blacksun.com/.

[DBPT96] Koen De Bosschere, Daniel Perron, and Paul Tarau. LogiMOO: Prolog Technology for Virtual Worlds. In *Proceedings of PAP'96*, pages 51–64, London, April 1996.

[DBT96] K. De Bosschere and P. Tarau. Blackboard-based Extensions in Prolog. *Software — Practice and Experience*, 26(1):49–69, January 1996.

[DTL97] Veronica Dahl, Paul Tarau, and Renwei Li. Assumption Grammars for Processing Natural Language. In Lee Naish, editor, *Proceedings of the Fourteenth International Conference on Logic Programming*, pages 256–270, MIT press, 1997.

[DTRS98] Veronica Dahl, Paul Tarau, Stephen Rochefort, and Marius Scurtescu. Assumption Grammars for Knowledge Based Systems. *Informatica. An International Journal of Computing and Informatics*, 22(4), 1998. Special Issue on NLP and Agent Communication.

[Int] Intel. Moondo. http://www.intel.com/iaweb/moondo/index.htm.

[Son] Sony. Cyber Passage. http://vs.sony.co.jp/VS-E/vstop.html.

[Tar96] Paul Tarau. Logic Programming and Virtual Worlds. In *Proceedings of INAP96*, Tokyo, November 1996. Keynote Address.

[Tar97] Paul Tarau. BinProlog 5.75 User Guide. Technical Report 97-1, Département d'Informatique, Université de Moncton, April 1997. Available from *http://clement.info.umoncton.ca/BinProlog*.

[TDB96] Paul Tarau and Koen De Bosschere. Virtual World Brokerage with BinProlog and Netscape. In Paul Tarau, Andrew Davison, Koen De Bosschere, and Manuel Hermenegildo, editors, *Proceedings of the 1st Workshop on Logic Programming Tools for INTERNET Applications*, JICSLP'96, Bonn, September 1996. http://clement.info.umoncton.ca/lpnet.

[Wor] Worlds. AlphaWorld. http://www.worlds.net/products/alphaworld.

Efficient Computation Of Frequent Itemsets In A Subcollection Of Multiple Set Families

Hong Shen
School of Computing and Information Technology, Griffith University
Nathan, QLD 4111, Australia
AND
Weifa Liang
Department of Computer Science, Australia National University
Canberra, ACT 2600, Australia
AND
Joseph Ng
Department of Computer Science, Hong Kong Baptist University
Kowloon, Hong Kong

Keywords: Algorithm, data mining, frequent itemset.

Edited by: Rudi Murn
Received: March 26, 1998 **Revised:** September 16, 1998 **Accepted:** February 26, 1999

Many applications need to deal with the additive and multiplicative subcollections over a group of set families (databases). This paper presents two efficient algorithms for computing the frequent itemsets in these two types of subcollections respectively. Let T be a given subcollection of set families of total size m whose elements are drawn from a domain of size n. We show that if T is an additive subcollection we can compute all frequent itemsets in T in $O(m2^n/(pn) + \log p)$ time on an EREW PRAM with $1 \leq p \leq m2^n/n$ processors, at a cost of maintaining the occurrences of all itemsets in each individual set family. If T is a multiplicative subcollection, we can compute all itemsets in T in $O(mk/p + \min\{\frac{m'}{p}2^n, n3^n \log m'/p\})$ time on an EREW PRAM with $1 \leq p \leq \min\{m, 2^n\}$ processors, where $m' = \min\{m, 2^n\}$. These present improvements over direct computation of the frequent itemsets on the subcollection concerned.

1 Introduction

Given a family of sets T, e.g. a transaction database, each containing a set of items, a fundamental problem in data mining is to find all frequent (synonyms: large, interesting) itemsets in T with a support not smaller than a predefined *minimal support* (threshold), where the *support* of an itemset is the ratio of its frequency of occurrence in T to the size of T [2]. Centralized on the *apriori* approach [1], there have been various algorithms and parallel implementations proposed for this problem and its variants [8, 12, 5, 3, 9, 11, 6, 7, 4].

In many applications, we often need to compute the frequent itemsets *collectively* across a subcollection of several set families constructed in a well-defined way. For instance, community service in a large campus may need to know the consumption figure of a particular group of food across all shops on the campus by individual students. This requires an *additive* subcollection of the transaction databases of these shops. The traditional set *intersection* and *union*, with duplicates being counted, are two special cases of additive subcollection: intersection is the case when that group contains only one item, and union when the group covers all items in the whole database. On the other hand, a multi-category grade maintenance system in a university student administration system may require to list all students who have passed one course in each category satisfying a predefined course structure (patterns). This requires a *multiplicative* subcollection on all grade databases of individual courses. The standard relational-join in databases is a special case when the course structure covers all combinations of the courses from different categories. More formally, we have the following definition for these two types of subcollections.

Let $T_0, T_1, \ldots, T_{k-1}$ be k given set families, and R be a k-ary relation (e.g. k-parameter equation). Denote by $\oplus$ and $\otimes$ the operators for additive and multiplicative subcollections respectively. We define the additive subcollection $T_\oplus$ and multiplicative subcollection $T_\otimes$ on $T_0, T_1, \ldots, T_{k-1}$ as follows:

$$\begin{aligned} T_\oplus &= T_0 \oplus T_1 \oplus \cdots \oplus T_{k-1} \\ &= \{t_i \mid t_i \in T_i, 0 \leq i \leq k-1, \\ &\quad R(t_0, t_1, \ldots, t_{k-1}) holds\}. \qquad (1) \\ T_\otimes &= T_0 \otimes T_1 \otimes \cdots \otimes T_{k-1} \\ &= \{\langle t^0, t^1, \ldots, t^{k-1}\rangle \mid t^i \in T_i, 0 \leq i \leq k-1, \\ &\quad R(t^0, t^1, \ldots, t^{k-1}) holds\}. \qquad (2) \end{aligned}$$

In $T_\otimes$, each element is a k-tuple (sets) satisfying relation R. For any element $t_i = \langle t_i^0, t_i^1, \ldots, t_i^{k-1} \rangle$ in $T_\otimes$, we define $t \subseteq t_i$ iff $t \subseteq t_i^j$ for all $0 \leq j \leq k-1$.

In this paper, we consider an interesting problem of computing the frequent itemsets across a well-defined subcollection of additive and multiplicative types on a group of set families. We show how to compute these frequent itemsets efficiently by applying the relevant bit-vector operations. We organize the paper as follows. As the main technical body of the paper the next two sections present algorithms for computing the frequent itemsets in additive and multiplicative subcollections respectively. We conclude the paper in Section 4 with some open problems for future research.

2 Algorithm for the additive subcollection

As shown in [10], efficient parallel solution to the problem of finding all frequent itemsets in T_i requires to first compute the frequencies of all itemsets and then "filter" them according to the values of their frequencies. We now show how to apply this algorithm to the problem we are addressing and to produce efficient solution to our problem.

Let $T_\Sigma = T_0 \parallel T_1 \parallel \ldots \parallel T_{k-1}$, where "$\parallel$" is the operator of simple set concatenation. Assume that all elements of T_i are drawn from (itemset) *domain* I, i.e. $I = \cup_i T_i$, and let U contain all subsets generatable on I. Clearly U covers all possible subsets in any T_i and $|U| = 2^n - 1$. In our approach, we spend an extra space A of $|U||T|/d$ words, where d is the machine word-length, to record all frequent itemsets' occurrences in T. A is organized as an $|U| \times |T|$ bit-array, where $A[i][j] = 1$ if the ith itemset of U occurs in the jth element (set) of T, and $A[i][j] = 0$ otherwise. Computing A can be viewed as a precomputation which is invoked only once. We also assume that $T_\oplus = T_0 \oplus T_1 \oplus \ldots, \oplus T_{k-1}$ is given as input in the form of a bit-vector V of $|T|$ bits, where $V[i] = 1$ if $T[i] \in T_\oplus$ and $V[i] = 0$ otherwise. In most practical applications, $T_\oplus$ can be easily computed without incurring much cost because each T_i is usually stored in certain way of classification in the hard disk. The basic idea of our approach is to use these bit-vectors to reduce the computation.

Algorithm AddFrequentSets (A, V, L)
{*Input A and V, output L containing all frequent itemsets in $T_0 \oplus T_1 \oplus \ldots, \oplus T_{k-1}$.*}

for $i = 0$ **to** $|U| - 1$ **do**
1. $A'[i][0..|T|-1] := A[i][0..|T|-1] \wedge V[0..|T|-1]$;
2. Compute the number of 1's in $A'[i][0..|T|-1]$ and store it in $C[i]$;
3. **if** $C[i]/|T_\oplus| \geq \delta$ **then** $L := L \cup \{I_i\}$.

{*L is initialized to $\emptyset$.*}

We now analyze the time complexity of the algorithm. We assume that we are given a machine with word length d that can perform basic arithmetic and logic operations and also extracting the number of 1's in a single word in one step (constant time). The latter assumption is reasonable because extracting the number of 1's in a word should not be more difficult than performing an arithmetic/logic operation (e.g. multiplication). Step 1 of the algorithm requires $O(|U||T|/d)$ time since the logic AND operator"$\wedge$" is carried out word-by-word throughout all $|U||T|/d$ words in A. Step 2 requires also $O(|U||T|/d)$ time for word-by-word extracting the number of 1's and adding each row's together. Step 3 takes only constant time. Therefore the whole algorithm requires $O(|U||T|/d)$ time.

We say that a computation model is *conservative* if it has a word of $\Theta(\log N)$ bits for processing data of magnitude N. Clearly this word length is the minimum for processing such data as it is required to store each datum in a single word and process it in a single step. Since $|U| = 2^n - 1$ and $|T| = m$, in our case a conservative machine should have $\max\{\log|U|, \log|T|\} \geq n$ bits. Throughout this paper, unless otherwise stated, all the computation models are conservative. Thus we have

Lemma 1 *For a group T of set families of total size m drawn from an domain of size n, an additive subcollection of T, by spending an extra $m2^n/n$ space to maintain all subset occurrences in T we can compute all frequent itemsets in a given additive subcollection of T in $O(m2^n/n)$ sequential time.*

If we are given an EREW PRAM with $1 \leq p \leq m2^n/n$ processors, it is easy to see that Steps 1 and 3 in the algorithm can be completed in $m2^n/(pn)$ time, while Step 2 requires $O(m2^n/(pn) + \log p)$ by parallel summation. This results in the following theorem:

Theorem 1 *Given an EREW PRAM with $1 \leq p \leq m2^n/n$ processors, if all subset occurrences in T are known, we can compute all frequent itemsets in a given additive subcollection of a group of set families of total size m drawn from a domain of size n in $O(m2^n/(pn) + \log p)$ time.*

The following corollary is immediate:

Corollary 1 *Algorithm AddFrequentSets can be completed in $O(n)$ time on an EREW PRAM with $m2^n/n^2$ processors.*

From [1] we know that directly computing all frequent itemsets in $T_\oplus$ rather than taking advantage of the frequencies of all itemsets requires $O(m2^n)$ sequential time and $O(m2^n/p + n\log m)$ parallel time using p processors for any p. So our algorithm is more efficient than this direct approach.

Consider the case when all sets in T are distinct, which algorithm AddFrequentSets can easily accommodate by deleting all duplicate elements and attaching a counter of duplicates to each element in T. In this case $m \leq 2^n - 1$, where $m = 2^n - 1$ when T contains all possible subsets of U. In this worst scenario, we can narrow the upper bound of time complexity of the apriori further:

Lemma 2 *When all sets in T are distinct, the apriori requires time $\Theta(4^n/\sqrt{n})$ in the worst case, where n is the domain size from which all elements of T's are drawn.*

Proof

By [1], we know that the *apriori* approach works as follows. For $I = \bigcup T_i$ place all subsets of I in a lattice of $|I| + 1$ levels with source $\emptyset$ at level 0 and sink I at level $|I|$, where nodes at level i represent all subsets containing i elements — i-sets, and there is an edge from node x at level i to node y at level $i + 1$ if and only if $x \subset y$; starting from the source level-by-level compute the supports from T for all subsets at each level and delete those whose supports are smaller than δ together with all branches rooted at them (pruning).

We now modify the apriori approach to reduce the cost for computing the support of each subset in the lattice. Assume that the pruning process proceeds in "top-down" fashion from level 1 to level $n - 1$ as in the apriori, but computing the support of each subset is done in the way of "bottom-up" with the method to be shown below. We arrange all sets of T in cardinality increasing order into also n levels from 1-set to n-set. This can be done by simply sorting T by cardinality. We project T to the lattice by marking those subsets in the lattice which appear also in T, and associate a support counter with each marked subset of the lattice. We then compute the support of each subset x at level i from those at level $i + 1$ in the lattice by examining subset inclusion for $i = n - 1, n - 1, \ldots, 1$, rather than examining it with all sets in T as in the apriori. This is realized by projecting all marked subsets at level $i + 1$ to their adjacent subsets at level i by means of incrementing a subset's support counter by the value of the projected subset's support counter if it gets a projection, where an unmarked subset gets marked if it has an adjacent subset which is marked. Furthermore, for those subsets at level $i + 1$ that share a common adjacent subset at level $i + 2$, an amount of the sharing degree minus one should be deducted from the support counter of each subset at level i which they support, because this amount was the duplicate support counted when projecting these $(i+1)$-sets to i-sets. The above projection alone requires $\binom{n}{i+1}$ subset-inclusion examinations in the worst case when T contains all 2^n subsets constructed from I (including the extra empty-set at the source), since level $i + 1$ in this case contains all $\binom{n}{i+1}$ distinct $(i+1)$-sets constructed from I. Each such examination can be completed in $O(1)$ time with a machine word-length of n bits. Since there are $\binom{n}{i}$ subsets in level i for all $1 \le i \le n$, the total number of subset-inclusion examinations is given by the following formula:

$$\sum_{i=1}^{n-1} \binom{n}{i}\binom{n}{i+1} < \sum_{i=0}^{n} \binom{n}{i}^2 = \binom{2n}{n}. \qquad (3)$$

Applying the Stirling's formula, we can approximate this number to $(\frac{2}{\pi n})^{1/2} 2^{2n}$. This results in the lemma. □

Remark Note that the unmodified apriori would require simply $\Theta(4^n)$ time in the worst case. □

In the same worst case ($m = 2^n$), our algorithm requires $\Theta(4^n/(pn) + \log p)$ time, which is clearly still more efficient than directly using apriori even after the above modification.

3 Algorithm for the multiplicative subcollection

In this section we consider the problem of computing all frequent itemsets in a multiplicative subcollection of set families, and present an efficient algorithm for it.

Let $T_\otimes = T_0 \otimes T_1 \otimes \ldots \otimes T_{k-1} = \{t_0, t_1, \ldots, t_{m-1}\}$, where $t_i = \langle t_i^0, t_i^1, \ldots, t_i^{k-1}\rangle$ and $t_i^j \in T_j$, $0 \le i \le m-1$, $0 \le j \le k-1$. $T_\otimes$ can be generated by invoking some standard database query operation, *join* for instance. As stated in Section 1, we define $t \subseteq t_i$ iff $t \subseteq t_i^j$ for all $0 \le j \le k-1$.

Assume that $T_\otimes$ is given by the user. A straightforward solution to our problem is to apply an existing algorithm, e.g. the *apriori* [1] or subset statistics [10], to find all frequent itemsets on $T_\otimes$ directly. This would require in the worst case $O(2^n km) = O(k2^{(k+1)n})$ time by [1], since m can be as large as 2^{kn}. This is obviously too expensive even for very small k, and therefore is not applicable in practice. We shall present an efficient algorithm that runs in a time which is almost same as required for computing all frequent itemset on a single set family when m is not excessively large.

Our algorithm is based on two observations. First, with the definition of set inclusion on t_i, an itemset occurs in t_i iff it occurs in the intersection of all t_i's component-sets t_i^j. So examining the occurrence of an itemset in t_i is concluded to examining its occurrence in the corresponding intersection-set. Second, no matter how large m is, all the component-sets of t_i for all i have the same domain D of size n, so there can be at most 2^n different intersection-sets among all t_i's. Therefore we need only to consider at most 2^n such sets while taking the number of duplicates of each set into account when counting the occurrences of each set.

We use an $m \times k \times n$ bit-array T to represent $T_\otimes$, where $T[i][0..k-1][0..n-1]$ represents t_i, $T[i][j][0..n-1]$ represents t_i^j. $T[i][j][h] = 1$ indicates that t_i^j contains the hth element in the domain – $D[h]$, $T[i][j][h] = 0$ indicates $D[h] \notin t_i^j$. Our algorithm is described as follows:

Algorithm MultiFrequentSets (T, D, L)
{*Input T representing $T_0 \otimes T_1 \otimes \ldots, \otimes T_{k-1}$ and itemset domain D, output L containing all frequent itemsets in T.*}

1. **for** $i = 0$ **to** $m - 1$ **do**
   for $j = 0$ **to** $k - 1$ **do**
   $T'[i][0..n-1] := T'[i][0..n-1] \wedge$
   $T[i][j][0..n-1]$;

{*Compute $T'[i][0..n-1]$ representing $t'_i = t_i^0 \cap t_i^1 \cap \cdots \cap t_i^{k-1}$, where $T'[i][0..n-1]$ is initialized to 1, $0 \leq i \leq m-1$.*}

2. **for** $i = 0$ **to** $m-1$ **do**
   Distribute $T'[i][0..n-1]$ to the bucket with index of value $T'[i][0..n-1]$;
   {*There are 2^n buckets, each corresponding a subset on D.*}

3. **for** $i = 0$ **to** $2^n - 1$ **do**
   if the ith bucket is not empty **then**
   Collect the contents of this bucket to T'' and keep the number of duplicates (size of the bucket) in dup_i;
   {*T'' is an $m' \times n$ bit array containing all distinct sets of T', $m' \leq \{m, 2^n\}$.*}

4. **for** $i = 0$ **to** $m' - 1$ **do**
   $D'[0..n-1] := D'[0..n-1] \vee T''[i][0..n-1]$;

5. $D'[0..n'-1] := D'[0..n-1]$ after deleting all 0 bits in $D'[0..n-1]$;
   {*D' initialized to 0, is the itemset domain of T'', $n' \leq n$.}

6. Compute all frequent itemsets of T'' on D', where an itemset t's occurrence in any set $T''[i][0..n-1]$ is dup_i instead of 1 as usual counting in [1], if $t \subseteq T'[i][0..n-1]$.

The correctness of the algorithm can be seen from the comments interspersed with the algorithm. We proceed with analysis of its time complexity. Assume that we are given a conservative machine with word length of n bits. Step 1 of the algorithm takes $O(mk)$ time. Steps 2 and 3 require time $O(m)$ and $O(m + 2^n)$ respectively. Steps 4 and 5 require time $O(m') = O(\min\{m, 2^n\})$ and $O(n)$ respectively. Step 6 can be done either in $O(m'2^{n'}) = O(m'2^n)$ time using the algorithm of [1], or in $O(n'3^{n'} \log m') = O(n3^n \log m')$ time using the result of [10], where $m' \leq \min\{m, 2^n\}$. So we have

Lemma 3 *Given a multiplicative subcollection $T_\otimes$ of size m of k set families on a domain of size n, all frequent itemsets in $T_\otimes$ can be computed in $O(mk + \min\{m'2^n, n3^n \log m'\})$ time sequentially, where $m' = \min\{m, 2^n\}$.*

In the parallel environment, assume that we are given an EREW PRAM with $1 \leq p \leq \min\{m, 2^n\}$ processors. Then Step 1 can be done in $O(mk/p)$ time. Steps 2 and 3 need $O(1)$ and $O(\log m\lceil 2^n/m \rceil)$ time respectively. Steps 4 and 5 take $O(m'/p + \log m')$ and $O(\log n)$ time respectively. Step 6 requires $O(\frac{m'}{p}\sum_{i=0}^{n'-1}\binom{n'}{i}) = O(\frac{m'}{p}2^{n'})$ time by [1], since at level $i+1$ of the lattice [10] each processor holding a block of m'/p $(i+1)$-sets of T'' needs to examine each of these sets with all $\binom{n'}{i}$ i-sets at level i for subset inclusion. This step can also be done in $O(n'3^{n'} \log m'/p)$ time by [10]. Hence we have the following theorem:

Theorem 2 *Given a multiplicative subcollection $T_\otimes$ of size m of k set families on a domain of size n, we can compute all frequent itemsets in $T_\otimes$ in $O(mk/p + \min\{\frac{m'}{p}2^n, n3^n \log m'/p\})$ time on an EREW PRAM with $1 \leq p \leq \min\{m, 2^n\}$ processors, where $m' = \min\{m, 2^n\}$.*

4 Concluding remarks

Given a group of set families (databases), two types of subcollections of them encountered in practical applications, the additive subcollection and multiplicative subcollection, have been studied. This paper addresses the problem of how to compute all frequent itemsets whose occurrences in these two types of subcollections exceed a predefined threshold. We have proposed two efficient algorithms. The first algorithm computes all frequent itemsets in additive subcollection that uses bit-vector techniques and takes advantage of the occurrences of all itemsets in each individual set family. The second algorithm computes all frequent itemset in the multiplicative subcollection by first computing the intersection of all component-sets of each set and then computing the frequent itemsets on the resulting set of intersection sets after necessary simplification. Both algorithms are considerably more efficient than their counterparts of straightforward approaches to solving these two problems respectively, in both sequential and parallel environments.

Future research includes studies on efficient algorithms for finding all frequent itemsets in other types of collections and combinations of set families.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Jorgeesh Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *20th International Conference on Very Large Data Bases, September 12–15, 1994, Santiago, Chile proceedings*, pages 487–499, Los Altos, CA 94022, USA, 1995. Morgan Kaufmann Publishers.

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 22(2):207–216, June 1993.

[3] Luc Dehaspe and Luc De Raedt. Mining association rules in multiple relations. In Nada Lavrač and Sašo Džeroski, editors, *Proceedings of the 7th International Workshop on Inductive Logic Programming*, Volume 1297 of *LNAI*, pages 125–132, Berlin, September 17–20, 1997. Springer.

[4] A. A. Freitas and S. H. Lavington. Parallel data mining for very large relational databases. *Lecture Notes in Computer Science*, 1067:158–163, 1996.

[5] Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, and Takeshi Toluyama. Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 25(2):13–23, 1996.

[6] Eui-Hong Han, George Karypis, and Vipin Kumar. Scalable parallel data mining for association rules. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 26(2):277–288, 1997.

[7] Andreas Mueller. Fast sequential and parallel algorithms for association rule mining: A comparison. Technical Report CS-TR-3515, University of Maryland, College Park, August 1995.

[8] Jong Soo Park, Ming-Syan Chen, and P. S. Yu. An effective hash-based algorithm for mining association rules. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 24(2):175–186, June 1995.

[9] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB '95: proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, Sept. 11–15, 1995*, pages 432–444, Los Altos, CA 94022, USA, 1995. Morgan Kaufmann Publishers.

[10] H. Shen. Fast parallel subset statistics and its applications in data mining. Technical report, 1998.

[11] R. Srikant and R. Agrawal. Mining generalized association rules. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB '95: proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Switzerland, Sept. 11–15, 1995*, pages 407–419, Los Altos, CA 94022, USA, 1995. Morgan Kaufmann Publishers.

[12] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. *Proceedings of 1996 ACM SIGMOD International Conference on Management of Data*, pages 1–12, Motreal, Canada, 1996.

SISTER: A Flexible System For Image Retrieval

Monica Mordonini and Agostino Poggi
Dipartimento di Ingegneria dell'Informazione, University of Parma
Parco Area delle Scienze, 181/A,
43100- Parma, Italy
phone +39 10 905728, fax +39 10 905723, E-mail poggi@CE.UniPR.IT

Keywords: image retrieval by content, pictorial information, inductive classifier, pattern matching, image database management system

Edited by: Rudi Murn
Received: December 23, 1998 **Revised:** May 15, 1999 **Accepted:** September 14, 1999

SISTER is a system for the storing and retrieval of large collections of images on the basis of both textual information and the image content. In particular, SISTER allows the user to formulate queries for different image categories on the basis of color information and specific attributes of the image category; this is possible, because such a system can be easily adapted to support new image categories specializing the acquisition and the retrieval subsystems. SISTER is composed of three parts: i) an image acquisition subsystem automatically extracting the attributes from images, ii) a database management subsystem maintaining the description of images, and iii) a retrieval subsystem allowing the user to retrieve images through a user-friendly graphical interface. The acquisition subsystem extracts image attributes combining image processing and inductive classification modules. Classification modules allow the computation of image attributes that cannot be directly extracted through image processing techniques, but whose value derives from the value of some other image attributes extracted by the acquisition subsystem.

1 Introduction

Since the early 90's the research in Image Retrieval based on their content is going to grow together with the demand of inquire large image and video collections. In fact the planning and the development of a search system that effectively operates on large collections of multimedia data is a complex task: the possibility to use the information contained in the data is closely correlated to their organization that has to allow an efficient phase of browsing and search through the database. Mainly the queries on multimedia data can be subdivided in two types: queries by textual and visual keys. Many systems in Image Retrieval use search techniques based on the association of textual keys with the content of each image and each video belong to the database: often these keys are not the most appropriated to describe the information searched by the user and, moreover, it can't be possible to apply a manual annotation approach to a large scale image collections. These facts led to put the researchers attention on the image or video itself, that is on the original information contained in its file description. These techniques, commonly called content-based retrieval techniques, analyze and classify the images on the base of their visual content (such as space color, shape, texture, spatial relations between the objects, regions of interest and changes of scene) [27]. The interest in the advance in this research activity is been evidenced by the many "special issues" that the leading journals in Database Management and Computer Vision have been dedicated to the topic (see, for example, [13, 23, 20, 29, 15]). Various prototypes of Image Retrieval systems have been realized both in research and commercial area; in most of them the queries and data retrieval are carried out through one or more of the following options [5]:

- random browsing
- search by example
- search by sketch
- search by text
- navigation with customized image categories

Among the most known architectures of Image Retrieval we can remind:

- QBIC (Query By Image Content), developed at the IBM Almaden Research Center, is the first commercial system of Image Retrieval on the basis of their visual content. It allows the queries in term of color (disposition and percentage), texture, visual examples [9]; this system was employed as search engine of a databases prototype for art images retrieval on the basis of their visual content rather than on their textual description [14].

- Virage, developed at the Virage Inc., allows the search by means of color, composition, texture and structure like QBIC. Moreover, it is able to act composite interrogations on the basis of an arbitrary combination of the previous "atomic" queries [1].

- RetrievalWare, developed at the Excalibur Technologies Corp., supports queries from texture, brightness, color. It places in particular evidence the use of adaptive techniques, like Neural Networks, to carry out the visual query [8].

- Photobook, developed at the MIT Media Lab., is based on a wide set of tools for image processing and browsing techniques, often interactive with the end-user, that simplify the search for images [25, 18, 26].

- VisualSEEk, developed at the Columbia University, supports queries based on visual features and their spatial relations. The visual features used in this system are mainly based on the color and the use of wavelet applied to the texture [31].

- Netra, developed in the project UCSB Alexandria Digital Library (ADL), extracts information by color, shape, texture in various regions of the image in order to find similar regions in the database [19].

- MARS (Multimedia Analysis and Retrieval System), developed at the University of Illinois at Urbana-Champaign, would represent a possible system which is able to carry out the integration between Database Management Systems and Information Retrieval. The research is focused to how organize various visual features to obtain an architecture that can dynamically accommodate to different applications and different users [21, 22, 28].

Other systems of multimedia data classification and content-based image retrieval can be found in [7, 32, 6, 11, 3, 12, 16, 30].

None of the known systems for content-based image retrieval is built to be specialized for the retrieval of images using queries based on a set of specialized attributes for each different image category and to automatically extract those attributes. In this paper, we present a system, called SISTER, for the storing and retrieval of images that allows the automatic extraction of image features and the formulation of specific queries for different image categories and improves image attributes acquisition combining image processing and inductive classification modules. The next section introduces the structure of the system. Section three presents two case studies, that is, the specialization of the system for fashion sketches and for photo portraits. Finally, section four presents the main contributions of our work and our future research direction.

2 SISTER

SISTER (System for Image SToragE and Retrieval) is a system for archiving images based on an image database management system and two subsystems for the acquisition and retrieval of images (see figure 1).

SISTER offers the following features: i) an easy retrieval of images, because the user can ask about image attributes through a user-friendly graphical interface and the attributes he/she can ask depend on the image category, that is, each image category (e.g., landscapes, portraits, ...) has its set of attributes; ii) a fast retrieval of images, because image attributes are extracted from images in the acquisition phase and not in the retrieval phase; iii) an automatic acquisition and storing of images, because a large part of image attributes are not manually extracted, but by the acquisition subsystem; iv) an easy management of different categories of images, because the system allows an easy specialization of both the acquisition and retrieval subsystems to manage new image categories.

2.1 Image Acquisition Subsystem

The acquisition subsystem automatically extracts attributes from images and stores them into a database. The architecture of the acquisition subsystem is displayed in figure 2. Such a subsystem has a pipe-and-filter organization connecting a set of image processing modules grouped in three different levels of computation. The first level extracts general attributes from the image. The second level is different for each image category and extracts specific attributes. Finally, the third level collects general and specific attributes and stores them in a system database (there is a database for each image category).

The first level of computation extracts automatically color information from images. In particular, it extracts:

- a color histogram for the 256 colors of the image GIF palette (in terms of hue, saturation and brightness), and for each color:

- a qualitative value of the centre of the color area. The possible values are: up (i.e., the centre in the half upper part of the image), left, down, right, middle, horizontal-middle, vertical-middle, left-up, right-up, left-down and right-down;

- a qualitative value of the variance, indicating if the color is concentrated in a part of the image or it is distributed in all the image. The possible values are: concentrated, low-concentrated, scattered and high-scattered;

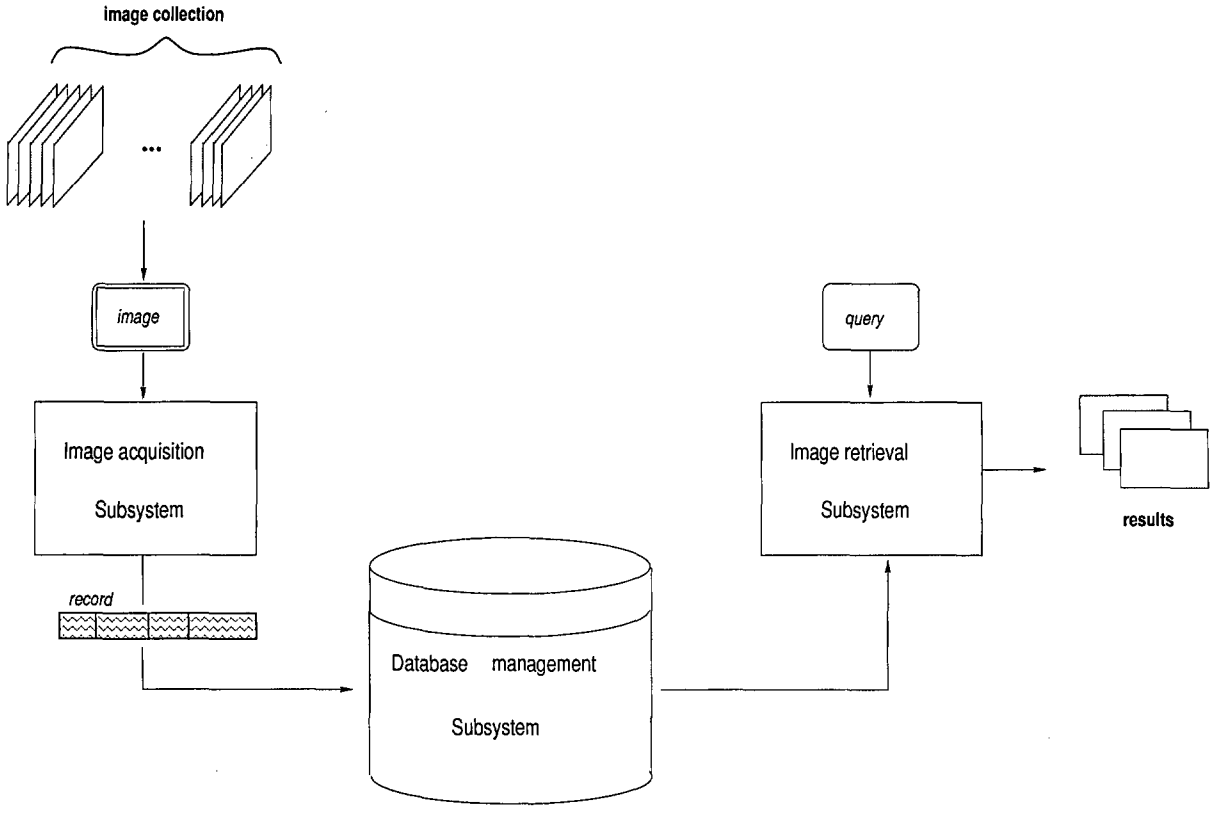


Figure 1: SISTER architecture.

– a qualitative value of the granularity, indicating if the color fills a limited or a wide number of image areas. The possible values are: high, medium and low.

Given that specific attributes are obviously different for each category of image, each category of image needs a different second level of computation. Each of such levels can be easily built by composing i) some image processing modules offered by the system; ii) some new image processing modules and sometimes iii) some classification modules built through a rule extraction module.

Classification modules are useful to extract some attributes that can be induced or at least predicted by the value of other attributes (for example, an elongated horizontally light blue area in the upper half of an image suggests the presence of sky). Such classification modules are based on a sequence of "if-then" rules that are built by a rule extraction module based on an inductive algorithm working on a set of training examples [2].

2.2 Image Database Management Subsystem

The image database management subsystem can use different relational database management systems running on different machines; in particular, the current implementation uses ORACLE and MySQL relational database management systems running on two machines one to maintain fashion sketches images data and the other to maintain photo portrait images data.

The use of a different database for each image category allows to guarantee good retrieval response times. The use of relational databases is not a limit because image elaboration is performed in the acquisition phase and so a database management system more suitable to do elaboration on database elements like an object-oriented database system is not necessary. Moreover, in our case, a relational database management system allows a simpler development of both the data models for the different image categories and the SQL queries to retrieve images from the databases.

2.3 Image Retrieval Subsystem

The image retrieval subsystem is based on two graphical user interfaces, called selection and presentation interface and color query interface.

The selection and presentation interface is composed of a module for user profile management and for image category selection and a module for query construction and results presentation (see figure 3).

The first module allows the users to keep trace of their previous queries in each image category database and to adapt the interface to the chosen image family. Each image class has, in fact, an own set of image attributes.

The second module allows the definition and the execu-

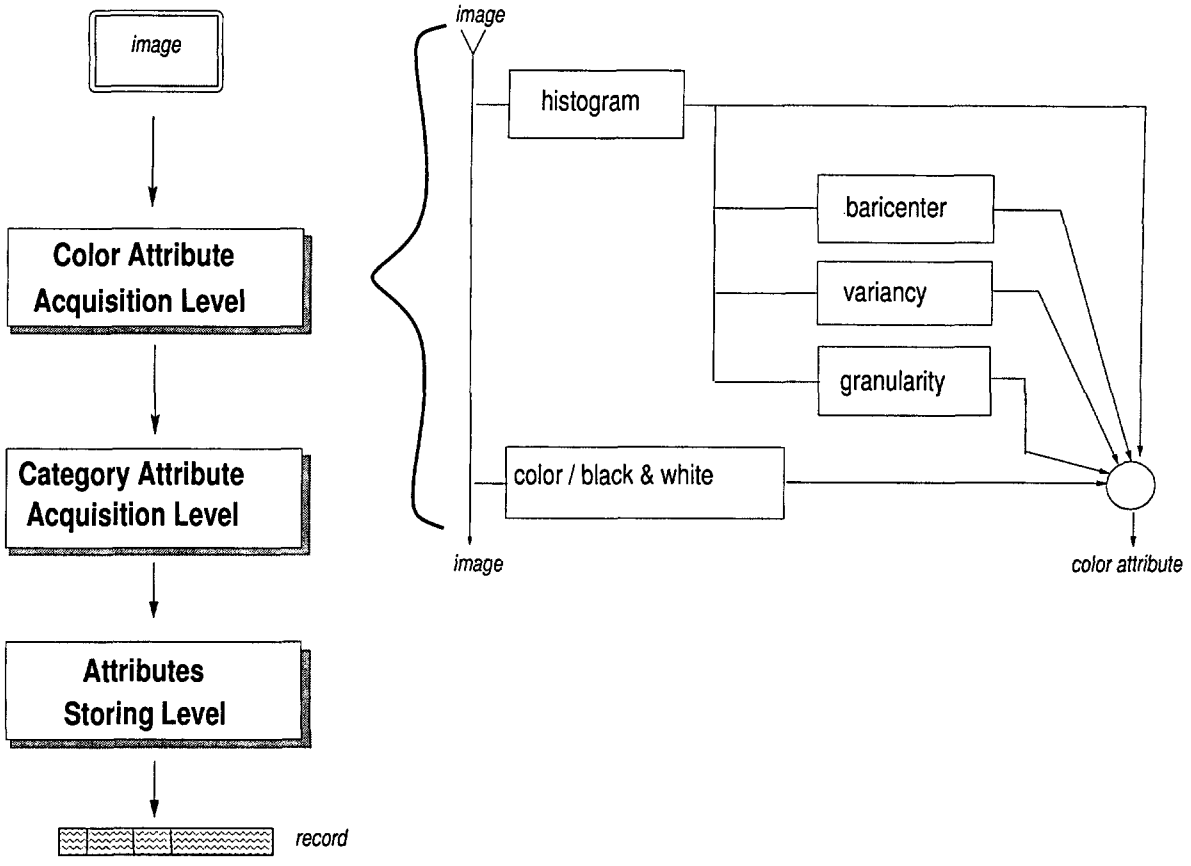


Figure 2: Image acquisition subsystem.

tion of attribute queries and the calling of the color query interface. Each image attribute can assume one among a finite set of values, therefore, the interface presents a sequence of drop-down-list windows that allow a user to select the value for each attribute among the possible values.

The color query interface (see figure 4) is independent from the chosen image category and allows users to ask about the presence of one or more colors in the image optionally giving for each color: the area filled by the color in the image, a range for the value of color and its area a qualitative value for the position of the baricenter, for variancy and for granularity.

Finally the results are presented in the main window of the selection and presentation interface and the user can take a look at them and at the attribute values of the central image.

2.4 Implementation

SISTER has been implemented by taking advantage of object-oriented programming features and, in particular, by using C++ and Java languages. The acquisition subsystem and its interface towards the database are implemented in C++. The modules of the acquisition subsystem are C++ objects built on the top of the ImageMagick image processing library [1]. Such a realization allows: i) good performance because it is based on a well-known and optimized image processing library and ii) good reuse of software because new modules can be obtained as specialization of pre-existent modules.

The image retrieval subsystem is implemented in Java. The software is based on a limited set of classes; in particular, each attribute shown in the attribute query interface is managed by an instance of the same class that displays the corresponding drop-down-list window and, during the query definition, builds the corresponding piece of SQL query. Such a solution has two important advantages: i) the possibility of an easy remote access through a WWW browser supporting Java, because the interface is defined as a Java applet, and ii) an easy specialization for different image categories because a new attribute causes the introduction of a new instance of the same class with initialization parameters, the attribute name and the list of possible values, that is, the declaration of the attributes and of their possible values.

[1] ImageMagick image processing library is available from http://www.wizards.dupont.com/cristy/ImageMagick.html.

Figure 3: Selection and presentation interface.

3 Experimentation

We experimented SISTER with two different image categories: fashion sketches and photo portraits. The fashion sketches database contains a thousand of images acquired from the original sketches of the "Sorelle Fontana" atelier. The photo portraits database contains about 300 of images directly acquired from a camera, images from the Vision and Modeling Group at the MIT Media Lab and images acquired via a scanner from an old collection of photo portraits (Cattani's collection 1927-1948).

In the case of fashion sketches, we specialized the acquisition subsystem to extract information about the dress type, the dress length and the sketch drawer (see figure 5).

The value of the first two attributes are computed through two simple image processing algorithms able to distinguish long from short dresses and entire from broken dresses, that is, dresses composed of a jacket and a skirt with different colors and/or texture. The last attribute, the drawer, is computed by combining a classification module with a pattern matching module. This is possible because, each drawer has an own style of sketching that allows an expert to recognize him/her from the other drawers working for the atelier. In particular, the 12 drawers of our "Sorelle Fontana" database can be partially classified through the value of some attributes that can be easily found by the system image processing modules: 1) colored or black and white sketch, 2) the background texture, 3) the color of eyes, and 4) the color of the mouth.

Starting from a training set of a hundred of images we built the classification module that we used for finding the drawer of the other images of the database. The result of the classification was a 15% average error because the limited set of attributes we used could not always discriminate between the different drawers.

Drawers can be often recognized from the shapes they use for some details as, for example, eyes and mouth.

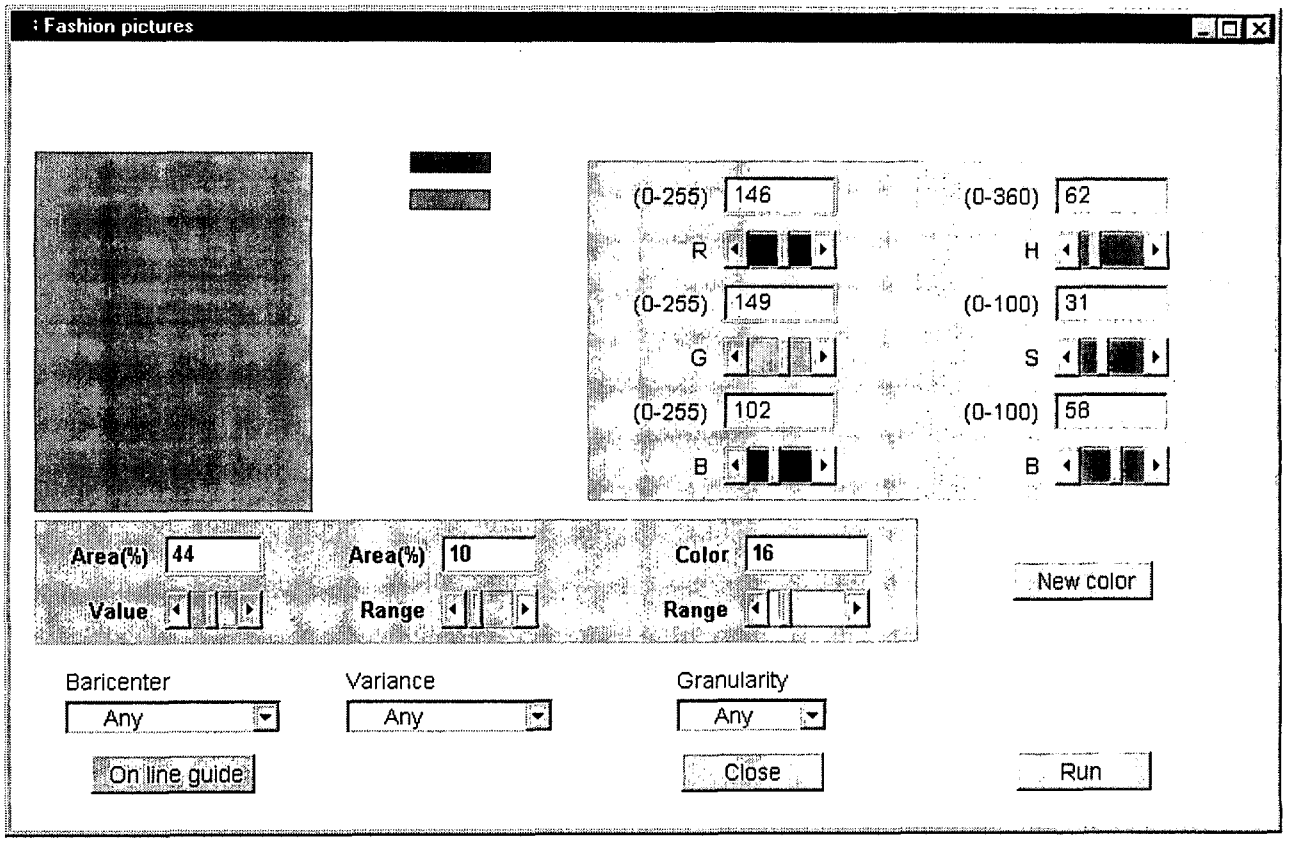


Figure 4: Color query interface.

Shape information cannot be easily used by the classification module, therefore, we used the classification module to determinate a subset of possible drawers and then we used a pattern matching module based on Freeman's chains [10] to compare the contour of the mouth of the current sketch with the typical contours of the mouths designed by the drawers selected by the classifier. In this case, the average error is reduced to 2%[2].

Figure 6 shows an example of query to the fashion sketches database. The user can choose the attributes in the keyword list placed on the right of the image presentation area and then clicks the query button. The status of the query and its result is indicated on the top area, whereas in the image presentation area the first seven images that have the chosen attributes are reported. In the example illustrated in figure 6, the user selected the drawer name, the dress length and the dress type from the keyword list. The system found 15 images that have these attributes and they could be visualized in the image presentation area by clicking the next and the precedent buttons or on the image themself.

In the case of photo portraits, we specialized the acquisition subsystem to extract the presence of mustache, beard and pointed beard, the type of forehead, the presence and the type of glasses and the hair color (see figure 7).

The extraction of such attributes was simplified because we reuse some image processing modules developed for the fashion sketch images. For example, the module used to identify the head for fashion sketches images has been used with success for such a kind of images too.

Figure 8 shows an example of query to the photo portraits database of the Vision and Modeling Group at the MIT Media Lab. The query interface is the same of the fashion sketches interface. In the example illustrated in figure 8, the user selected the glass presence and the color of the hair from the keyword list. The system found 8 images that have these attributes and visualized the first seven in the image presentation area.

4 Conclusions

In this paper, we present a system, called SISTER, for the storing and retrieval of images. SISTER allows the user to formulate specific queries for different image categories and the automatic extraction of image attributes through image processing and classification modules.

Sometimes, classification modules are useful because they may allow the extraction of attributes that cannot be directly extracted through image processing techniques and because they may allow the reduction of the error percentage made by the image processing modules in the acquisi-

[2]We also experiment the pattern matching module without the classification module; the result was longer execution time because the matching is performed on the mouths of all the drawers and a 10% average error.

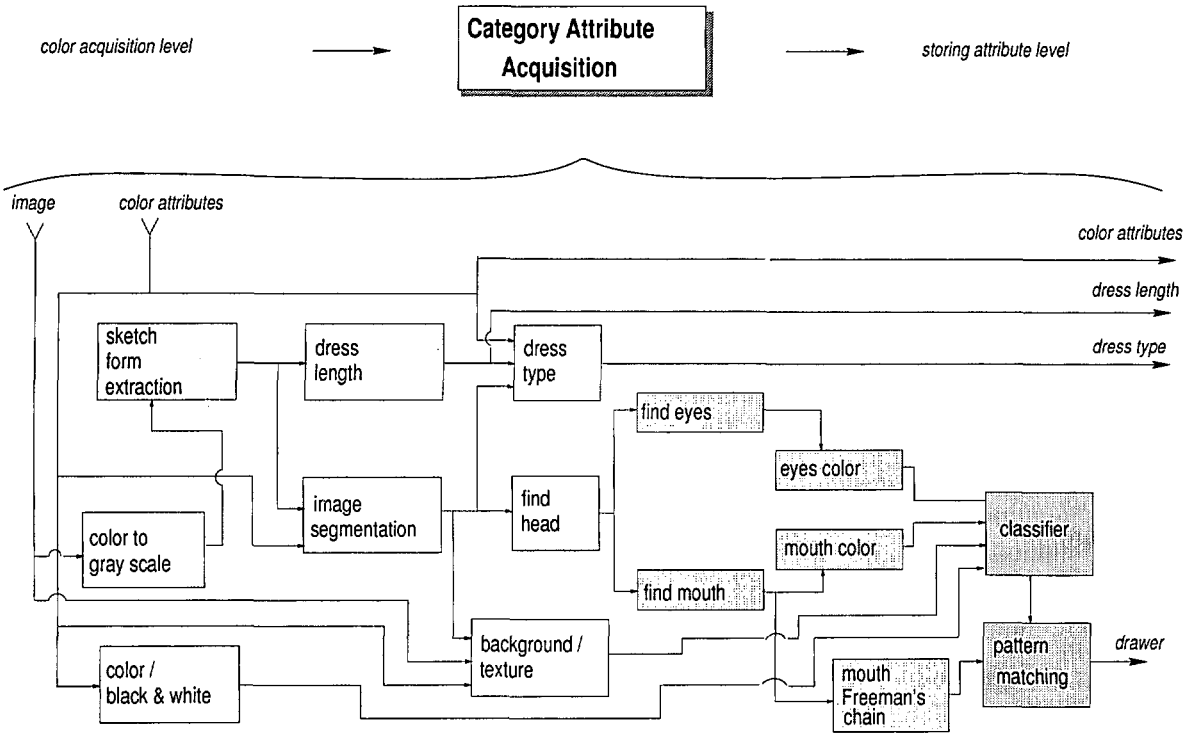


Figure 5: Fashion sketches attribute acquisition level.

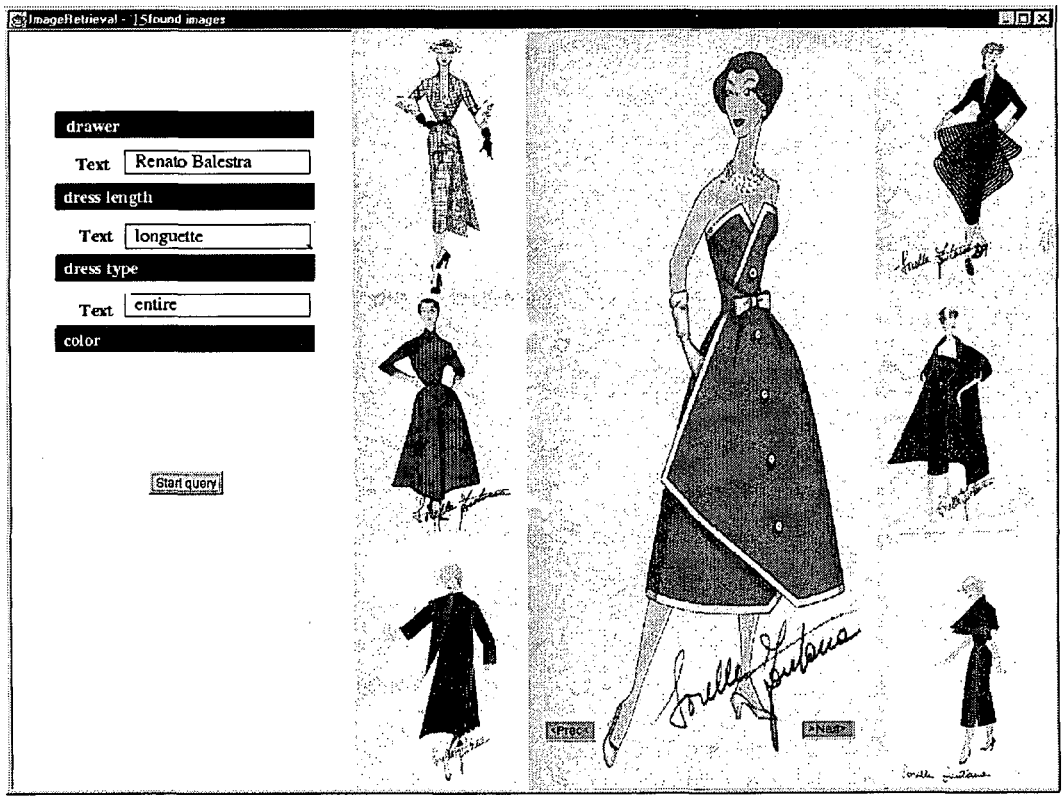


Figure 6: Example image of a query to the fashion sketches database.

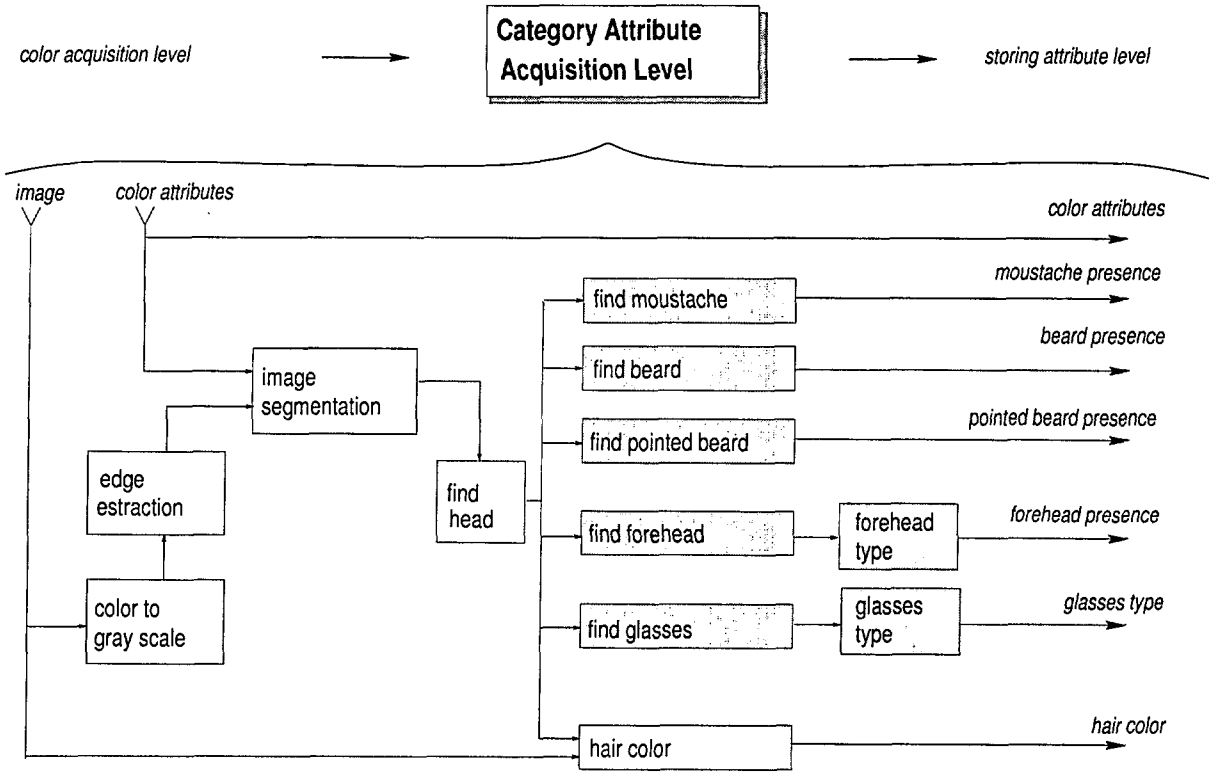


Figure 7: Photo portraits attribute acquisition level.

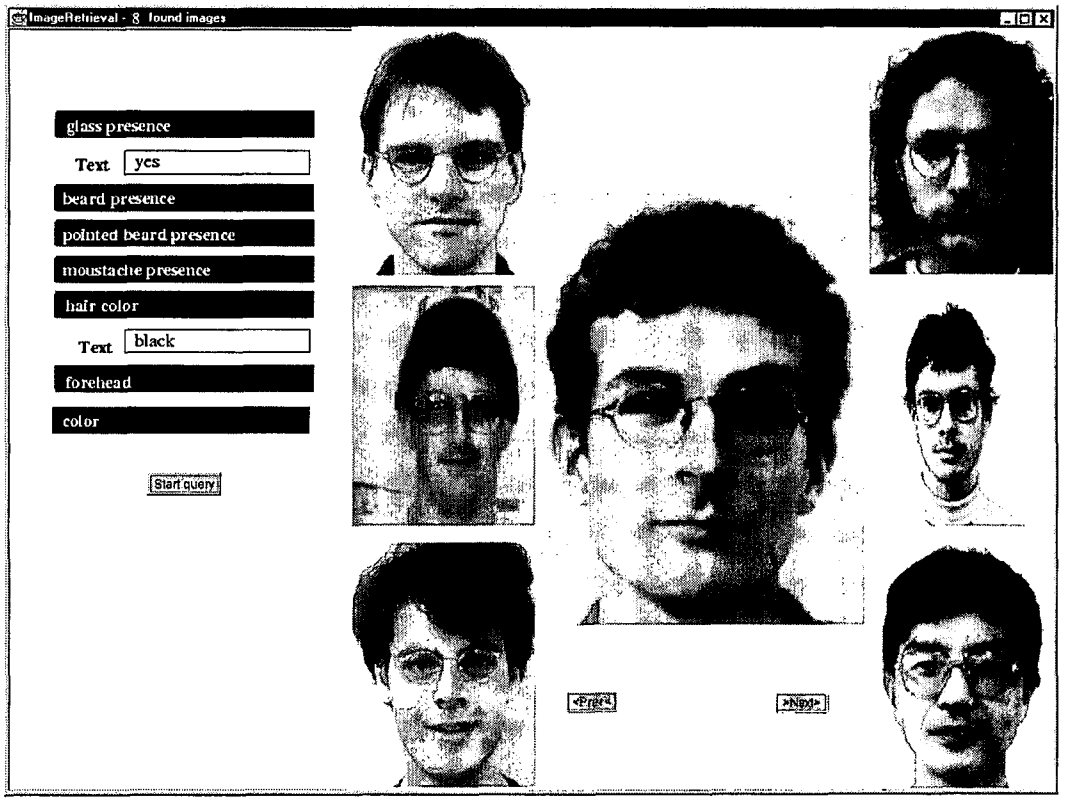


Figure 8: Example image of a query to the photo portraits database of the Vision and Modeling Group at the MIT Media Lab.

tion of image attributes.

SISTER can be easily specialized to manage different image categories. The acquisition subsystem can be adapted to process a new image category connecting some predefined and/or new image processing modules to the color attributes acquisition level. The retrieval subsystem can be adapted by simply declaring the specific image category attributes and their possible values.

We experimented the system with two different image categories: fashion sketches and photo-portraits. The development of their acquisition subsystems required few weeks of work of a student, while the development of the retrieval subsystem required few minutes. Moreover, the retrieval subsystem is very easy to use even by people without any knowledge on computers, in fact, a test performed on twenty of such persons shows that all of them are able to used it after few minutes.

Our current research directions are: the experimentation of the system with other image categories (landscapes and sport images), the introduction of a module for face recognition based on eigenfaces [24] for the photo portrait database, and the development of a visual environment (like AVS [4] and Khoros [17]) for the realization of the acquisition subsystems.

5 Acknowledgements

This work has been partially supported by the Italian National Research Council (CNR) through "Progetto Finalizzato Beni Culturali".

References

[1] A.Gupta and R.Jain. Visual information retrieval. *Comm. of ACM*, 40(5):70–79, 1997.

[2] D.T. Pham amd M.S. Aksoy. RULES: a simple rule extraction system. *Expert Systems with Applications*, 8(1):59–65, 1995.

[3] A. Del Bimbo and P. Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Trans. on Patt. Anal. Mach. Int.*, 19(2):121–132, 1997.

[4] J. Caldwell, L. Goldberg, and H. Lord. Overview of AVS5. In *Proc. of the 2nd International AVS User Group Conference*, Orlando, Florida, 1993. AVS '93.

[5] N.S. Chang, A. Eleftheriadis, and R. McClintock. Next-generation content representation, creation and searching for new media applications in education. In *IEEE Proceedings*, 1998. In Press.

[6] P. Charlton and B. Huet. Using multiple agents for content-based image retrieval. Technical report, Westminster University, London, 1995.

[7] C.Jacobs, A. Finkelstein, and D.H. Salesin. Fast multiresolution image querying. In *Proc. SIGGRAPH'95*, Los Angeles, 1995.

[8] J. Dowe. Content-based retrieval in multimedia imaging. In *Proc. SPIE Storage and Retrieval for Image and Video Database*, 1993.

[9] M. Flickner et al. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[10] H. Freeman. Computer processing of line drawing images. *ACM Comput. Surveys*, 6(1):57–97, 1974.

[11] S.I. Gallant and M. F. Johnston. Image retrieval using image context vectors: first results. In *Proc. SPIE Symp. on Electronic Imaging: Science Technology*, San Jose, CA, 1995.

[12] T. Gervers. Pictoseek: A content-based search system for the world wide web. In *Proc. of VISUAL'97*, San Diego, CA, 1997.

[13] V. N. Gudivada and J. V. Raghavan. Special issue on content-based image retrieval systems. *Computer*, 28(9), 1995.

[14] B. Holt and L. Hardwick. Retrieving art images by image content: the uc davis qbic project;. *Proc. of ASLIB*, 46(10):243–248, 1994.

[15] R. Jain. Special issue on visual information management. *Comm. of ACM*, 40(12), 1997.

[16] M.L. Kersten and M.A. Windhouwer. A feature database for multimedia objects. In *Proc. of ERCIM DBRG*, Schloss Birlinghoven, Germany, 1998.

[17] K. Konstantinides and J. Rasure. The Khoros Software Development Environment for Image and Signal Processing. *IEEE Journal of Image Processing*, 1993.

[18] F. Liu and R.W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. Patt. Recog. and Mach. Intell.*, 18(7):722–733, 1996.

[19] W.Y. Ma and B.S. Manjunath. Netra: A toolbook for navigating large image databases. *Proc. IEEE Int. Conf. on Image Processing*, 1997.

[20] A.D. Narasimhalu. Special section on content-based retrieval. *Multimedia Systems*, 3(1), 1995.

[21] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and T. Huag. Supporting similarity queries in MARS. In *Proc. of ACM Multimedia*, 1997.

[22] M. Ortega, Y. Rui, S. Mehrotra, and T. Huag. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Trans. on Circuit and System for Video Technology*, 8(5):644–655, 1998.

[23] A. Pentland and R. Picard. Special issue on digital libraries. *IEEE Trans. Patt. Recog. and Mach. Intell.*, 18(8), 1996.

[24] A. Pentland and M. Turk. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1993.

[25] R.W. Picard and A. Pentland. Photobook: Content-based manipulation of image database. In *Proc. SPIE Storage and Retrieval Image and Video Database II*, pages 34–37, San Jose, CA, 1994.

[26] R.W. Picard, A. Pentland, and S. Sclaroff. Photobook: Content-based manipulation of image database. *Int. Journal of Computer Vision*, 18(3):233–254, 1996.

[27] Y. Rui, T. Huang, and S. Chang. Image retrieval: past, present, and future. *Journal of Visual Communication and Image Representation*, 1998. In Press.

[28] Y. Rui, T. Huang, and S. Mehrotra. Browsing and retrieving video content in a unified framework. In *Proc. of IEEE MMSP'98 Workshop*, 1998.

[29] B. Schatz and H. Chen. Building large-scale digital libraries. *IEEE Computer*, 1996.

[30] G. Sheikholeslami, W. Chang, and A. Znang. Semantic clustering and querying on heterogeneous features for visual data. In *Proc. of ACM Multimedia*, Bristol, U.K., 1998.

[31] J.R. Smith and S.F. Chang. Visualseek: A fully automated content based image query system. In *Proc. of ACM Multimedia*, 1996.

[32] V.Ogle and M.Stonebraker. Chabot: Retrieval from a relational database of images. *IEEE Computer*, 28(9):40–48, 1995.

Tuning of Fuzzy Logic Controller with Genetic Algorithm

Borut Zupančič, Marko Klopčič and Rihard Karba
University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia
Phone: +386 61 1768 417, Fax: +386 61 1264 631
E-mail: borut.zupancic@fe.uni-lj.si

Keywords: control system, fuzzy logic controller, optimization, genetic algorithm

Edited by: Anton P. Železnikar
Received: January 6, 1999 **Revised:** July 22, 1999 **Accepted:** September 13, 1999

The paper deals with the fuzzy logic controller (FLC) tuning by the aid of the optimization with genetic algorithm (GA). Because of the complexity of calculations the Sugeno 0th order FLC was used. The GA optimization tuned 25 consequent parameters while the membership functions remained fixed. Because of large number of parameters traditional optimization methods were not successful. Beside appropriately selected GA parameters the choice of appropriate reference signal of the control system was extremely important. Namely it must be selected so that the whole ranges of both FLC inputs are used. It is recommended to plot trajectory of FLC inputs to see which parts of the truth table is appropriately covered by the inputs and to find out which consequent parameters can not be optimized. Filtering of FLC characteristic is another useful method, which makes output characteristics smoother and so improves responses. The efficiency of the proposed approach were verified and validated on a hydraulic control system.

1 Introduction

The life cycle of a control system demands several optimizations in several steps. Probably the most demanding steps are those in the phases of process modelling and controller design. Sometimes these optimizations are very simple, based on experiences, tunning rules or simulation trials [3, 4]. Sometimes better result are obtained by conventional optimization techniques [1]. This approach is extremely important for control systems with lower number of loops and with so called parametric controllers, e.g. traditional PID controllers. Such optimization is extremely efficient when no more than app. 10 parameters must be tuned. However more advanced control algorithms contain usually much more parameters, which must be appropriately tuned. More complex control algorithms result in better efficiency, when systems are complex, nonlinear or time varying, multivariable, highly oscillating, with significant delays etc. Conventional optimization algorithms are not able to properly handle such problems, so there is a constant search for new and better methods. In this search researchers also started looking at how nature and people handle similar problems. Such thinking led to fuzzy logic and artificial neural nets as important elements for advanced control algorithms and genetic algorithms as new robust optimization techniques based on natural evolution.

One of the important advantages of Fuzzy Logic Controllers (FLC) in comparison with conventional linear controllers is that they provide an ability of non-linear control behaviour. However, the design of such a controller is not an easy task because there are many parameters, which are usually set with designer experiences or with simulation studies. Another approach is to use optimization. But FLC has several inputs (very usually two: for control error and its derivative) and each input has several membership functions. Beside there are many rules in the FLC knowledge base in which many so called consequent parameters appear.

As traditional optimization methods are too sensitive to the number of parameters, an advantage with optimization based on genetic algorithm (GA) was expected.

Modern tools, which were used for this study (MATLAB-SIMULINK, Fuzzy Logic Toolbox, Genetic Algorithm Toolbox [11]) give wide possibilities for efficient design and experimentation.

2 Description of genetic algorithm

GA [2], [6], [7] used for optimization has four standard operations:

- evaluation of individuals,
- reproduction,
- crossover and
- mutation

These operations are repeated until the terminating condition is met. In our case the optimization was stopped after a specified number of generations were evaluated. Table 1 shows the characteristic parameters of GA.

As a reproduction mechanism a method called deterministic roulette wheel was used [11]. Parameters were binary

number of generations	N_g
crossover probability	p_c
number of crossover points	N_c
mutation probability	p_m
number of individuals in gen.	N_p

Table 1: GA parameters

coded (with 12 bits). The most important parameter of the GA is fitness function, which is given by Eqs. (1), (2) and (3). Eq. (1) is well known criterion function often used in control systems design.

$$f = \int_0^{t_{max}} |e(t)|dt \tag{1}$$

$e(t)$ is the error between reference and controlled variable. Absolute error is used, as error is usually an oscillating signal. To get faster convergence, relative differences between fitness values for particular individuals are further increased by subtracting the minimal fitness in a generation k from the criterion defined with Eq. (1).

$$f_1(k) = f(k) - \min_{i=1...N_p}(f(i)) + 1 \tag{2}$$

So the offset of the fitness function is removed. One is added to prevent the value zero of the fitness function $f_1(k)$. The differences between individuals are further increased by the transformation into final fitness function

$$f_2(k) = \left(\frac{f_1(k)}{\frac{1}{N_p}\sum_{i=1}^{N_p} f_1(i)} \right)^3 \tag{3}$$

Individuals with fitness above the average fitness get higher fitness value, while the ones below get lower fitness value. So with Eqs. (2) and (3) the relative differences between fitness values are greatly increased, enabling better individuals to have more offspring.

3 Description of the fuzzy logic controller

The controller used in experiments was Sugeno 0^{th} order type of FLC [14]. This type was chosen, because it is simpler in comparison with other types of FLCs also from the calculation complexity point of view. Namely each optimization performs many simulation runs and the CPU time needed for one simulation depends very much on the time needed for controller action evaluation. On the other hand its properties satisfied one of the important requirements - the ability to control non-linear process [14].

The FLC has usually two inputs x and y for control error and its derivative. Membership functions in our studies were equally spaced as it is shown in Figure 1. They were fixed during optimization.

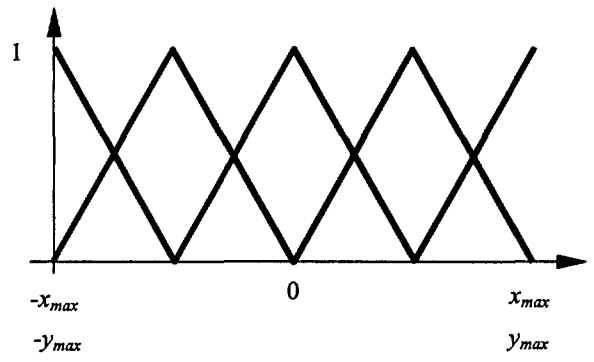


Figure 1: Membership functions of FLC input variables

Knowledge base of the FLC consisted of $P \times Q$ rules as is depicted in Figure 2.

1. IF $x = MF_{x1}$ AND $y = MF_{y1}$ THEN $u_{1,1} = C_{1,1}$
2. IF $x = MF_{x1}$ AND $y = MF_{y2}$ THEN $u_{1,2} = C_{1,2}$
3. IF $x = MF_{x1}$ AND $y = MF_{y3}$ THEN $u_{1,3} = C_{1,3}$

Q IF $x = MF_{x1}$ AND $y = MF_{yQ}$ THEN $u_{1,Q} = C_{1,Q}$
$Q+1$ IF $x = MF_{x2}$ AND $y = MF_{y1}$ THEN $u_{2,1} = C_{2,1}$

$P \times Q$ IF $x = MF_{xP}$ AND $y = MF_{yQ}$ THEN $u_{P,Q} = C_{P,Q}$

Figure 2: Knowledge base of FLC

x	the first input variable
y	the second input variable
MF_{xi}	i^{th} membership function of the first input variable
MF_{yi}	j^{th} membership function of the second input variable
u_{ij}	output of the $(i,j)^{th}$ rule
C_{ij}	consequent parameter
P	the number of membership functions of the first input variable
Q	the number of membership functions of the second input variable

The fuzzy logic operation and inference mechanism were realized as product [12].

$$u_{i,j} = C_{i,j} m_{xi} m_{yj} = C_{i,j} r_{i,j} \tag{4}$$

where

m_{xi}	membership grade of the first input
m_{yj}	membership grade of the second input
$r_{i,j}$	fulfilment of the $(i.j)^{th}$ rule, calculated as the product of membership grades

The controller output is calculated as the integral of the weighted average of consequent values

$$u(t) = \int_0^t \frac{\sum_{i=1}^{P}\sum_{j=1}^{Q} r_{i,j}C_{i,j}}{\sum_{i=1}^{P}\sum_{j=1}^{Q} r_{i,j}}dt = \int_0^t \frac{\sum_{i=1}^{P}\sum_{j=1}^{Q} u_{i,j}}{\sum_{i=1}^{P}\sum_{j=1}^{Q} r_{i,j}}dt \quad (5)$$

With integration the fuzzy logic controller, which is actually a non-linear PD controller, was transformed into a PI one in order to eliminate steady state error.

4 Optimization

In our optimization study five equally spaced membership functions ($P = 5, Q = 5$) were used for each input variable (as shown in Figure 1). With optimization 25 consequent parameters $C_{i,j}$, each was coded with 12 bits, were determined. It is obvious that many troubles could be expected using conventional optimization techniques. The procedure is shown in Figure 3.

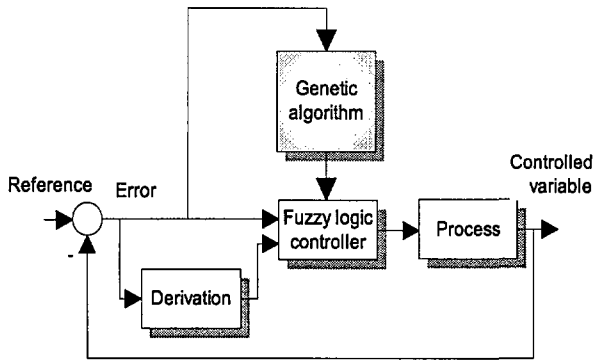


Figure 3: Optimization of fuzzy controller

FLC design is much more complex than linear controller design, because its output characteristic is non-linear. Output values for all possible input values have to be defined. The range of input values is divided into smaller intervals. The number of these intervals depends on the number of membership functions. Table 2 shows the truth table for control error e and its derivative de/dt. The fuzzyfied values are NB (negative big), NM (negative medium), ZE (zero), PM (positive medium) and PB (positive big).

Eq. 5 shows that control signal u is influenced only by those consequent parameters $C_{i,j}$ which have appropriate non zero membership grades $r_{i,j}$. In other words, only the rules with membership functions, which are defined on the domains of the current input variables are active. So it is obvious that the optimization of controller parameters $C_{i,j}$ is efficient only when the control error and its derivative cover the whole area defined by both variables during transient responses (simulation runs). Unfortunately as close

de/dt \ e	NB_E	NM_E	ZE_E	PM_E	PB_E
NB_{dE}	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$	$C_{1,4}$	$C_{1,5}$
NM_{dE}	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	$C_{2,4}$	$C_{2,5}$
ZE_{dE}	$C_{3,1}$	$C_{3,2}$	$C_{3,3}$	$C_{3,4}$	$C_{3,5}$
PM_{dE}	$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	$C_{4,4}$	$C_{4,5}$
PB_{dE}	$C_{5,1}$	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	$C_{5,5}$

Table 2: Trajectory of error and its derivative drawn on the FLC's truth table

loop system is optimized, it is not possible to select directly the appropriate signals $e(t)$ and $de(t)/dt$. Instead one has to select appropriately the reference signal. In Table 2 a typical trajectory of error and its derivative caused by a simple reference step change shows that with such input many consequent parameters can not be appropriately optimized. To overcome this problem, several types of reference signals were tested:

- a signal consisting of sinusoidals with different amplitudes and frequencies,
- a signal consisting of steps with different amplitudes and delays,
- a square wave signal with increasing amplitude.

After substantial testing the square wave signal with increasing amplitude seemed to be the best solution. It is shown in Figure 4, together with control error for a typical example, while the appropriate trajectory is shown in Figure 5.

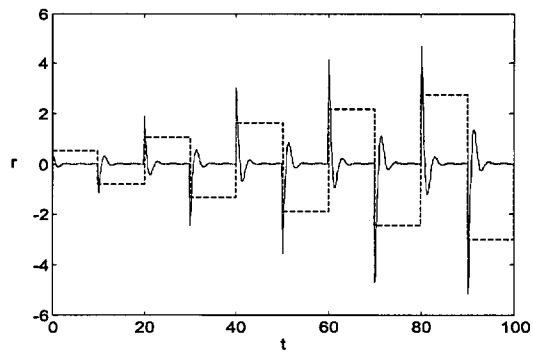


Figure 4: Reference signal (dashed line) and control error

As the whole phase area is fulfilled, it is possible to optimize all consequent parameters. However, in most cases GA finds solution close to the optimum, not the exact optimum. In our case this means, that some consequent parameters are slightly smaller than they should be, while the

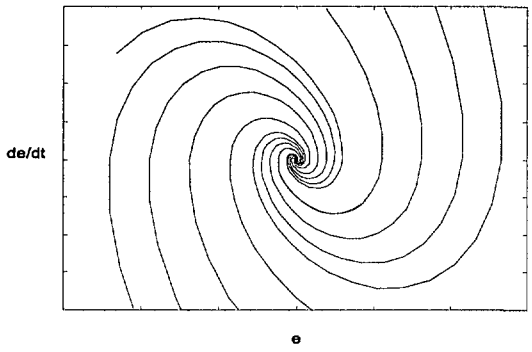


Figure 5: Trajectory in the error - derivative plane

others (perhaps the neighbours, see Table 2) are slightly larger. If the FLC output characteristics is considered as a non-linear function of two independent variables e and de/dt the surface is not very smooth as it has many local minima and maxima. Such surface can not assure the appropriate performance of the control system. As this inconvenience can be considered as a kind of noise introduced by stochastic features in GA, the idea to use a kind of filtering arose.

The idea of filtering is, to calculate the consequent parameter value by averaging in which the parameter itself and all parameter's neighbours are included (see Table 2)

$$C'_{i,j} = \frac{\frac{a}{4}(C_{i-1,j-1} + C_{i-1,j+1} + C_{i+1,j-1} + C_{i+1,j+1})}{a+b+c} + \frac{\frac{b}{4}(C_{i,j-1} + C_{i,j+1} + C_{i-1,j} + C_{i+1,j})}{a+b+c} + \frac{c \cdot C_{i,j}}{a+b+c} \quad (6)$$

where

- $C'_{i,j}$ new (filtered) value of the consequent parameter
- $C_{m,n}$ current values of consequent parameters ($m = i-1, i, i+1$, $n = j-1, j, j+1$)
- a, b, c weights (parameters of the filter)

Using the filtering, the best results are obtained with optimization in several steps. After each step the filtering is used, what means that the new individuals are calculated from all individuals of the last generation of previous optimization step. Values of the filtering parameters a, b and c depend mostly on the type of process. There is no strict rule how to set them, but in our examples the starting values were set to 1 ($a = 1, b =$ 1and $c = 1$). In some experiments b and c were intensified during optimization steps.

5 Experimental results: Optimization of the FLC controller of a hydraulic system

Our laboratory hydraulic set-up consists of three tanks and a main reservoir [5]. The transfer function which describes the relation between the input flow of liquid (incoming flow in the first tank) and the level in the third tank (controlled variable) is

$$G_P(s) = \frac{Y(s)}{U(s)} = \frac{1}{s^3 + 2s^2 + 3s + 1} \quad (7)$$

As mentioned FLC was Sugeno 0^{th} order with two inputs (error and its derivative), for each input five equally spaced membership functions were defined (Figure 1). Knowledge base was described with 25 rules (Figure2). The controller output was calculated with Eq.(5). Optimization with GA was used to calculate optimal values of 25 consequent parameters $C_{i,j}$, each was coded with 12 bits. GA selects the best controllers from the generation and performs other operations (crossover and mutation). Selection is made on the basis of fitness values that depend on the control error (see Eqs. (1), (2), (3)). The important parameters of GA are shown in Table 3.

number of generations	N_g	70
crossover probability	p_c	1
number of crossover points	N_c	3
mutation probability	p_m	0.01
number of individuals in gen.	N_p	30

Table 3: GA parameters

The overall scheme is shown in Figure 3.

The first generation of individuals was initialized with random numbers - no knowledge about the process was included. The optimal control system performance is shown in Figure 6.

The small oscillations are caused by rough FLC output characteristic which can be confirmed by Figure 7, where gray scale is used to denote the profile of the plane (darker means less).

After this study the filtering was introduced. Three optimization iterations were performed, each with 70 generations. After each optimization the filtering was used. After the first optimization and filtering the criterion function was 672, after the second 275 and after the third 262. As the change between the second and the third iteration was not significant, the iterative procedure was terminated. Other examples also confirm that three iterations are usually reasonable. Figures 8, 9 and 10 represent the optimal results (reference sp, controlled variable y and control signal u) after the first, second and the third iteration of optimization.

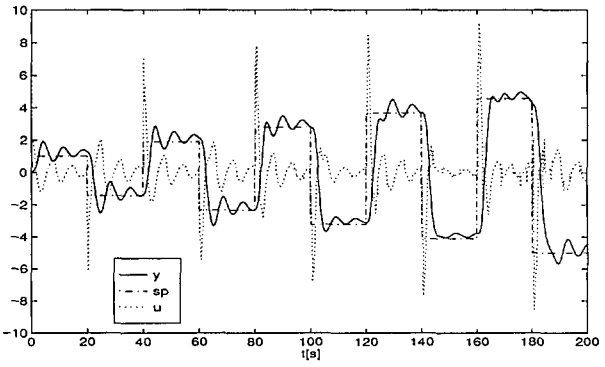


Figure 6: Optimal results (without filtering, y ... controlled variable, u ... control variable, sp ... reference)

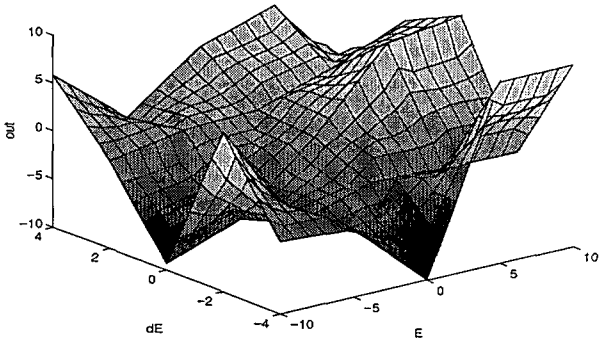


Figure 7: Output characteristic of the FLC optimized with GA (without filtering)

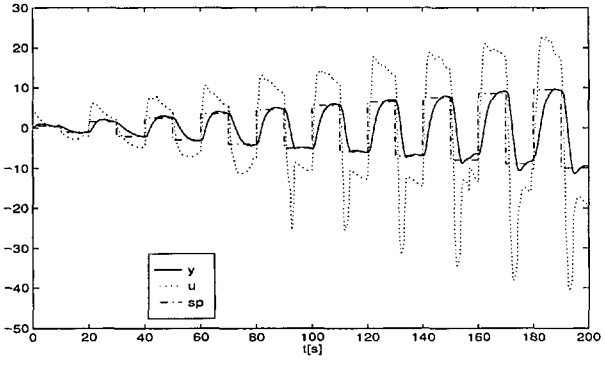


Figure 8: Results after the first iteration of optimization

It can be seen, that responses at rising and falling edges are not the same, because FLC characteristic is not symmetrical (linear). The response in Figure 10 is much better as it is fast and with a small overshoot. As the fitness function depends only on the control error, the values of controller signal u are very high in the points of reference change. If such values are unacceptable for actuator, the control variable u should be somehow included in the criterion function.

Figure 11 depicts the output characteristic of the FLC.

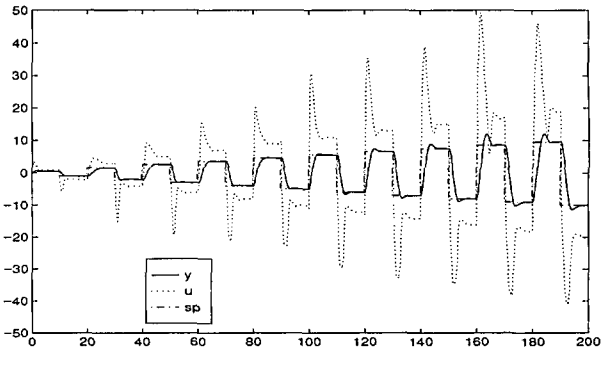


Figure 9: Results after the second iteration of optimization

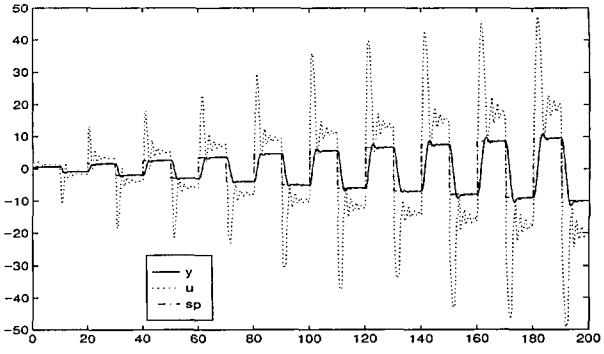


Figure 10: Results after the third iteration of optimization

Using filtering it is much smoother.

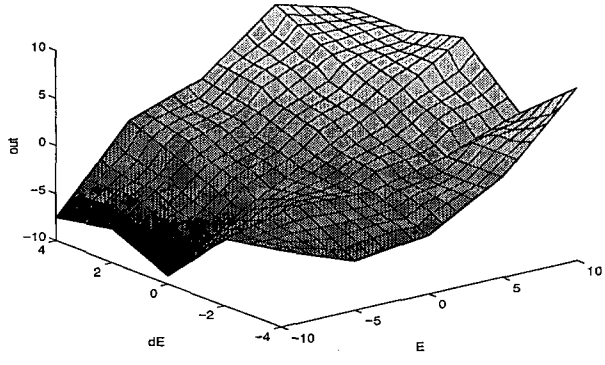


Figure 11: Output characteristic of the FLC optimized with GA (with filtering)

6 Conclusion

Genetic algorithms seem to be an efficient optimization approach in complex control systems with many tuning parameters. In fuzzy logic control systems there are many parameters, which can influence the behaviour: the number and the shape of membership functions, consequent pa-

rameters, etc. So conventional optimization techniques are usually not enough efficient or even unusable.

In our presented study the consequent parameters of FLC were optimized with GA. Experiences show that some of the parameters can have more or less random values after the optimization if some facts are not taken into account. To avoid such situation, appropriate type of reference signal should be used. It is also recommended to plot trajectory of FLC inputs (e.g. plane e, de/dt) to see which parts of the truth table is appropriately covered by the inputs and to find out which consequent parameters can not be optimized. Observation of FLC output characteristics is also useful because smooth shapes mean that the optimization produce at least near optimum values.

Filtering of FLC characteristic is another useful method, which makes output characteristics smoother and so improves responses. The procedure also decreases bad influence of parameters which are not satisfactory optimized, because their values get closer to the average of other optimized parameters.

However in the future more effort should be devoted to additional experiments with different reference or disturbance signals, which are more similar to shapes, which occur in reality. A lot of possibilities give also different types of FLCs as well as the study of the influence of different approaches in GA.

References

[1] Dixon L.C.W. (1972) *Nonlinear optimization*. The English universities press limited.

[2] Goldberg D. E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley Publishing Company.

[3] Åstrom. K. J. & Hägglund T. & Hang C. C. (1990), Automatic tuning and adaptation for PID controllers - a survey. *Automatica*, 20, p. 645-651.

[4] McMillan G. K. (1990) *Tuning and Control Loop Performance - A Practitioner's Guide (2nd ed.)*, Instrument Society of America, NC.

[5] Matko D. & Karba R. & Zupančič B. (1992) , *Simulation and Modelling of Continuous Systems: A Case Study Approach*. Prentice Hall Int., London.

[6] Beasley D. & Bull D. & Martin R. (1993) An Overview of Genetic Algorithms: Part 1, Fundamentals. *University Computing*, 15, 2, p. 58-69.

[7] Beasley D. & Bull D. & Martin R. (1993) An Overview of Genetic Algorithms: Part 2, Research Topics. *University Computing*, 15, 4, p. 170-181.

[8] Punch W. F. & Goodman E. D. & Min Pei & Lai Chia-Shun & Hovland P. & Enbody R. (1993) , Further research on Feature Selection and Classification Using Genetic Algorithms. *ICGA93*, Champaign III, p. 557-564 .

[9] Beasley D. & Bull D. & Martin R. (1993) A Sequential Niche Technique for Multimodal Function Optimization, Technical Report No. 93001. *University of Bristol*, Bristol, UK.

[10] Beasley D. & Bull D. & Martin R. (1993) , Reducing Epistasis in Combinatorial Problems by Expansive Coding. *Proc. 5^{th} International Conference on Genetic Algorithms*, Ed. S. Forrest, Morgan Kaufman, p. 400-407

[11] Prong van Hoogeveen J. W. D. (1995) *Optimization using Genetic Algorithms*. Ph.D. thesis, Technical university Delft. 3-14.

[12] Jager R. (1995) *Fuzzy Logic in Control*. Ph.D. thesis, Technical university Delft.

[13] Zhao J. &, Gorez R. & Wertz V. (1996) Genetic algorithms for the elimination of redundancy and/or rule contribution assessment in fuzzy models *Mathematics and Computers in Simulation*, 41, p. 139-148.

[14] Babuška and Verbruggen H. B. (1996) An Overview of Fuzzy Modelling for Control *Control Eng.Practice*, 4, 11, p. 1593-1606.

PNNI And The Optimal Design Of High-Speed ATM Networks

Abdella Battou
Center for Computational Science,
U.S. Naval Research Laboratory. Washington D.C.
AND
Bilal Khan and Sean Mountcastle,
ITT Industries Systems Division,
at the Center for Computational Science,
U.S. Naval Research Laboratory. Washington D.C.
E-mail: {battou,bilal,mountcas}@cmf.nrl.navy.mil

Keywords: ATM Network Design, PNNI Simulation, PRouST

Edited by: Mohsen Guizani
Received: January 15, 1999 **Revised:** August 25, 1999 **Accepted:** July 31, 1999

In addition to being well-dimensioned and cost-effective, a high-speed ATM network must pass some performance and robustness tests. We propose an approach to ATM network topology design that is driven by the performance of its routing protocol, PNNI. Towards this end, we define performance indicators based on the time and traffic required for the protocol to first enter and subsequently return to the meta-stable state of global synchrony, *in which switch views are in concordance with physical reality. We argue that the benefits of high call admission rate and low setup latency are guaranteed by our indicators. We use the PNNI Routing and Simulation Toolkit (PRouST), to conduct simulations of PNNI networks, with the aim of discovering how topological characteristics such as the diameter, representation size, and geodesic structure of a network affect its performance.*

1 Introduction

The size of operational ATM clouds continues to grow at an increasing pace. Both in anticipation of this changing scale, and to insure smooth inter-operation of these networks, the ATM Private Network-Network Interface standard (PNNI) was recently adopted (ATM Forum 1996). PNNI defines a set of protocols for *hierarchical* networks of ATM switches, and is designed to provide efficient and effective routing. In the long term, however, the degree to which PNNI succeeds in this regard will depend crucially on two factors:

First, because PNNI does not mandate specific policies for call admission, route selection, or topology aggregation, these aspects of the protocol remain "implementation-specific". Clearly the degree to which PNNI meets the challenges posed by tomorrow's ATM networks will depend significantly on the success of *switch designers* in devising effective algorithms for the admission and routing of connections, and for the aggregation of topology information.

Second, *network designers* must have the tools and information necessary to design ATM network topologies that are (i) capable of meeting anticipated traffic demands, and (ii) *optimized for performance under PNNI.* In this paper, we shall not address the first of these two issues, that of dimensioning networks to satisfy known costs and traffic demands. Our investigations begin at the point where a network designer, having been given projected traffic profiles and switch/fiber specifications, has arrived at a set of candidate ATM network topologies which appear equally adequate. We argue that although two topologies may appear indistinguishable in terms of the mathematics of QoS requirements, the PNNI protocol exhibits significant differentiation in their performance. Understanding how the PNNI protocol affects network performance is a necessary first step to determining how the adoption of PNNI *should* affect ATM network design. In subsequent sections, we shall describe our simulation experiments and begin developing guidelines for ATM network topology design that take into consideration the specific nature of the PNNI protocol.

2 PNNI Performance Indicators

There are many candidate performance criteria for evaluating the relative merits of network topologies. Here we shall assume that topology design is motivated by increasing the ATM network's call admission rate and decreasing the average connection setup latency. Additionally, we desire that the background traffic due to the PNNI protocol itself, should not be excessively high.

Setup Latency. In the absence of crankback, setup latency within a peergroup is seen to be linearly correlated with the number of hops in the selected path (Niehaus et

al. 1997) and thus may be estimated in the worst case by the network diameter. When crankbacks occur, each failed attempt at valid route selection contributes significant additional latency, required for backtracking to the entry border node, computing an alternate route and then re-traversing the peergroup along the new path. Reduction of setup latency thus requires minimizing the crankback frequency.

Crankbacks and Call Admission Rate. Recent results on PNNI aggregation schemes [2,4] indicate that ATM call admission rate and crankback frequency is directly proportional to the "distortion" present in switches' views of the network. In particular, the experimental data presented in (Awerbuch et al. 1998) confirms the intuition that when a switch has inaccurate (e.g. outdated) views of network topology and metric information, this increases the likelihood that calls entering the peergroup at that switch will be assigned sub-optimal routes. A larger discrepancy between a switch's local information and the underlying reality of the network's state, results in a larger fraction of calls originating at the switch being rejected en-route, hence undergoing crankbacks (and possibly even unwarranted rejection at the source).

Thus, beyond the problem of dimensioning, selecting topologies that will yield high ATM call admission rates and low average setup latency requires that one be able to characterize which topologies minimize the divergence in switches' views.

2.1 Our Approach

Local synchrony. We define *local synchrony* of a peergroup to be the state where every switch in the peergroup has knowledge of the same set of PNNI Topology State Elements (PTSEs). It follows from the logic of the PNNI NodePeer finite state machine, that if a peergroup reaches local synchrony, then all member switches agree about the topology metrics describing their peergroup. Within a PNNI peergroup, each member switch is responsible for originating and flooding accurate information about its internal state and the resource availabilities on its incident links. Thus, modulo any loss due to aggregation schemes, local synchrony may be interpreted as a state in which all members of the peergroup are in agreement not only amongst themselves, but also with the underlying reality of the peergroup's state.

Global synchrony. We define *global synchrony* of a connected ATM network to be the state where every peergroup at every level has reached local synchrony, and the PNNI network hierarchy has reached a unique apex. Admittedly, the notion of global synchrony is "artificial" in the sense that it may be rarely achieved in real dynamic networks where connections are constantly arriving and departing. However, in a simulated network this state is attainable, and we shall use it to probe the rate of PNNI information propagation.

When a switched virtual circuit (SVC) is established in an ATM network, bandwidth availability is altered for links that the SVC traverses. Assuming this change is significant, updated information is re-originated and flooded by each switch incident to the affected links. If the network had reached global synchrony prior to the SVC setup request, these re-originations cause the network to fall out of a state of global synchrony for a brief time, until the new information has reached every node. This naturally leads us to consider:

- **Resynchronization time:** Average time required for the network
  to return to global synchrony, after a single, isolated, random SVC setup.
- **Resynchronization traffic:** Average PNNI traffic required for the network
  to return to global synchrony, after a single, isolated, random SVC setup.

The basic "trial" involved in measuring the above **resynchronization parameters** is as follows: allow the network to reach global synchrony, then inject a connection request between randomly chosen source and destination nodes and measure the time required and bytes transmitted before the network returns to a state of global synchrony. By repeating this trial a large number of times, we obtain an average value, which we call the resynchronization time.

To illustrate the importance of resynchronization time, let us consider two extreme scenarios. First, consider a network where the average time between SVC requests (i.e. 1/ SVC arrival rate) is much higher than the network resynchronization time. In this situation, changes in bandwidth availability induced by an SVC setup will, on average, have time to flood to all other switches in the network prior to the arrival of the next SVC setup request. Thus, routing decisions for each SVC will, on average, be made according to completely accurate information at the originating switch. If an SVC setup is rejected or experiences unacceptably high latency, this undesirable behavior is attributable solely to inadequate dimensioning of the network, because there is no legitimate way to fulfill the request.

In contrast, consider a network where the average time between SVC requests is much lower than the network resynchronization time. In this situation, changes in bandwidth availability induced by an SVC setup have not yet propagated to many switches in the network by the time the next SVC setup request arrives. Thus, the routing decision for an SVC is likely to be made according to stale information. The extent to which the information is stale is determined by the extent to which network resynchronization time exceeds average inter-SVC arrival time. If an SVC setup is rejected or experiences unacceptably high latency, this undesirable behavior may be due to inadequate dimensioning of the network or it may be due to suboptimal routes and unwarranted rejections induced by inaccurate information at the switches.

Major changes in network topology, such as network partitioning due to link/node failure, or re-merging of components upon subsequent recovery, will cause the PNNI

hierarchy to undergo severe restructuring. We define the **boot parameters** below to be indicators of the time and traffic required to reinstate consistent routing information after such catastrophic changes.

- **Boot time:** Time at which the network first reaches global synchrony.
- **Boot traffic:** PNNI traffic required for the network to reach
  global synchrony for the first time.

We take the parameters above as a worst-case estimate of the time and traffic resources required to return to global synchrony. By comparison, the resynchronization parameters described earlier aim to measure the same quantities for the average case, i.e. during normal (stable) operation of the network.

We contend that a network designer, given two topologies that are equivalent with regards to meeting anticipated QoS requirements, must take into consideration their resynchronization and boot times. In particular, reducing these two quantities increases the fraction of time that the network spends in a state of global synchrony. At the same time, the designer must keep a watchful eye on the Boot traffic and Resynchronization traffic to make sure that not too much of the network bandwidth is being expended by PNNI itself.

One way in which the designer might determine the values of the four parameters mentioned above is to take physical measurements of them on two live networks, each configured in the appropriate topology. But this would be difficult to do accurately because of the inherent problems in distributed measurement, and moreover it would completely defeat the intention of *design before implementation!* Alternately, one could simulate the candidate PNNI networks to determine both time and traffic for boot and resynchronization; we follow this latter approach here.

3 Experiments

3.1 The Simulation Environment

Our simulation experiments were carried out using the PNNI Routing and Simulation Toolkit (PRouST), which was developed by the Signaling Group at the Naval Research Laboratory's Center for Computational Sciences. PRouST is a faithful and complete implementation of version 1.0 of the PNNI standard and can be used both to simulate large networks of ATM switches as well as to emulate live ATM switch stacks. In particular, PRouST includes the Hello, NodePeer, and Election finite state machines, all relevant packet encoding and decoding libraries, routing database management, and full support for hierarchy. In addition, a "plug-in" interface for call admission, path selection and aggregation policies is provided. For inter-switch signaling PRouST makes use of the NRL Signaling Entity for ATM Networks (SEAN), which is a complete simulation/emulation library implementing version 4.0 of the ATM User-Network-Interface (UNI) standard. The fidelity of PRouST's PNNI implementation has been demonstrated extensively in live interoperability tests with commercial switches. Both PRouST and SEAN are written in C++ over the Component Architecture for Simulation of Network Objects (CASiNO) described in (Mountcastle et al. 1999), and both are available in the public domain.

In the simulations that follow, all network links operate at the OC3 rate. Link jitter varies uniformly between $+10$ μs and -10 μs for each transmitted message. PNNI messages that enter the switch control port experience a latency of 10 ms. The backplane of the switch routes data traffic on existing virtual circuits at OC48 rates. These figures, while artificial, are projections of current switch vendor specifications. We found that jitter did not noticeably alter the outcome of our simulations from one trial to the next. The variations were typically less than 1% and for boot and resynchronization time, and less than 5% for boot and resynchronization traffic. The values we have listed in our tables are mean values.

An outline of results presented: We seek to understand what factors influence resynchronization and boot parameters. To this end, we start by simulating single-peergroup networks with grid, chain, ring and star topologies; these results are presented in sections 3.2 and 3.3. We compare these very particular families of topologies with similar experiments using all possible topologies on 7 nodes–these results are described in sections 3.4-3.5. In addition, we simulate 100 randomly generated topologies on 20 nodes, the results of which are described in section 3.6. Finally, in sections 3.7 and 3.8 we address the impact of peergroup size and hierarchy, by simulating several different hierarchic configurations of linear networks.

3.2 Boot and Resynchronization Time

We start by investigating the characteristics of network topology that influence boot and resynchronization times. The NodePeer protocol floods PNNI Topology State Elements over a link whenever the switches incident to the link have discrepant databases. Thus, information will flow from each switch radially until it has reached every other switch in the peergroup. Given this, network boot time and worst-case resynchronization times should be linearly dependent on the diameter of the peergroup.

Simulation Results We simulated PNNI networks of increasing size, specifically, chains, grids rings, and star networks. The tables (1,2,3,4) show the results of the simulations, and are depicted in figures 1 and 2. The figures indicate that PNNI boot time grows linearly in network diameter; for each unit increase in diameter PNNI boot time increases by approximately 21.4 ms, while resynchronization times increase by 10.7 ms.

Chain Length	Network Diameter	Boot Time seconds	Resynch. Time seconds
2	1	30.0864	0.0108
4	3	30.1290	0.0323
6	5	30.1719	0.0537
8	7	30.2146	0.0755
10	9	30.2588	0.0962
20	19	30.4562	0.2035
30	29	30.6859	0.3104

Table 1: Chains—Boot/resynch. times

Grid Size	Network Diameter	Boot Time seconds	Resynch. Time seconds
2x2	2	30.1073	0.0212
3x3	4	30.1498	0.0428
4x4	6	30.1926	0.0636
6x6	10	30.2785	0.1067

Table 2: Grids—Boot/resynch. times

Ring Length	Network Diameter	Boot Time seconds	Resynch. Time seconds
10	5	30.1654	0.0518
20	10	30.2694	0.1037
30	15	30.3732	0.1555
40	20	30.4932	0.2172
50	25	30.6002	0.2688

Table 3: Rings—Boot/resynch. times

3.3 Boot and Resynchronization Traffic

To study PNNI boot and resynchronization traffic, we will introduce the notion of the representation size of a network: the minimum number of PNNI topology state elements required to fully describe its topology. In our subsequent discussion, we take the representation size of a network to be twice the number of links plus the number of switches.

Because the NodePeer protocol is responsible for transmission of the peergroup's current representation to each component switch, traffic required to reach *initial* global synchrony should be bounded below by a function proportional to the product of the representation size and the number of switches. For chains, grids and rings the number of nodes and the representation size are linearly related, so this product is quadratic in the representation size.

Resynchronization involves flooding updated information about links affected by the SVC, to all members of the peergroup. In the best-case, flooding takes place over a spanning tree and the total traffic required to resynchronize is proportional to the number of switches in the peergroup. In the worst case, when flooding is occurs over every edge in the network, traffic due to resynchronization is proportional to the number of links in the network. For this reason, we consider resynchronization traffic in PNNI networks as a function of representation size.

Star Size	Network Diameter	Boot Time seconds	Resynch. Time seconds
5	2	30.0872	0.02172
10	2	30.0886	0.02179
15	2	30.0894	0.02188
20	2	30.0896	0.02191
25	2	30.0909	0.02198

Table 4: Stars—Boot/resynch. times

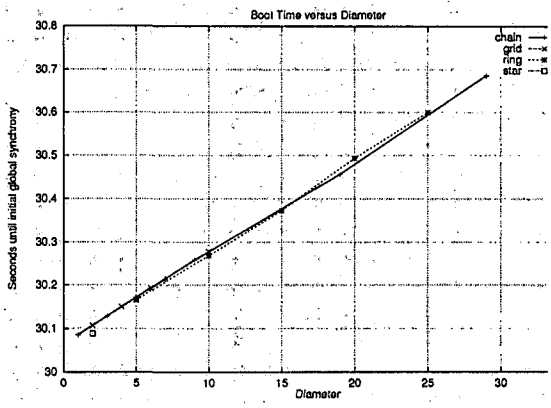


Figure 1: Boot time vs. network diameter

Simulation Results We interpret the traffic data collected from the simulations of the previous section. This data is shown in tables (5,6,7,8). The traffic required to reach initial global synchrony in each of these cases, grows super-linearly with the representation size, as can be seen in figures 3 and 4). Resynchronization traffic also manifests super-linear growth as a function of representation size.

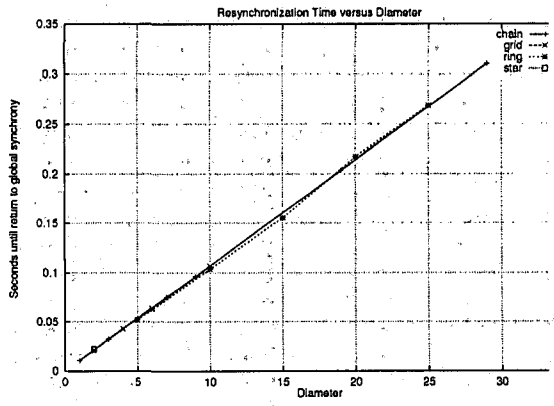


Figure 2: Resynch. times vs. network diameter

Chain Length	Repres. Size	Boot Traffic bytes	Resynch. Traffic bytes
2	4	10292	2967
4	10	56028	13399
6	16	133976	30127
8	22	248124	53247
10	28	394476	82575
20	58	1629364	324487
30	88	3702444	719548

Table 5: Chains—Boot/resynch. traffic

Grid Size	Repres. Size	Boot Traffic bytes	Resynch. Traffic bytes
2x2	12	73284	14412
3x3	33	182232	43240
4x4	64	1674880	260640
6x6	156	3521512	816003

Table 6: Grids—Boot/resynch. traffic

Ring Length	Repres. Size	Boot Traffic bytes	Resynch. Traffic bytes
10	30	201812	2720
20	60	765388	5440
30	90	1379620	108160
40	120	2167876	816680
50	150	3353088	1286000

Table 7: Rings—Boot/resynch. traffic

3.4 General Tests I: All 7 Node Networks

In order confirm the validity of the above conclusions, we conducted the same experiments on the class of all (853 topologically distinct) 7 node connected graphs[1].

Simulation Results Because the 7 node networks all have relatively small diameter, there was little differentiation in their boot and resynchronization times. As the graphs in figures 5 and 6 indicate, boot and resynchronization times were clustered at discrete values spaced 10ms apart. We note that 10ms is the time required for processing of a single PNNI message at the control port of a switch. While the plot does not immediately illustrate this, the distribution of the data points was not uniform over the cluster points; we have plotted the average as a function of diameter to emphasize this. Somewhat surprisingly, on average, resynchronization time seems to grows super-linearly

[1]The software used to generate all non-isomorphic 7 node graphs was the NAUTY program developed by Brendan McKay.

Star Size	Repres. Size	Boot Traffic bytes	Resynch. Traffic bytes
5	16	30864	9144
10	31	129024	43976
15	46	292724	104776
20	61	526464	191624
25	76	820152	302496

Table 8: Stars—Boot/resynch. traffic

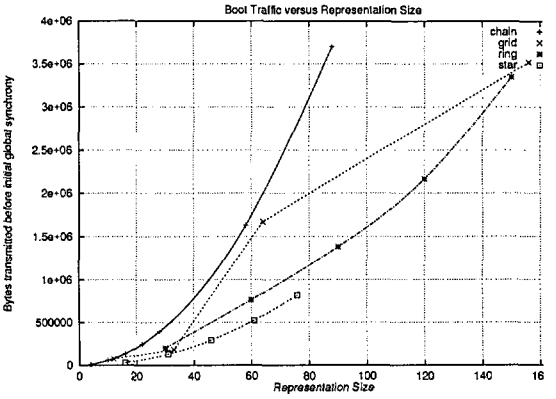


Figure 3: Boot traffic vs. representation size

with network diameter. More simulations need to be conducted over a larger class of graphs to determine the cause of this phenomenon. The PNNI boot traffic for these simulations is shown in figure 7, and is linear with an approximate growth rate of 6000 bytes per unit of representation.

3.5 Network Geodesic Structure

The plot shown in figure 8 indicates that that over 7 node graphs of any fixed representation size, there is considerable variation in PNNI resynchronization traffic. We attempted to understand the cause for this differentiation by determining, for each representation size, which topologies

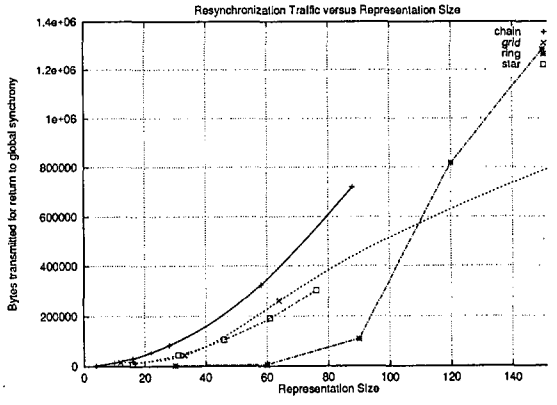


Figure 4: Resynch. traffic vs. representation size

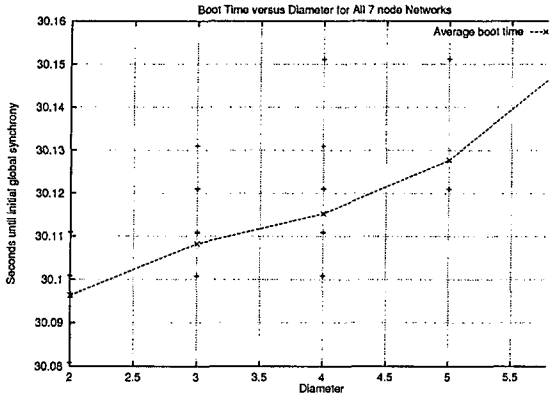


Figure 5: Boot times for 7 node networks

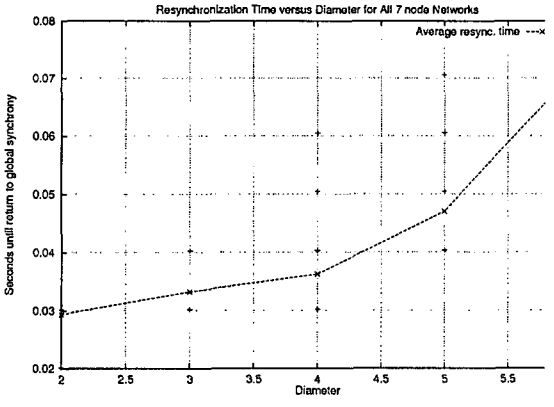


Figure 6: Resynch. times for 7 node networks

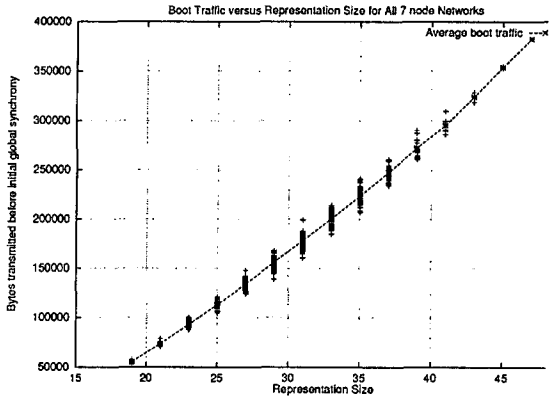


Figure 7: Boot traffic for 7 node networks

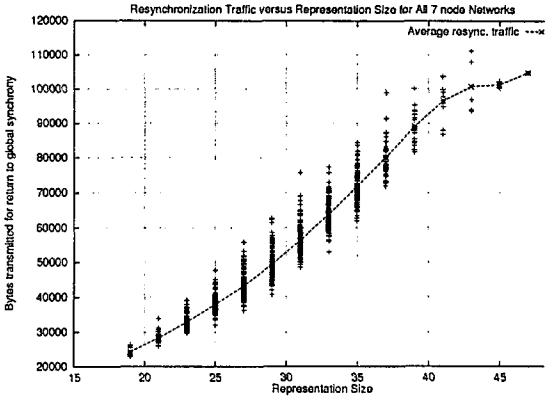


Figure 8: Resynch. traffic for 7 node networks

achieved the highest and lowest resynchronization traffic. Figure 9 is a partial list of the best and worst performers (only those with representation sizes between 19 and 33 are shown).

We noted that the worst performers had a large number of redundant geodesics between pairs of switches. That is to say, topologies with the worst performance contained a multiplicity of shortest paths between nodes. Scrutiny of the simulations revealed that this multiplicity results in a redundant flooding in the PNNI NodePeer protocol. Figure 10 illustrates the redundant flooding of a PTSE along one edge when two nodes are connected by more than one shortest path. The fact that the worst performers in figure 9 contain many unchorded 4-cycles, whereas the best performers contained many triangles, supports this conclusion.

3.6 General Tests II: Randomly Generated 20 Node Networks

We also conducted measurements of boot and resynchronization parameters for randomly generated networks.

A random topology is generated as follows: 20 nodes are assigned random locations on a grid. Links are added via a random process that repeatedly generates a random node-pair and adds a link between them with probability that decays exponentially with the Euclidean distance between the two nodes (Waxman 1998). Links that will cause the degree of a node to exceed eight are rejected by the random process in order to keep the graph reasonably sparse. The process of adding links terminates when the graph is connected and every node has degree at least 2. In this manner, we generated 8000 random topologies on 20 nodes. These were then sorted into classes, based on the diameter of the generated network, and 10 networks were chosen randomly from each class.

Simulation Results The results of simulations of 100 random 20-node networks are shown in figures 11 and 12. The data indicates that for any given value of diameter, there is a significant variation in the resynchronization and boot times. We were not able to determine the topological structure property that is responsible for this variance, but hope to address the question in our future work. The results of previous sections indicate that if we are given two networks from the same restricted family (such as grids, for example) then diameter alone can serve to predict the approximate value of the resynchronization time. In the set of experiments presented in this section, we realize that, unfortunately, this simple estimation criterion does not hold as well over large mixed families of graphs (e.g. our random set). On the other hand, we remark that the average value of resynchronization and boot times (over the ran-

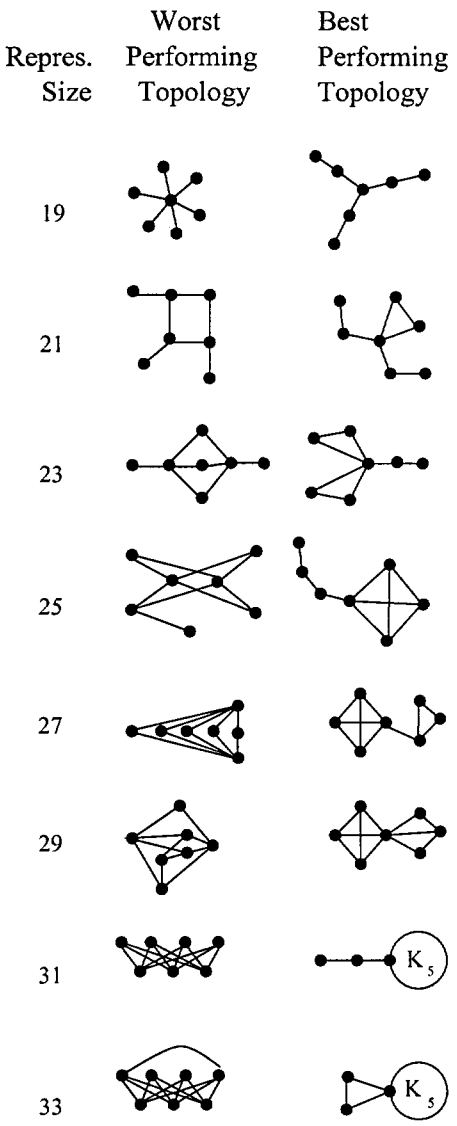


Figure 9: Best and worst 7 nodes topologies (in terms of boot traffic

domly chosen topologies) does increase linearly with diameter.

3.7 Peergroup Size

Another aspect of network design we consider is choice of peergroup size.

First, we note that partitioning a network into peergroups will result in more SVCs needing to be established between leaders at the next higher level. It also necessitates logical Hello finite state machines to stabilize over these SVCs and for the election process to conclude at the higher level. These two factors together are responsible for a discrete jump in the PNNI boot time when one moves from single peergroup to multiple peergroup configurations. As we begin to decrease the size of the peergroups, each requires less information to reach local synchrony, since detailed information about distant peergroups is being represented by a single logical node at the higher level. This causes boot traffic and resynchronization traffic to decrease. As we make peergroup size smaller still, an SVC setup on average traverses a larger number of peergroups, which in turn triggers re-aggregation of links at the higher level. Thus beyond a certain point, decreasing peergroup size causes an increase in resynchronization traffic and time.

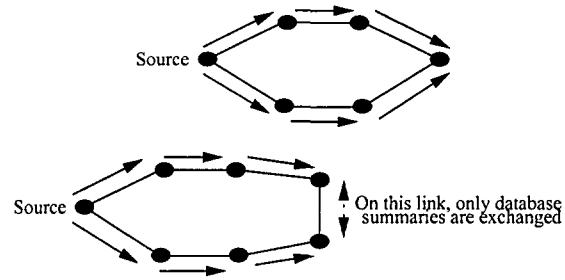


Figure 10: Effects of geodesic structure on traffic

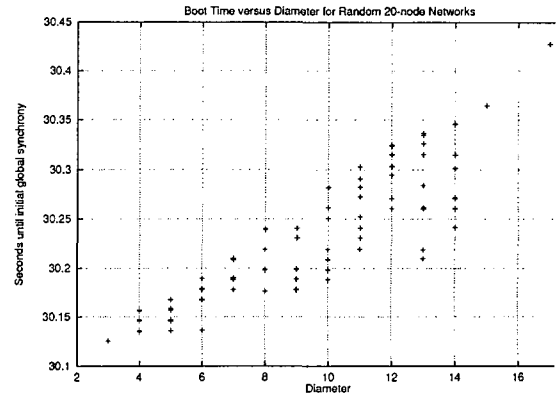


Figure 11: Boot times for 100 random 20-node networks

Simulation Results We simulated chains of 16 and 32 switches, configured with peergroups of sizes 2,4 and 8 (the 16 node examples are illustrated in figures 13). The data from the simulations is presented in tables 9 and 10.

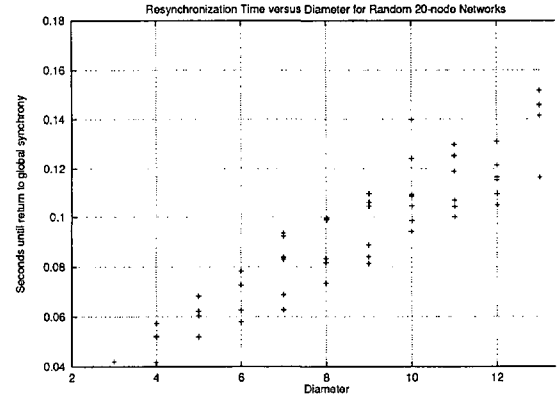


Figure 12: Resynch. times for 100 random 20-node networks

In the 32 node chain, going from a single peergroup of 32 nodes to 4 peergroups of 8 nodes produces a jump in the PNNI boot time of over 59 seconds. On the other hand, boot traffic decreases by a factor of 3, and resynchronization traffic by a factor of 6.4. Reduction of the peergroup size from 4 to 2, causes an increase in resynchronization traffic by 30%. This happens because higher level links are re-aggregated and flooded downward in response to the change in resources at the lower level.

Topology name	16-16	16-4	16-2
# of PGs	1	4	8
PG size	16	4	2
Boot time	30.38s	90.26s	90.27s
Boot traffic	1034K	639K	587K
NNI boot traffic	0	1.5K	1.8K
Resync. traffic	0.30s	0.23s	0.24s
Resync. traffic	494K	156K	233K

Table 9: Varying peergroup size for a 16 node chain

Topology name	32-32	32-8	32-4	32-2
# of PGs	1	4	8	16
PG size	32	8	4	2
Boot time	30.73s	75.43s	75.39s	75.49s
Boot traffic	4217K	1401K	1121K	1420K
NNI boot traffic	0	3.1K	3.7K	3.9K
Resync. time	0.64s	0.46s	0.46s	0.51s
Resync. traffic	1981K	306K	689K	898K

Table 10: Varying peergroup size for a 32 node chain

3.8 Hierarchy

Finally, we consider how the presence of hierarchy affects our performance indicators.

First, we note that a network with many levels of hierarchy requires SVCs to be established between leaders in the same higher level peergroup. It also necessitates logical Hello finite state machines to stabilize over these SVCs and for the election process to conclude in each peergroup at each level. These two factors are the principal cause of

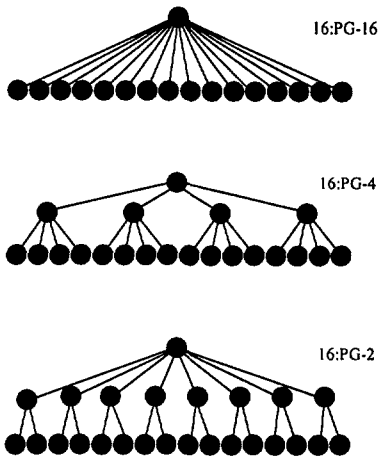


Figure 13: Examples of peergroups simulated.

the discrete jump seen in PNNI boot time and boot traffic as one considers networks with a greater number of levels. On the other hand, hierarchy localizes the side effects of network changes. In particular, it reduces the number of switches to which updated information must be flooded. Thus hierarchical addressing lowers both resynchronization time and traffic.

Simulation Results We simulated chains of 16 and 32 switches, configured with various hierarchical structure (The 16 node scenarios are depicted in figure 14). The data collected is presented in table 11. A 3-level configuration of the 32 switch chain boots in 75.4897 seconds, while the 5 level one requires 150.3220 seconds to reach initial global synchrony. The 5 level hierarchy has the advantage of improving both resynchronization time and traffic by a factor of 1.3 and 1.8 respectively.

Topology name	16/3	16/5	32/3	32/5
Chain length	8	8	32	32
Hierarchy structure	16,8,1	16,8,4, 2,1	32,16,1	32,16, 8,4,1
Boot time	90.27s	120.32s	75.49s	150.32s
Boot traffic	587K	611K	1420K	1717K
NNI boot traffic	1.8K	3.1K	3.9K	6.8K
Resync. time	0.24s	0.16s	0.51s	0.38s
Resync. traffic	233K	215K	1191K	638K

Table 11: Simulation results for several hierarchy configurations.

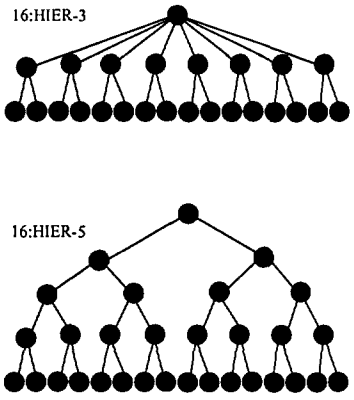


Figure 14: Two examples of hierarchic structures simulated.

4 Conclusion and Future Work

The simulations conducted using the PNNI Routing and Simulation Toolkit (PRouST) confirmed that topological characteristics such as the diameter, representation size, and geodesic structure do affect the boot and resynchronization times and traffic. These four indicators determine the discrepancy between switches' views of the network and its underlying physical state, and are thus critical to call admission rates and crankback frequency in ATM networks. Partitioning a network into peergroups reduces the boot and resynchronization traffic and introducing hierarchy results in improved resynchronization time and traffic, although it has a minor side-affect of increasing the boot time.

Peergroup size and hierarchy structure are two of the most importance choices a network designer must make. In our future research efforts will focus further on these two parameters. As we have shown in [section 3.7] there is an optimal value beyond which reduction of peergroup size results in increased resynchronization traffic, due to the re-aggregation and downward flooding of higher level links in response to changes at lower levels. We intend to precisely quantify both this optimal value, and the relative tradeoffs between peergroup size, hierarchy structure and resynchronization traffic and time.

5 Acknowledgements

We would like to thank David Talmage and Jack Marsh at the NRL Signaling Group for their contributions to PRouST, SEAN and CASiNO. Also we wish to acknowledge Sandeep Bhat for his assistance, particularly with the internals of the NodePeer finite state machine.

References

[1] ATM Forum. (1996) Private Network-Network Interface Specification, Version 1.0.

[2] Baruch Awerbuch, Yi Du, Bilal Khan, and Yuval Shavitt. (1998) Routing Through Networks with Hierarchical Topology Aggregation. *Journal of High Speed Networks*, vol. 7(1) p.57–73.

[3] W. C. Lee, Topology aggregation for hierarchical routing in ATM networks. *Computer Communications Review*, p. 82-92.

[4] S. Mountcastle, et al. (1999) CASiNO: A Component Architecture for Simulating Network Objects. *Proceedings of the 1999 Symposium on Performance Evaluation of Computer and Telecommunications Systems, July 11-15, 1999*. p. 261-272.

[5] D. Niehaus, et al. (1997) Performance Benchmarking of Signaling in ATM Networks. *IEEE Communications Magazine*, vol. 35 number 8, p. 134-142.

[6] Bernard M. Waxman. (1988) Routing of multipoint connections. *Journal on Selected Areas of Communications*, vol. 6 p. 1617-1622.

Intelligent Agent Technology

The first Asia-Pacific conference on intelligent agent technology with accompanying tutorials and workshops was held at Hong Kong Baptist University, 14-17 December 1999. Over 120 participants attended the conference. The acceptance rate for long conference papers was 27.8% and 22% for single- track WS papers.

The WS on agents in electronic commerce, WAEC'99, was chaired by Yiming Ye, IBM T.J. Watson Research Center. At the WS, researchers presented several ideas how to incorporate intelligent agents into electronic commerce. The WS was nicely concentrated on a specific subject and dominated by IBM researchers. Among the invited speakers was also Head of the IBM Deep Blue team that defeated Kasparov.

The conference had much broader spectrum than the WS. While major agents conferences are already specialized into specific agent subfields, the first Pacific conference was a uniform one with the aim to bring together all researchers in the agent area. Papers were grouped into the following categories: agent architectures, multi-agent cooperation, distributed intelligence, formal agent theories, knowledge discovery and data mining agents, personalized Web agents, software agents, mobile agents, agent-supported enterprise. Program chairs were Jiming Liu and Ning Zhong. They were also editors of the conference proceedings, published as a book by World Scientific.

There were four invited presentations: Ohsuga, Bradshaw, Zytkow, Ling. Dr. Ohsuga in his presentation "How can AI systems deal with large and complex problems - Model building as problem solving" analyzed reasons for increasing introduction of complex problems in our lives and into AI. The major way to tackle with complex problems is to design advanced problem models that enable automatic program generation.

Dr. Zytkow is one of the best-known researchers in the automatic discovery field. Learning by discovery is the most challenging subfield of machine learning, sometimes treated as stand-alone area related to intelligence and top human creative capabilities. Since agents have to be autonomous by definition, they have to create new knowledge, i.e. autonomously persue knowledge in purposeful interaction with the internal world. More autonomous means more able to make discovery in more complex circumstances. Zytkow's approach presented in "Robot-discoverer: a role model for any intelligent agent" might be welcome since at least some of the agents today seem to be only labeled as agents in order to join the mainstream. There is little doubt that agents capable of discovery - creating new knowledge, are more advanced than agents without this possibility.

Dr. Bradshaw from Boeing presented an invited paper "Steps toward the permanent colonization of cyberspace". Bradshaw's book on intelligent agents is among most often used textbooks on agents worldwide. His presentation was also one of the most futuristic. Bradshaw pointed out that agent research is a new field and that it quickly progresses and changes along the way. Today, there are several types of agents and several areas of agent research. As an interesting new application, the space ball capable of checking status of vital functions in a station was presented. The ball will soon be applied in real-life circumstances. However, it seems that the application was man-dominated thus not leaving room for full agent autonomy. The major part of the presentation was devoted to long-lived heterogeneous colonies of agents colonizing the cyber space. Such colonies need life- support services, essentials, legal and social services.

Dr. Ling, head of the Microsoft Richmond research laboratories, presented an invited paper on "Intelligent agents: embodied and disembodied". In the first part of the presentation he analyzed Bayesian networks and the advantages of this methodology. A limited version of Bayesian network is applied in the Microsoft Office Assistant; a slightly more advanced version is implemented in the troubleshooting agents. In the second part of his presentation, Ling presented embodied agents, capable of animation. They imitate visual effects of feelings by showing different facial expressions or by body language. This might be a desirable function in videoconferencing. Overall, in this way computers become more personal, closer to and more acceptable by humans.

The panel was another very interesting event. Bradshaw presented the old Apple animation showing a futuristic agent capable of information gathering and communication with a user. Many of these functions have not fulfilled yet. In other words it means that some challenges that agent face are the old challenges of artificial intelligence - how to simulate human-acceptable intelligence on computers.

Throughout the event, there were lots of interesting papers on agent research. One of the dominating research directions is the multi-agent research, the other is dealing with information overflow. Several papers dealt with information gathering from heterogeneous sources such as mobile telephone, the Web and TV.

We presented our employment agent in the single-track WS session. It seems that our idea of a uniform information gathering from any Internet employment database through normal HTML is original. Other approaches are usually based on agent languages such as KQML, or database wrappers. But these approaches demand cooperation from the database while in our approach the agent takes care of modifying its queries according to database forms.

Matjaž Gams

First Announcement and Call for Papers

JCKBSE 2000

Fourth Joint Conference on Knowledge-Based Software Engineering
Brno, Czech Republic, September 12–14, 2000

Organized by:

- Czech Society for Computer Science
- Department of Computer Science and Engineering, Faculty of Electrical Engineering and Computer Science, Brno University of Technology, Czech Republic
- SIG-KBSE, The Institute of Electronics, Information and Communication Engineers, Japan

In cooperation with:

- IEEE Tokyo Section Computer Chapter
- Japanese Society of Artificial Intelligence
- Russian Association for Artificial Intelligence
- Slovak Society for Computer Science
- The Institution of Electrical Engineering - Slovakia Centre
- VEMA Brno, computers and projects, Ltd., Brno

Steering Committee:

- Christo Dichev, IIT, Bulgarian Academy of Sciences
- Morio Nagata, Keio University
- Pavol Navrat, Slovak University of Technology
- Vadim L. Stefanuk, IITP, Russian Academy of Sciences
- Haruki Ueno, Tokyo Denki University

About the Conference

JCKBSE aims to provide a forum for researchers and practitioners to discuss the latest developments in the areas of knowledge engineering and software engineering. Particular emphasis is placed upon applying knowledge-based methods to software engineering problems. The Conference originated in order to provide a forum in which the latest developments in the field of knowledge-based software engineering could be discussed. Although initially targeting scientists from Japan, the CIS countries and countries in Central and Eastern Europe, JCKBSE warmly welcomes participants from all countries. JCKBSE 2000 will continue with this tradition and is anticipating even wider international participation. Furthermore, the scope of the conference as indicated by its topics has been updated to reflect the recent development in all the three covered areas, i.e. knowledge engineering, software engineering, and knowledge-based software engineering. The conference will include several invited talks and a plenary talk by a distinguished speaker.

Topics

- architecture of knowledge, software and information systems including collaborative, distributed, multi-agent and multimedia systems, internet and intranet,
- requirements engineering, domain analysis and modeling, formal and semiformal specifications,
- knowledge engineering for domain modeling, system family engineering, and software product lines,
- intelligent user interfaces and human-machine interaction,
- knowledge acquisition and discovery, data mining,
- automating software design and synthesis,
- object-oriented and other programming paradigms, metaprogramming,
- reuse, re-engineering, reverse engineering,
- knowledge-based methods and tools for software engineering, including testing, verification and validation, process management, maintenance and evolution, CASE,
- decision support methods for software engineering,
- applied semiotics for knowledge-based software engineering,
- knowledge systems methodology, development tools and environments,
- practical applications and experience of software and knowledge engineering,
- information technology in control, design, production, logistics and management,
- enterprise modeling, workflow,
- knowledge management for business process,
- intelligent agents for software engineering,
- program understanding, programming knowledge, learning of programming, modeling programs and programmers,
- knowledge-based methods and tools for software engineering education,
- software engineering and knowledge engineering education, distance learning.

Program Committee

Bailes P. (Australia), Banaszak Z. (Poland), Benczur A. (Hungary), Bielikova M. (Slovakia), Brusilovsky P. (USA), Devedzic V. (Yugoslavia), Dichev Ch.V. (USA),

Dicheva D. (USA), Dochev D. (Bulgaria), Ehrlich A. (Russia), Eisenecker U. (Germany), Far Behrouz H. (Japan), Fukazawa Yoshiaki (Japan), Gams M. (Slovenia), Gladun V. P., (Ukraine), Hashimoto Masa-aki (Japan) – co-chair, Hori Masahiro (Japan), Hruska T. (Czech Republic) – co-chair, Kaijiri Kenji (Japan), Kang Kyo-Chul (Korea), Khoroshevsky V.F. (Russia), Komiya Seiichi (Japan), Koyama Teruo (Japan), Kumeno Fumihiro (Japan), Lloyd-Williams M. (UK), Lozovskiy V. S. (Ukraina), Moessenboeck H. (Austria), Molnar L. (Slovakia), Nagata Morio (Japan), Navrat P. (Slovakia), Okamoto Toshio (Japan), Oomori Yasumasa (Japan), Osipov G.S. (Russia), Pechersky Y. (Moldova), Stefanuk V. L. (Russia), Sugawara Kenji (Japan), Tan Chew Lim (Singapore), Ueno Haruki (Japan), Welzer T. (Slovenia), Yamada Hiroyuki (Japan), Yamamoto Shuichiro (Japan), Young Gilbert H. (Hong Kong), Zendulka J. (Czech Republic).

Venue

Brno, the metropolis of the South Moravian region is the second largest city in the Czech Republic with a population of more than 400,000. From the north-east and the north-west the town is surrounded by promontories of the Drahany Uplands and the Czech-Moravian Highlands, while to the south Brno's streets run into gently undulated plains around the massif of the Palava Hills.
Brno is connected by a motorway and a railway with Prague and Bratislava and a railway and a good road with Vienna and Ostrava. There is also an international airport in Brno.
The conference is taking place in the Santon Hotel, which is located in one of the most popular recreational areas in Brno – the Brno Dam Lake. The distance from public transport facilities is approx. 200m. In Brno, it is possible to visit numerous cultural and social events, theatres, concert halls, musea and historical monuments.

Proceedings

All accepted papers will be published in the conference proceedings and will be available at the conference.
The official language of the conference is English.

Paper Submission

Full papers should not exceed 8 pages. Short papers should not exceed 4 pages. Papers will be reviewed according to: technical quality, originality, clarity, appropriateness to the conference focus, and adequacy of references to related work.
Authors should submit the papers electronically. For details see the conference web site. In addition, one copy of a manuscript should be sent, too. For details see the conference web site.

Important Dates

February 1, 2000 – Preliminary registration
March 1, 2000 – Paper submission deadline
May 1, 2000 – Notification of acceptance
May 20, 2000 – Camera-ready deadline
September 12–14, 2000 – Conference dates

Correspondence Address

JCKBSE 2000
Department of Computer Science and Engineering,
Brno University of Technology,
Božetěchova 2, 612 66 Brno, Czech Republic,
fax: +420-(0)5-4121 1141
e mail: jckbse@fee.vutbr.cz
www: http://www.fee.vutbr.cz/UIVT/JCKBSE/

Fees

The cost of an integrated package including the conference fee, accomodation (three nights), meals (starting with the dinner on September 11 and finishing with the lunch on September 14), and social activities is 350 USD (before July 20, 2000). After July 20, 2000 the cost of the integrated package is 400 USD.
For participants from the Central and Eastern Europe including the newly independent states, the cost of the integrated package is 150 USD. Ask about the fee if your country fits this condition.
For a limited number of students, the program committee will grant an additional financial support.
The fee for an accompanying person is 250 USD.
The cost of the integrated package including some special services (bus transport from the Vienna or Prague airport to Brno and back, a guided tour, a boat trip on the dam) is 550 USD. This program can be arranged for groups only.
We cannot refund a part of the fees in case you cannot stay for the whole period of the conference.

THE MINISTRY OF SCIENCE AND TECHNOLOGY OF THE REPUBLIC OF SLOVENIA

Address: Trg OF 13, 1000 Ljubljana,
Tel.: +386 61 178 46 00, Fax: +386 61 178 47 19.
http://www.mzt.si, e-mail: info@mzt.si
Minister: Lojze Marinček, Ph.D.

Slovenia realises that that its intellectual potential and all activities connected with its beautiful country are the basis for its future development. Therefore, the country has to give priority to the development of knowledge in all fields. The Slovenian government uses a variety of instruments to encourage scientific research and technological development and to transfer the results of research and development to the economy and other parts of society.

The Ministry of Science and Technology is responsible, in co-operation with other ministries, for most public programmes in the fields of science and technology. Within the Ministry of Science and Technology the following offices also operate:

Slovenian Intellectual Property Office (SIPO) is in charge of industrial property, including the protection of patents, industrial designs, trademarks, copyright and related rights, and the collective administration of authorship. The Office began operating in 1992 - after the Slovenian Law on Industrial Property was passed.

The Standards and Metrology Institute of the Republic of Slovenia (SMIS) By establishing and managing the systems of metrology, standardisation, conformity assessment, and the Slovenian Award for Business Excellence, SMIS ensures the basic quality elements enabling the Slovenian economy to become competitive on the global market, and Slovenian society to achieve international recognition, along with the protection of life, health and the environment.

Office of the Slovenian National Commission for UNESCO is responsible for affairs involving Slovenia's co-operation with UNESCO, the United Nations Educational, Scientific and Cultural Organisation, the implementation of UNESCO's goals in Slovenia, and co-operation with National commissions and bodies in other countries and with non- governmental organisations.

General Approaches – Science Policy

Educating top-quality researchers/experts and increasing their number, increasing the extent of research activity and achieving a balanced coverage of all the basic scientific disciplines necessary for:
- quality undergraduate and postgraduate education,
- the effective transfer and dissemination of knowledge from abroad,
- cultural, social and material development,
- promoting the application of science for national needs,
- promoting the transfer of R&D results into production and to the market,
- achieving stronger integration of research into the networks of international co-operation (resulting in the complete internationalisation of science and partly of higher education),
- broadening and deepening public understanding of science (long-term popularisation of science, particularly among the young).

General Approaches – Technology Policy
- promotion of R&D co-operation among enterprises, as well as between enterprises and the public sector,
- strengthening of the investment capacities of enterprises,
- strengthening of the innovation potential of enterprises,
- creation of an innovation-oriented legal and general societal framework,
- supporting the banking sector in financing innovation-orientated and export-orientated business
- development of bilateral and multilateral strategic alliances,
- establishment of ties between the Slovenian R&D sector and foreign industry,
- accelerated development of professional education and the education of adults,
- protection of industrial and intellectual property.

An increase of total invested assets in R&D to about 2.5% of GDP by the year 2000 is planned (of this, half is to be obtained from public sources, with the remainder to come from the private sector). Regarding the development of technology, Slovenia is one of the most technologically advanced in Central Europe and has a well-developed research infrastructure. This has led to a significant growth in the export of high-tech goods. There is also a continued emphasis on the development of R&D across a wide field which is leading to the foundation and construction of technology parks (high -tech business incubators), technology centres (technology-transfer units within public R&D institutions) and small private enterprise centres for research.

R&D Human Potential

There are about 750 R&D groups in the public and private sector, of which 102 research groups are at 17 government (national) research institutes, 340 research groups are at universities and 58 research groups are at medical institutions. The remaining R&D groups are located in business enterprises (175 R&D groups) or are run by about 55 public and private non-profit research organizatios.

According to the data of the Ministry of Science and Technology there are about 7000 researchers in Slovenia. The majority (43%) are lecturers working at the two universities, 15% of researchers are employed at government (national) research institutes, 22% at other institutions and 20% in research and development departments of business enterprises.

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 700 staff, has 500 researchers, about 250 of whom are postgraduates, over 200 of whom have doctorates (Ph.D.), and around 150 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 **Slove**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.

In the last year on the site of 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.

At the present time, part of the Institute is being 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 is being developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park will take the form of a shareholding company and will host an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of 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 Economic Relations and Development, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 61000 Ljubljana, Slovenia
Tel.:+386 61 1773 900, Fax.:+386 61 219 385
Tlx.:31 296 JOSTIN SI
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Contact person for the Park: Iztok Lesjak, M.Sc.
Public relations: Natalija Polenec

CONTENTS OF *Informatica* Volume 23 (1999) pp. 1–579

Papers

infrastructure for support of digital signatures. Informatica 23:477–481.

KREMIEN, O., K. MICHAEL & E. IRIT. 1999. Preserving mutual interests in high performance computing clusters. Informatica 23:41–48.

KRISPER, M. & T. ZRIMEC. 1999. Modelling of an information society in transition - Slovenia's position in the CE countries. Informatica 23:467–471.

KWONG, P. & S. MAJUMDAR. 1999. Scheduling of I/O in multiprogrammed parallel systems. Informatica 23:67–76.

LAURINI, R., K.-J. LI, S. SERVIGNE & M.-A. KANG. 1999. Modeling an auditory urban database with a field-oriented approach. Informatica 23:169–185.

LIN, W.-M. & W. XIE. 1999. Minimizing communication conflicts with load-skewing task assignment techniques on network of workstations. Informatica 23:57–66.

LIOTOPOULOS, F.K. 1999. Issues on gigabit switching using 3-stage Clos networks. Informatica 23:335–346.

MALEKOVIĆ, M. 1999. Agent properties in multiagent systems. Informatica 23:283–288.

MANOLAKOS, E.S. & D.G. GALATOPOULLOS. 1999. JavaPorts: An environment to facilitate parallel computing on a heterogeneous cluster of workstations. Informatica 23:97–105.

MARISITS, T., S. MOLNÁR & G. FODOR. 1999. Supporting all service classes in ATM: A novel traffic control framework. Informatica 23:305–315.

MARUGESAN, S. 1999. Intelligent agents on the Internet and Web: Applications and prospects. Informatica 23:437–443.

MICKLE, M.H. 1999. On the determination of absolute network performance. Informatica 23:383–387.

MORDONINI, M. & A. POGGI. 1999. SISTER: a flexible system for image retrieval. Informatica 23:549–558.

NANČOVSKA, I., A. JEGLIČ, D. FEFER & L. TODOROVSKI. 1999. Equation discovery system and neural networks for short-term DC voltage prediction. Informatica 23:513–520.

NONG, G., M. HAMDI & J.K. MUPPALA. 1999. Performance evaluation of a scheduling algorithm for multiple input-qued ATM switches. Informatica 23:369–381.

OMONDI, A.R. 1999. Floating-point arithmetic and the IEEE-754 standard, I: Number-system design. Informatica 23:413–429.

PROVOST, F. & A. POHORECKYJ DANYLUK. 1999. Problem definition, data cleaning, and evaluation: A classifier learning case study. Informatica 23:123–136.

RAKOVIĆ, D., M. TOMAŠEVIĆ, E. JOVANOV, V. RADIVOJEVIĆ, P. ŠUKOVIĆ, Ž. MARTINOVIĆ, M. CAR, D. RADENOVIĆ, Z. JOVANOVIĆ-IGNJATIĆ & L. ŠKARIĆ. 1999. Electroencephalographic (EEG) correlates of some activities which may alter consciousness: The transcendental meditation technique, musicogenic states, microwave resonance relaxation, healer/healee interaction, and alertness/drowsiness. Informatica 23:399–412.

RAYHAN, A., F. EGLUIBALY, & A. ALMULHEM. 1999. Fault-tolerant ATM switch using logical neighborhood network. Informatica 23:325–334.

SANG, J. 1999. High-performance cluster computing over gigabit/fast Ethernet. Informatica 23:19–27.

SETZ, T. 1999. Fault tolerant execution of computer-intensive distributed applications in LIPS. Informatica 23:87–95.

SHA, D. & V.B. BAJIĆ. 1999. Adaptive on-line ANN learning algorithm and application to identification of non-linear systems. Informatica 23:521–529.

SHEN K., W. LIANG & J. NG. 1999. Efficient computation of frequent itemsets in a subcollection of multiple set families. Informatica 23:543–547.

SICHERL, P. 1999. A new perspective in comparative analysis of information society indicators. Informatica 23:455–460.

ŠILC, J. & B. ROBIČ. 1999. Asynchronous microprocessors. Informatica 23:239–247.

VEHOVAR, V. & M. KOVAČIČ. 1999. Measuring information society: some methodological problems. Informatica 23:461–465.

VENKATESAN, R., Y. EL-SAYED, R. THUPPAL & H. SIVAKUMAR. 1999. Performance analysis of pipelined multistage interconnection networks. Informatica 23:347–357.

ZAZULA, D., B. VIHER, D. KOROŠEC, E. AVDIČAUŠEVIČ, M. LENIČ & B. POTOČNIK. 1999. Conceptual interactive learning tools based on computer simulators.

Editorials

A Conference Report

Calls for Papers

Professional Societies

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit three copies of the manuscript with good copies of the figures and photographs to one of the editors from the Editorial Board or to the Contact Person. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible directly on the manuscript, from typing errors to global philosophical disagreements. The chosen editor will send the author copies with remarks. If the paper is accepted, the editor will also send copies to the Contact Person. The Executive Board will inform the author that the paper has been accepted, in which case it will be published within one year of receipt of e-mails with the text in Informatica LaTeX format and figures in `.eps` format. The original figures can also be sent on separate sheets. Style and examples of papers can be obtained by e-mail from the Contact Person or from FTP or WWW (see the last page of Informatica).

Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the Contact Person.

QUESTIONNAIRE

☐ Send Informatica free of charge

☐ Yes, we subscribe

Please, complete the order form and send it to Dr. Rudi Murn, Informatica, Institut Jožef Stefan, Jamova 39, 61111 Ljubljana, Slovenia.

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 five 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 the European computer science and informatics 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://ai.ijs.si/informatica/
http://orca.st.usm.edu/informatica/

Referees:

Witold Abramowicz, David Abramson, Adel Adi, Kenneth Aizawa, Suad Alagić, Alan Aliu, Richard Amoroso, John Anderson, Hans-Jurgen Appelrath, Grzegorz Bartoszewicz, Catriel Beeri, Daniel Beech, Fevzi Belli, Istvan Berkeley, Azer Bestavros, Balaji Bharadwaj, Jacek Blazewicz, Laszlo Boeszoermenyi, Damjan Bojadžijev, Jeff Bone, Ivan Bratko, Jerzy Brzezinski, Marian Bubak, Leslie Burkholder, Frada Burstein, Wojciech Buszkowski, Rajkumar Bvyya, Netiva Caftori, Jason Ceddia, Ryszard Choras, Wojciech Cellary, Wojciech Chybowski, Andrzej Ciepielewski, Vic Ciesielski, David Cliff, Travis Craig, Noel Craske, Matthew Crocker, Tadeusz Czachorski, Milan Češka, Honghua Dai, Deborah Dent, Andrej Dobnikar, Sait Dogru, Georg Dorfner, Ludoslaw Drelichowski, Matija Drobnič, Maciej Drozdowski, Marek Druzdzel, Jozo Dujmović, Pavol Ďuriš, Hesham El-Rewini, Warren Fergusson, Pierre Flener, Wojciech Fliegner, Vladimir A. Fomichov, Terrence Forgarty, Hans Fraaije, Hugo de Garis, Eugeniusz Gatnar, James Geller, Michael Georgiopolus, Jan Goliński, Janusz Gorski, Georg Gottlob, David Green, Herbert Groiss, Inman Harvey, Elke Hochmueller, Rod Howell, Tomáš Hruška, Alexey Ippa, Ryszard Jakubowski, Piotr Jedrzejowicz, A. Milton Jenkins, Eric Johnson, Polina Jordanova, Djani Juričič, Sabhash Kak, Li-Shan Kang, Roland Kaschek, Jacek Kierzenka, Jan Kniat, Stavros Kokkotos, Kevin Korb, Gilad Koren, Henryk Krawczyk, Ben Kroese, Zbyszko Krolikowski, Benjamin Kuipers, Matjaž Kukar, Aarre Laakso, Phil Laplante, Bud Lawson, Ulrike Leopold-Wildburger, Joseph Y-T. Leung, Xuefeng Li, Alexander Linkevich, Raymond Lister, Doug Locke, Peter Lockeman, Matija Lokar, Jason Lowder, Kim Teng Lua, Andrzej Małachowski, Bernardo Magnini, Peter Marcer, Andrzej Marciniak, Witold Marciszewski, Vladimir Marik, Jacek Martinek, Tomasz Maruszewski, Florian Matthes, Timothy Menzies, Dieter Merkl, Zbigniew Michalewicz, Roland Mittermeir, Madhav Moganti, Reinhard Moller, Tadeusz Morzy, Daniel Mossé, John Mueller, Hari Narayanan, Rance Necaise, Elzbieta Niedzielska, Marian Niedq'zwiedziński, Jaroslav Nieplocha, Jerzy Nogieć, Stefano Nolfi, Franc Novak, Antoni Nowakowski, Adam Nowicki, Tadeusz Nowicki, Hubert Österle, Wojciech Olejniczak, Jerzy Olszewski, Cherry Owen, Mieczyslaw Owoc, Tadeusz Pankowski, William C. Perkins, Warren Persons, Mitja Peruš, Stephen Pike, Niki Pissinou, Ullin Place, Gustav Pomberger, James Pomykalski, Gary Preckshot, Dejan Rakovič, Cveta Razdevšek Pučko, Ke Qiu, Michael Quinn, Gerald Quirchmayer, Luc de Raedt, Ewaryst Rafajlowicz, Sita Ramakrishnan, Wolf Rauch, Peter Rechenberg, Felix Redmill, David Robertson, Marko Robnik, Ingrid Russel, A.S.M. Sajeev, Bo Sanden, Vivek Sarin, Iztok Savnik, Walter Schempp, Wolfgang Schreiner, Guenter Schmidt, Heinz Schmidt, Dennis Sewer,Zhongzhi Shi, William Spears, Hartmut Stadtler, Olivero Stock, Janusz Stokłosa, Przemysław Stpiczyński, Andrej Stritar, Maciej Stroinski, Tomasz Szmuc, Zdzislaw Szyjewski, Jure Šilc, Metod Škarja, Jiří Šlechta, Chew Lim Tan, Zahir Tari, Jurij Tasič, Piotr Teczynski, Stephanie Teufel, Ken Tindell, A Min Tjoa, Wieslaw Traczyk, Roman Trobec, Marek Tudruj, Andrej Ule, Amjad Umar, Andrzej Urbanski, Marko Uršič, Tadeusz Usowicz, Elisabeth Valentine, Kanonkluk Vanapipat, Alexander P. Vazhenin, Zygmunt Vetulani, Olivier de Vel, John Weckert, Gerhard Widmer, Stefan Wrobel, Stanislaw Wrycza, Janusz Zalewski, Damir Zazula, Yanchun Zhang, Zonling Zhou, Robert Zorc, Anton P. Železnikar

EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering the European computer science and informatics 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 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.

Volume 23 Number 4 December 1999 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics