User Annotations as a Context for Related Document Search on the Web and Digital Libraries

Jakub Ševcech, Róbert Móro, Michal Holub and Mária Bieliková
Faculty of Informatics and Information Technologies, Slovak University of Technology
Ilkovičova 2, 842 16 Bratislava, Slovakia
E-mail: {jakub.sevcech, robert.moro, michal.holub, maria.bielikova}@stuba.sk

Keywords: annotation, related document, search, query construction

Received: November 23, 2013

In this digital age, a lot of documents that people read are accessed through the Web and read on-line. There are various applications and services which enable creating bookmarks, tags, highlights and other types of annotations while reading these electronic documents. Annotations represent additional information on a particular information source and indicate that documents or their sections are somehow interesting for the document reader. However, existing approaches lack immediate reward for content annotation. We propose a method for query construction enabling search for other documents related to the currently studied one using not only the document's content, but also user created annotations as indicators of user's interests. In our proposed approach, annotations are used to activate nodes in a graph created from the document's content employing spreading activation algorithm. We evaluate the proposed method in Annota - a service for bookmarking and collaborative annotation of Web pages and PDF documents displayed in a web browser. Along with its main purpose, Annota is designed to support scenarios useful for a novice researcher working together with his or her mentor. Based on Annota usage data we also analyzed properties of various types of annotations. Discovered annotation properties served as a basis for simulation we performed to determine optimal parameters of the query construction. We compared the proposed method to the commonly used tf-idf based method which our method outperformed when using annotations in the query construction process by improving the overall precision of the document retrieval. Therefore, annotations proved to be a viable source of information for user's interest detection.

Povzetek: Razvita je metoda, ki pri delu s spletnimi besedili uporablja oznake v besedilu.

1 Introduction

While reading printed documents, a common practice is to write down various types of notes. We use them as means of storing our thoughts, to highlight interesting parts of the document and to ease navigation in the printed document. Many tools and services allow us to create similar notes in electronic documents as well. We can create various bookmarks, tags, highlights and other types of annotations while surfing the Web or when reading electronic documents. In contrast to notes written in printed documents, electronic annotations are often objects of further processing and they can serve as means of improving intra and inter document navigation, to organize personal collections of documents, to search for documents, etc.

There is active research in the field of utilization of annotations [1] and patterns [2] their users follow when creating or making use of these annotations. Various types of annotations can be used for user interest identification [3], user modeling [4] and subsequently for personalization or additional support while searching for resources.

Annotations, created by user, can be considered a form of user's context he or she creates while reading documents and traveling in digital space [5]. This context can take various forms depending on the used annotation type, such as thoughts stored as short notes attached to the document as a whole, comments to specific sections of the document, or highlighted document sections that are in some way interesting to the reader. Many applications use annotations as a means of navigation between documents and for organizing content. For example, in [6] the authors describe an organization of learning materials and collaboration of students while learning to use an educational system that provides students the possibility to attach various types of annotations to learning objects. The study of various search tasks supported by a social bookmarking service

[*] This paper is based on J. Ševcech and M. Bieliková, *Query Construction for Related Document Search Based on User Annotations* published in the proceedings of the 3rd International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2013 conference).

deployed in a large enterprise is presented in [7]. The authors concluded that bookmarking services and annotations attached to documents can enhance document organization and social navigation.

User generated tags are one of the most commonly used methods for organizing the content, because of their utility and applicability for various content types. They have been successfully used to organize various media files, e.g. photos, videos, and documents in many real world applications such as bookmarking services Diigo[1] or Delicious[2]. Further types of annotations, such as highlights and comments can serve to create custom in-document navigation. The users can categorize or describe resources [8] and thus create navigation that fits their needs without relying on navigation provided by document's author.

User created annotations can be used not only to support inter or intra document navigation. Tags are used for folksonomy construction [9], annotations can play an important role in content enrichment and content quality improvement, e.g. in an educational system, as presented in [6]. In this system the authors use content error reports, user generated comments and questions, to improve course content and other types of annotations, such as tags and highlights, for the navigation and even the content summarization [10].

Currently, there are many services allowing users to annotate the documents. However, all of these applications motivate users to create annotations by a prospect of future improvement of inter or intra document navigation, i.e. users benefit from created annotations only after there are enough annotated documents, or when returning to previously annotated document. Problem with this approach is that there is a lack of immediate reward after annotation is created.

The rest of the paper is structured as follows. In section 2 we further analyze different approaches for utilization of annotations in the search process. Annota - a service for web page bookmarking and annotation, that allows users to insert various types of annotations to Web pages and PDF documents displayed in Web browser, is introduced in section 3. We describe multiple applications and usage scenarios that are supported by annotations the users attach to documents emphasizing Annota's unique features compared to existing similar systems. In section 4 we propose a method for query construction from currently studied document and its attached annotations as one of document annotation applications. This method produces a query that can be used in related document retrieval where the query is taking into account user's interest provided by created annotations. The query is created while the user is reading the documents and it is used to search for related documents to the currently studied one. The reward for user creating annotations is thus provided during the time of annotation creation. We evaluate the proposed method using synthetic as well as online experiments in the Annota system in section 5 and conclude by discussing the method's properties and implications for the area of research in section 6.

2 Related work

One of the possible employments of annotations in information processing is the document search. There are two possible approaches for exploitation of annotations in the search process. One is to use annotations while indexing documents by expanding documents in a similar way anchor texts are used [11], or using bookmarks and annotations as document quality indicators while ranking documents [12].

The second possible application of annotations is in the query expansion or in query construction process. An example of annotations used for query expansion is presented in [13], where tags attached to search results are used to expand initial query similarly to pseudo-relevance feedback. Multiple methods for query expansion in folksonomies are presented in [14]. Of particular interest are methods expanding queries by tags from folksonomies on the basis of semantic similarity between words of the query and these tags.

An example of annotations used as queries to retrieve related documents is presented in [15]. The authors asked users to read a set of documents and to create annotations in documents using a tablet. They used these annotations as queries in related document search. They compared search precision of these queries with relevance feedback expanded queries. Queries derived from user's annotations produced significantly better results than relevance feedback queries.

More often, when creating queries for related document retrieval, the document's content is used instead of attached annotations. In [16] authors used the most important phrases from the source document as queries for document retrieval. Another work dealing with search for related documents is described in [17] where the authors use related document search as a means of recommendation of citations into unpublished manuscripts. They use text-based features of the document to retrieve similar documents and citation features to establish authority of documents. Similar document retrieval has also its application in document recommendation. In work presented in [18] a list of documents similar to those visited by the users were used as a form of content based recommendation of related documents.

In popular search engines such as ElasticSearch[3] and Apache Solr[4], term frequency is used in the query construction process. They provide special type of query interface called "more like this" query, which processes source text and returns a list of similar documents. Internally, the search engine extracts the most important words using tf-idf metric from source text and it uses the

[1] Diigo, http://www.diigo.com/
[2] Delicious, https://delicious.com/
[3] ElasticSearch, http://www.elasticsearch.org/
[4] Apache Solr, http://lucene.apache.org/solr/

most important words as a query for related documents search.

In most applications, the similar document retrieval process consists of two phases. In the first step, queries in the form of the most important phrases and, more often, the most important terms are extracted from the document's content. In the second step, these queries are used to retrieve other documents. In order to retrieve these most important terms from document's content, many different methods are used. Mostly, they are based solely on the term frequency in the document (such as already mentioned tf-idf based method) but many other methods are applicable. One possible category of methods for query term extraction are methods based on ATR (automatic term recognition) algorithms [19].

In multiple works authors showed that annotations represent important source of information for document retrieval. Methods for query construction for document retrieval however, use only document's content and information about the document collection in query construction process. They do not utilize user created annotations as user's interest indicators when creating query for document retrieval. We believe that annotations used in query construction process can significantly improve related document retrieval precision.

In our work we propose and evaluate a method for query construction from the document content enhanced by user created annotations. Annotations are used as interest indicators to determine parts of the document the user is most interested in. Using user created annotations our method creates a keyword query for related document search taking into account the user's interests. Proposed method is used in social bookmarking service Annota to retrieve related documents to the currently studied document. Annotations are used in related document retrieval in time of their creation and they provide immediate motivation for additional annotation creation in the form of related document search.

3 Service for Web page annotation

We developed a service called Annota[5] [20], which allows users to attach annotations to arbitrary web pages or PDF documents displayed in a web browser. Annota was created as a system to study methods for document search, navigation and organization on the Web. We uniquely employ annotations created by users in various methods of information retrieval, especially in digital libraries. In this domain, Annota supports various scenarios of collaboration: between a novice researcher and his or her supervisor (mentor), or between more researchers working on a joint project.

A few projects for supporting researchers already exist. Mendeley[6] allows users to organize and annotate documents via a desktop application and web interface. ResearchGate[7] is specialized to connect researchers while allowing them to add their own publications, follow others and ask research-related questions.

Annota provides environment to collaboratively collect documents while attaching annotations to them. Annota's unique features include annotation of documents directly on the Web as well as support for collaborative features such as bookmark sharing within groups and following other users of the service. Annota is realized as a client-server system. Client is represented by a browser extension allowing annotation of web pages. Annotations are stored on the server together with the identification of the resource (its URL) and additional metadata. The browser extension allows users to create various types of annotations, such as:

- tags,
- highlights,
- comments attached to selected text, and
- notes attached to the document as a whole.

Although Annota can be used on every web page, our target domain are digital libraries used by researchers in the field of information technologies, for which we provide additional support and tools. Annota stores metadata on various entities from digital libraries (authors, papers, conferences, etc.). We get this information by parsing web pages of selected digital libraries the users of Annota visit. When a user bookmarks a page containing metadata about a paper, Annota creates bibliographic reference to it. We realized the possibility to insert annotations into arbitrary web pages, articles in digital libraries and PDF documents displayed in web browser, by bookmarking and sharing documents and annotations.

The Annota service allows users to organize documents by tags, folders or faceted trees. It is possible to search in texts of documents contained in the user's library or in the library of bookmarked documents of all users. Besides keyword search, Annota offers various means of information space exploration, such as cloud of important terms, content of which is adapted by users' navigation history, i.e. by their previous queries [21], or navigation leads in the search results' summaries.

An example of a web page annotated using Annota is displayed in Figure 1. The figure shows a widget, where it is possible to bookmark displayed page, insert tags, edit note and share the bookmark with groups the user is member of. Users are able to highlight text fragments of the web page and to attach comments to these text selections.

The basic scenario of the service usage follows a user studying a document. The user has the following possibilities for particular activities:

- Bookmarking documents.
- Organizing the collection of documents using tags attached to individual bookmarks.
- Organizing the collection of documents by inserting the bookmarks into folders.
- Highlighting parts of the text and creating other types of annotations.
- Sharing bookmarked document in a group the user is member of via group sharing feature.

[5] Annota, http://annota.fiit.stuba.sk/
[6] Mendeley, http://www.mendeley.com/
[7] ResearchGate, http://www.researchgate.net/

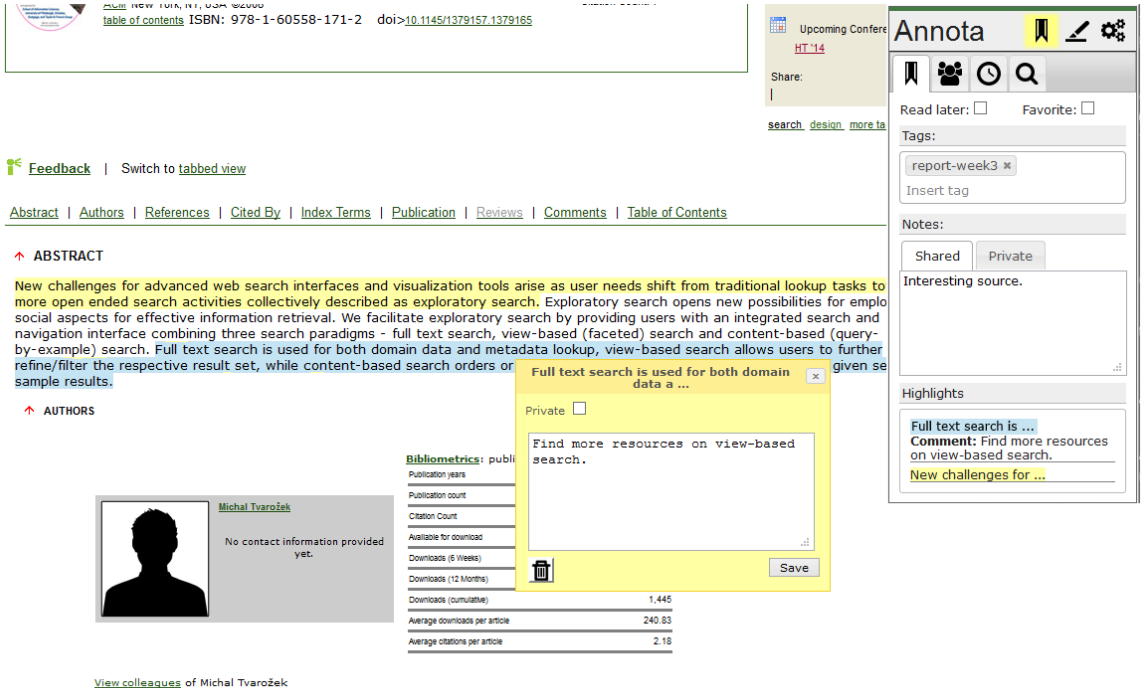

Figure 1: Web page in ACM DL annotated using bookmarking service Annota. It can be annotated collaboratively (highlights from different users are displayed in different colors).

- Following activity of other interesting users.

3.1 Annotation usage scenarios

Previously mentioned features are useful when a user works alone. However, research nowadays is being done by teams of collaborating people, sometimes composed by only two researchers (researcher novice and his supervisor or mentor), other times the teams are larger. In order to support collaboration of researchers within such teams, we support several scenarios of using Annota, namely:

- novice researcher scenario,
- paper authors scenario, and
- activity following scenario.

3.1.1 Novice researcher scenario

The novice researchers working on their projects obviously start by doing research on the state of the art in the research area of their interest. They usually read a lot of research articles, some of which are more useful and relevant to the target research topic than others. The researchers need tools to keep them organized in order to reference them later in their work. Moreover, the novice researchers need help from their respective mentors, who are expected to recommend useful resources their mentee should read.

Annota helps the novice researchers to organize the resources they read (using folders or tags) and to annotate the research papers. The researchers can use their own notes later, while writing the papers or preparing presentations.

Annota also allows the supervisors to create a group and invite the researcher they supervise to become a member of it. Then, the supervisors can share papers via this group, thus recommending important study material to their mentees. This is very helpful and thanks to that the novice researchers have a point from which to start searching for more information on their research topic.

Working in groups also enables the novice researchers to report the progress to their supervisor e.g. by using specialized tags (report-week3 for third week in semester as can be seen in Figure 1). Apart from one group per researcher, the supervisor might also pick the tactics of creating a group for all his mentees who share similar research topics. The users then share interesting research articles they have found together with their annotations and comments. The supervisors can add their own notes and help distinguish relevant publications or propose further readings.

In order to help the researchers with finding relevant papers, Annota allows them to search for papers already bookmarked by others. Since they also assign tags to the resources, the researcher might find an interesting paper on a certain topic easier than using only the search features provided by the digital library.

Annota can generate a report for the supervisor showing the activity of the group (or per user) for a selected period of time, containing overview of shared papers together with annotations. This allows the supervisors to continuously monitor the progress of the researchers they manage and effectively help them, a feature which is unique to Annota.

3.1.2 Paper authors scenario

In this scenario we consider a group of researchers doing research together. Part of every research is studying the work already done in the respective field. Collaborating researchers form a group in Annota and they can share

interesting publications with each other, comment and annotate them. These annotations can be later used when the researchers need to write a paper about their results, especially the "Related work" section.

A group in Annota needs not to be private. On the contrary, we encourage the groups to be public so that other users of Annota can see interesting resources the group has found together with their opinions. We allow every group to formulate its research goals in the form of tags (similarly to tags attached to documents). Users can find a group of their interest based on these tags.

3.1.3 Activity following scenario

We realize that collaboration and sharing of thoughts is very important for researcher in any field. Therefore, in Annota we allow its users to form social networks by following each other (a concept known mostly from Twitter[8]). When user A follows user B, the user A can see user B's newly added bookmarks and annotations, as well as other activities (joining of a group, following another user). When user A considers user B to be an authority in a field of his or her interest, this can keep the user A informed about the latest trends.

We do not limit the ability to follow someone just to Annota users. Since we gather freely available metadata about publications from various digital libraries, we allow the user to follow researchers, who are not Annota users. Furthermore, we allow them to follow interesting conferences, journals or publishers. This way the users are notified when their favorite researcher publishes a new paper, new issue of a journal or proceedings of their favorite conference are published, etc.

Moreover, the users of Annota can also follow the whole group, which enables them to see their newly added information. We believe the feature of following various entities (people, groups, publications, etc.) allows the whole community to grow and learn from each other and is an important feature to keep informed about the latest trends. Naturally, all the activities of the Annota users can be set to be private if they wish to keep their privacy. In such a case, nobody (not even the followers) sees them.

3.2 Creation of the Web page annotations

The browser extension created as a part of Annota service allows users to create annotations that link to document as a whole (tags, note) or to particular parts of the document (highlight, comment). As the extension is inserting annotations into web pages and they change frequently and without notification, we had to use a method for annotation linking to specified parts of the document that is resistant to changes in annotated document.

The key element in document annotation is the selection of a method to link documents and created annotations. Multiple systems supporting annotation creation assume that documents will not change after annotations are inserted. This is very strong assumption we cannot make in a domain such as web pages. We have to use method for annotation interlinking with document's content with regard to documents which may change over time. In [22] multiple criteria, which must be met by a robust method for locating annotations into documents, are defined. Some of the described criteria are:

- The method has to be robust to common changes in the referenced document.
- It has to be based on document's content.
- It has to work with uncooperative servers.
- The information necessary to locate annotation have to be relatively small compared to the document's content.

At the same time, in this work the authors suggest several approaches that meet these criteria. One of them is to use annotation context in form of surrounding text to place the annotation into the document. The method using document content to place annotations is defined also in Open Annotation Model [23]. It is tolerant to changes in the document content and when using approximate matching of strings it is also tolerant (to some extent) to changes in annotation context as well.

In order to attach annotations to document parts we use redundant representation of annotation location to support linking annotations into changing documents and to improve stability of annotation location. For locating annotation in the text, we store highlighted text with order of its in-text occurrence together with surrounding text. The combination of selected text and text occurrence order is tolerant to changes in the document's content except for changes in selected text and some changes before annotation location. With usage of approximate matching this method is to some extent tolerant to changes in selected text as well.

4 Method for query construction

Currently, the most common form of query used when searching for documents on the Web is the list of keywords. That is why the majority of methods for document retrieval using source document as query is transforming the document content into keyword queries. In order to retrieve words from the document to be used as query for related document search it is possible to use multiple different approaches. One of them is to extract most frequently occurring terms using the tf-idf metric or various ATR algorithms [19] as discussed in section 2. The tf-idf based method provides rather straightforward possibility to incorporate user created annotations: the source text of the document is extended by the content of created annotations, possibly with various weights for different types of annotations.

However, the method using the tf-idf for query word extraction takes into account only the number of occurrences of words in the source document (and document corpus). We believe that not only the number of word occurrences but also the structure of the source text is important when constructing a query for related

[8] Twitter, https://twitter.com

documents retrieval. Especially, if we suppose that while reading the document the users are usually interested in only a fraction of the document, this fraction is the place where they most probably attach an annotation.

We use user created annotations to increase weights of annotated parts of the document in query construction process and to attach additional content to the document. We proposed a method based on spreading activation in text of studied document transformed to a graph. The method uses annotations as interest indicators to extract parts of documents the user is most interested in. The proposed method consists of two phases:

1. Text to graph transformation that conserves word occurrence frequency in node degree and text structure in graph edges structure.
2. Graph nodes activation introduced by annotations attached to the document and query word extraction using spreading activation algorithm in created graph.

The text to graph transformation is based on word neighborhood in the text. The graph created from the text using words neighborhood conserves words importance in node degree, but it also reflects the structure of the source text in the structure of edges [23]. Such graph can be used for example the most important terms [24]. We use this graph to extract words are most important from point of view of the document reader and we use them as queries to retrieve similar documents.

4.1 Text to graph transformation

In order to transform text to a graph, it is first processed in several steps: segmentation, tokenization, stop-words removal and stemming. After these steps the initial text has a form of a list of words. Every unique word from this list is transformed into a single node of a graph. The edges of the graph are then created between two nodes if corresponding words in the text are neighbors or they are in the predefined maximal distance. This transformation is described by the following pseudocode:

```
tokens = text.downcase().split()
words = tokens.removeStopwords().stem()
length = words.size
nodes = words.uniq
edges = []
for(i=0;i<length;i++){
  for(j=i;i<min(i+dist,length-1);j++){
    edges.add(words[i], words[j])
  }   }
graph = Graph.new(nodes,edges)
```

As settings for maximal distance between words we used options described in [24], where two passages through the text with maximal distance set to two words and five words are used. By using these setting, the words with greater distance were connected and at the same time close words are better connected by bigger number of common edges. Created edges have the same weight but to speed up spreading activation process, we connected multiple edges between the same nodes and we set weights of the resulting edges to the number of connected edges.

4.2 Query word extraction

We use the graph representation of the text in order to find the most important nodes/words using spreading activation algorithm. This algorithm is commonly used for example to find the most related nodes in a graph to the initially activated node. The activation introduced into the initial node is spreading through the nodes of the graph and after the change in nodes activation is smaller than a specified threshold, the greatest amount of activation is concentrated in the most related nodes.

It is possible to use this algorithm for related nodes search, but also for other applications, such as keyword extraction [25]. We use the same intuition to extract the most important words to sections user is most interested in. We utilized user created annotations as their interest indicators. These annotations are used to introduce initial activation to nodes annotations are attached to. The initial activation is propagating through the graph and it is concentrating in most important words of the text. An example visualization of text transformed to graph and nodes activated by attached annotations is displayed on Figure 2. The node size reflects activation level and edge thickness number of edges between nodes. Colored nodes represents nodes with highest activation level, thus words selected as query for related document search.

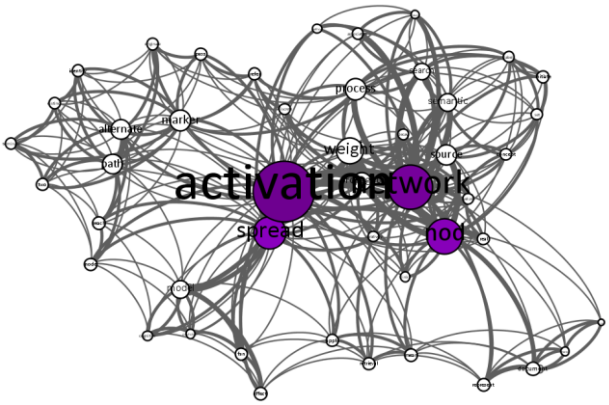

Figure 2: Example of text transformed to graph with activation spread across nodes.

When using annotations to insert initial activation into the document graph we consider separately annotations that are:

- highlighting parts of the document and
- inserting additional content into the document.

The proposed method takes into account both types. Those, which highlight parts of the document, contribute by activation to nodes representing words of highlighted part of the document and those enriching content of the document are extending the document graph by adding new nodes and edges and they are inserting activation to this extended part of the graph. When inserting activation to extended parts of the document we assume that some portion of the words used in the annotation content are located in the document text as well. The activation from

the extended part of the graph can then pass to the rest of the graph through common nodes.

When initial activation is spreading through the created graph, the nodes where activation is concentrating are the most important words of the graph and are considered words fit for the query. In our case, the activation is inserted into the graph through annotations attached to the document by its reader.

The proposed method is able to extract words, which are important for annotated part of the document, but it is also able to extract globally important words, that are important for document as a whole. The portion of locally and globally important words can be controlled by the number of iterations of the algorithm. With increasing number of iterations the activation is spreading from activated part of the document and extracted locally important words are changed to globally important words. When using this method it is thus important to determine when to stop the algorithm to find the best portion of globally and locally important words. It is also important to determine the right amount of activation inserted into the graph by various types of annotations. We determine these settings using simulation based on real user data while evaluating proposed method. The simulation is described in the next section of this paper.

The method for query word extraction uses annotations to insert initial activation into text transformed to graph. In case when no annotations are attached to the document, it is possible to extract globally important words from the document by activating the whole document's text.

5 Evaluation

In order to evaluate related document retrieval we performed both synthetic tests on dataset extracted from Wikipedia articles and online experiment with users of Annota bookmarking service.

5.1 Related document retrieval

We analyzed behavior of users of Annota while annotating documents using browser extension. Our experiments are based on usage data of 82 users who created 1 416 bookmarks and 399 in-text annotations during 4 months long period of using Annota on day-to-day basis. We studied multiple parameters of created annotations and we derived probabilistic distributions of these parameters. We studied properties such as the note length, number of highlights per user and per document, highlighted text length or probability of comment to be attached to highlighted text. We used extracted annotations properties and knowledge about their distributions in further evaluation. All observed parameters were following logarithmic or geometric distributions. Figure 3 displays an example of derived distribution for number of highlighted texts per document that follows logarithmic distribution.

Using various attributes of annotations and their probabilistic distributions we created a simulation, to find optimal weights for various types of annotations and number of iterations of proposed method for query construction from document text and attached annotations. We optimized query construction for document search precision.

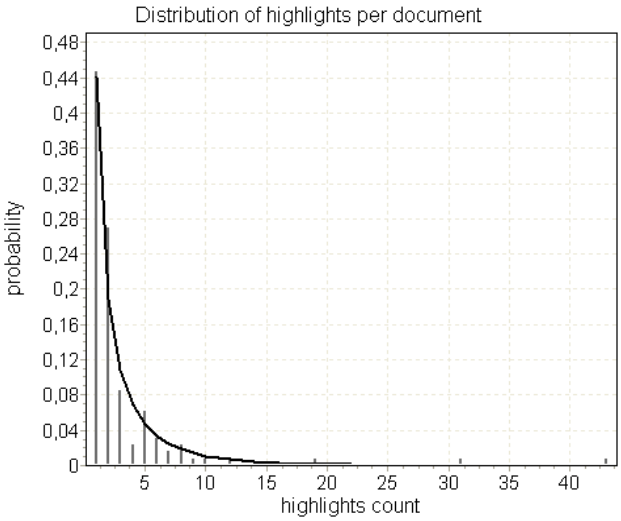

Figure 3: Logarithmic distribution of highlighted texts number per document.

The simulation was performed on the dataset we created by extracting documents from Wikipedia articles written in English. We constructed the source documents with aim to create documents containing several similar sections (from the point of view of used words) and with different topics. These generated documents simulate documents, where the user is interested in only a fraction of the content. In order to create such documents we used disambiguation pages in Wikipedia. The disambiguation page disambiguates multiple meanings of the same word and contains links to pages for each of these meanings. By combining abstracts of pages describing different meanings of the same word into single document, we simulate sections of the text describing multiple topics.

We downloaded all disambiguation pages and we selected random subset of these pages for which we downloaded pages they are linking to. Along with these disambiguated documents we downloaded all documents, having common category with at least one of disambiguated documents. We used search engine ElasticSearch to create an index of all downloaded documents and to search within this index. The parameters of created dataset are summarized in Table 1.

Table 1: Parameters of dataset used in simulation

Attribute	Number
All disambiguation pages	226 363
Selected disambiguation pages	86
Pages disambiguation pages are linking to	629
Categories	2 654
All downloaded pages	232 642

In the simulation we generated annotations in a way to correspond with probabilistic distributions extracted from the annotations created by users of the Annota service. From every disambiguation page and the pages it

was linking to, we created one source document by combining abstracts of all pages in random order. For every source document we selected one of the composing abstracts, which simulated one topic user is most interested in. We generated both annotations highlighting parts of the document and annotations inserting additional content for selected abstract. The highlights were randomly distributed over the whole abstract. To simulate the content of annotations extending content of the document (notes, comments) we used random parts of the page annotated abstract was extracted from.

Generated annotations along with source document content were used to create query using proposed method for query construction. The created query was used for related documents search in the index of all downloaded documents. When evaluating relevance of retrieved documents, we considered document to be relevant if it was from the same category as the page annotated abstract was extracted from.

We performed a simulation with several combinations of parameters and we implemented hill climbing algorithm to optimize parameter weights combination for the highest document search precision. Single iteration of performed simulation is described by following pseudocode:

```
for disambig in disambiguations do
  abstracts = disambig.pages.abstracts
  for abstract in abstracts do
    text = abstracts.shuffle.join(" ")
    graph = Graph.new(text)
    annot = Annotation.create(abstract)
    graph.activate(annot, weights)
    graph.spreadActivation()
    query = graph.topNodes
    results = ElasticSearch(query)
    cat = abstract.page.categories
    relevant = results.withCategory(cat)
  end
end
```

We compared search precision for proposed method and for tf-idf based method ("more like this" query) provided by ElasticSearch when searching for 10 most relevant documents. For the purpose of comparison of the proposed method with method based on tf-idf when using annotations in the query construction process, we performed an extension of the tf-idf based method to use annotations in query word extraction process. This method uses word frequency to find the most important words in the text. We extended the text of the document by text annotations were attached to and annotations content. We provided different weights for different annotations types by repeated extension of document by highlighted text and annotations content. We determined the optimal number of repetitions using parameter optimization with hill climbing algorithm, similarly to simulation for parameter estimation for method based on spreading activation in text transformed to graph.

Along with simulation using generated annotations for methods comparison, we performed two experiments to determine retrieval precision with no annotations and when whole abstract of the source document was highlighted. We aimed to determine the precision of compared methods when no annotations are available and while having complete information about user's interests. Results for simulations with generated annotations along with experiments with no annotations and with whole document fragment annotated are summarized in Table 2.

Table 2: Simulation results for spreading activation based method and tf-idf based method.

Method	Precision
Tf-idf based with no annotations	21.32%
Proposed with no annotations	21.96%
Tf-idf based with generated annotations	33.64%
Proposed with generated annotations	37.07%
Tf-idf based with whole fragment annotated	43.20%
Proposed with whole fragment annotated	53.34%

Proposed method based on spreading activation obtained similar or better results to tf-idf based method in all performed experiments. The results of experiments with no annotations, where only the content of the document was used to create query, suggests that proposed method provides similar, even better results for query word extraction. These results were achieved despite the fact that proposed method is using only information from the document content and not the information about other documents in the collection by contrast to tf-idf based method. The proposed method can thus be used as an alternative to tf-idf based method when creating query from document content.

The experiments with generated annotations and whole text fragment annotated suggests that proposed method outperforms tf-idf based method when annotations are used in query construction process. We performed a Student's t-test on 5% level of significance for pairs of proposed method and baseline method which showed statistically significant difference in mean precisions for compared methods when using generated annotations and whole abstracts annotated in query construction process (p-value < 0.01%).

The comparison of both methods without using annotations and using generated annotations in query construction process proved that annotations can increase precision of related documents retrieval. The experiment with whole document fragments annotated suggests that with increasing number of annotations the precision of generated queries increases for both used methods.

5.2 Online experiment in Annota bookmarking service

In order to compare the real increase of precision of document retrieval method with and without annotations we performed a qualitative user study where 8 volunteers were asked to annotate documents of their choice stored in Annota. Afterwards, we generated two queries, one with and another one without annotations. We retrieved two lists of documents using these queries and we presented them to volunteers in random order. They were

asked to select documents describing a topic related to the topic of source document from displayed lists and to select more relevant from two presented lists.

The volunteers annotated 11 unique documents. In 9 cases they selected the list created by the method using annotations as more relevant one. In one case the method using annotations created query in Slovak we found no relevant documents. This was caused by the fact that in this document all annotations were written in Slovak and all documents we searched in were in English. In one case the method not using annotations obtained better results. We obtained 34 relevant documents using method with annotations compared to only 15 documents returned by method without annotations.

Part of volunteers were writing annotations in Slovak, but to keep conditions the same as during document annotation out of the experiment, we allowed them to write annotations the same way they are used to. We asked one user to repeat the experiment on one document after he translated created annotations written in Slovak to English. When translated annotations were used in query construction all retrieved results were related to the source document.

In one case we asked the volunteer to repeat the experiment with increased number of annotations attached to the document. During this repeated experiment, the volunteer doubled the number of attached annotations. In the second retrieved list of documents, the number of relevant documents retrieved increased and the list included one exact match with the topic user was most interested in. With increasing number of annotations attached to document the precision of related document retrieval increases.

When using annotations to create a query, the proposed method retrieved more relevant documents in greater number than in the case when annotations were not used. Also, using annotations in to create queries, we retrieved more documents describing the same, as well as related topic as the source document.

We used a questionnaire about user's habits when annotating documents to determine how users of Annota are creating annotations into studied documents. The majority of participants are using annotations while reading printed or electronic documents. When annotating electronic documents, they use various tools to create bookmarks, to-do lists, saving documents for later, to insert highlights, comments and other types of annotations into documents. The most frequently used types of annotations are tags and in-text highlights. The purpose of creating annotations such as notes, comments and highlights is to summarize studied documents, describe documents, highlight most important sections, and store their thoughts about studied documents and as a form of in-document navigation to support fast recollection of document when returning to previously studied document. Interviewed volunteers confirmed our assumption that using annotations users are indicating those parts of the document they are most interested in.

6 Conclusions

Annotations represent a significant source of information on interesting or important parts of the documents. Their importance increases with possibilities for manipulating documents on the Web in the same way as printed documents and with possibilities for further processing and utilizing of the created annotations. We introduced Annota – a service for bookmarking and annotating Web documents while focusing on the domain of digital libraries. We described several scenarios where annotations can be useful. We studied users' behavior while annotating documents on the Web and proposed a method for query construction from document's content and attached annotations. For this purpose we considered document's content and its structure by using text to graph transformation and query terms extraction using spreading activation introduced by attached annotations.

The simulation based on probabilistic distributions of various parameters of annotations created by the users of Annota proved, that using annotations when creating queries for related document retrieval can increase retrieval precision and with increasing number of attached annotations the precision rises.

We compared two methods for query word extraction. The method based on spreading activation in document text transformed to graph outperforms tf-idf based method when creating query for related documents search from source document and attached annotations. The proposed method achieved comparable results to tf-idf based method when no annotations were used in the process of query construction. It is thus possible to use it even when no annotations are attached to the document with comparable precision as commonly used method when extracting words suitable for query for related document retrieval. The spreading activation based method outperformed baseline method when annotations attached to documents were used in query construction process. The proposed method does not use information from other documents, only information from the content of the source document and its attached annotations. It is thus search engine independent and can be used to create queries for any search engine accepting queries in the form of a list of keywords.

Performed user study showed that users insert annotations into document sections they are most interested in and they use annotations to summarize documents, highlight most important parts of documents and to store their thoughts. In connection with comparison of related document retrieval precision of proposed method and commonly used method when using annotations in query construction process and when no annotations were used, we showed that annotations can be used as user interest indicators in query construction for related document retrieval and that they improve related document retrieval precision.

Acknowledgement

This work was partially supported by the projects VG1/0675/11, VG1/0971/11 and APVV-0208-10. The authors wish to thank other members of Annota team, namely Roman Burger, Martin Lipták, Juraj Kostolanský, Peter Macko and Samuel Molnár for their contribution to design and implementation of selected components of Annota.

References

[1] M. Agosti, N. Ferro (2007). A formal model of annotations of digital content. *ACM Trans. Inf. Syst.*, vol. 26, no. 1.

[2] S.A. Golder, B.A. Huberman (2006). Usage patterns of collaborative tagging systems." *Journal of Information Science*, vol. 32, no. 2, pp. 198-208.

[3] X. Wu, L. Zhang, Y. Yu (2006). Exploring social annotations for the semantic web. *Proc. of the 15th Int. Conf. on World Wide Web (WWW '06)*, ACM, pp. 417-426.

[4] R. Wetzker, C. Zimmermann, C. Bauckhage, S. Albayrak (2010). I tag, you tag: translating tags for advanced user models. *Proc. of the 3rd ACM Int. Conf on Web Search and Data Mining (WSDM '10)*, ACM, pp. 71-80.

[5] P. Návrat (2012). Cognitive traveling in digital space: from keyword search through exploratory information seeking. *Central European Journal of Computer Science,* vol. 2, no. 3, pp. 170-182.

[6] M. Šimko, M. Barla, V. Mihál, M. Unčík, M. Bieliková (2011). Supporting Collaborative Web-based Education via Annotations. *World Conf. on Educational Multimedia, Hypermedia and Telecommunications*, pp.2576-2585.

[7] D. Millen, M. Yang, S. Whittaker, J. Feinberg (2007). Social bookmarking and exploratory search. *ECSCW 2007*, Springer, London, pp. 21–40.

[8] C. Körner, R. Kern, H. P. Grahsl, M. Strohmaier (2010). Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. *Proc. of the 21st ACM Conf. on Hypertext and Hypermedia*, ACM, pp. 157-166.

[9] C. Cattuto, C. Schmitz, A. Baldassarri, et al. (2007). Network properties of folksonomies. *AI Comm.*, vol. 20, no. 4, pp. 245-262.

[10] R. Móro, M. Bieliková (2012). Personalized Text Summarization Based on Important Terms Identification. *23rd Int. Workshop on Database and Expert Systems Applications*, IEEE, pp. 131–135.

[11] X. Zhang, L. Yang, X. Wu, et al. (2009). sDoc: exploring social wisdom for document enhancement in web mining. *Proc. of the 18th ACM conf. on Inf. and knowledge management*, ACM, pp. 395–404.

[12] Y. Yanbe, A. Jatowt, S. Nakamura, K. Tanaka (2007). Can social bookmarking enhance search in the web? *Proc. of the 7th ACM/ IEEE-CS joint conf. on Digital libraries*, ACM, pp. 107–116.

[13] C. Biancalana, A. Micarelli (2009). Social tagging in query expansion: A new way for personalized web search. *Computational Science and Engineering*, vol. 4. IEEE, pp. 1060-1065.

[14] R. Abbasi (2011). Query expansion in folksonomies. *Semantic Multimedia*, Springer Berlin Heidelberg, pp. 1-16.

[15] G. Golovchinsky, M.N. Price, B.N. Schilit (1999). From reading to retrieval: freeform ink annotations as queries. *SIGCHI Bulletin*. ACM Press, 1999, pp. 19–25.

[16] Y. Yang, N. Bansal, W. Dakka, et al. (2009). Query by document. *Proc. of the 2nd ACM Int. Conf. on Web Search and Data Mining (WSDM '09)*, ACM, pp. 34–43.

[17] T. Strohman, W. B. Croft, D. Jensen (2007). Recommending Citations for Academic Papers. *Proc. of the 30th Annual Int. SIGIR Conf. on Research and Development in Inf. Retrieval*, ACM, pp. 5–6.

[18] M. Kompan, M. Bieliková (2010). Content-based News Recommendation. *E-Commerce and Web Technologies*, Lecture Notes in Business Information Processing, vol. 61, part 2, Springer, pp.61-72.

[19] Z. Zhang, J. Iria, C. A. Brewster, F. Ciravegna (2008). A comparative evaluation of term recognition algorithms. *Proc. of 6th Int. Conf. on Language Resources and Evaluation*, Marrakech Morocco.

[20] J. Ševcech, M. Bieliková, R. Burger, M. Barla (2012). Logging activity of researchers in digital library enhanced by annotations. *Proc. of 7th Workshop on Int. and Knowledge oriented Tech.*, pp. 197-200. (in Slovak)

[21] S. Molnár, R. Móro, M. Bieliková (2013). Trending words in digital library for term cloud-based navigation. *Proc. of the 8th Int. Workshop on Semantic and Social Media Adaptation and Personalization (SMAP '13)*, IEEE CS, to appear.

[22] T. A. Phelps, R. Wilensky (2000). Robust intra-document locations. *Computer Networks*, vol. 33, no. 1, pp. 105-118.

[23] P. Ciccarese, M. Ocana, L. J. Garcia Castro, S. Das, T. Clark (2011). An open annotation ontology for science on Web 3.0. Journal of Biomedical Semantics, vol. 2, no. 2.

[24] D. Paranyushkin (2011). Visualization of Text's Polysingularity Using Network Analysis. *Prototype Letters*, vol. 2, no. 3, pp. 256–278.

[25] G. K. Palshikar (2007). Keyword extraction from a single document using centrality measures. *Pattern Recognition and Machine Intelligence*, Springer Berlin Heidelberg, pp. 503-510.