

Query Construction for Related Document Search Based on User Annotations

Jakub Ševcech, Mária Bieliková
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Ilkovičova, 842 16 Bratislava, Slovakia
Email: {name.surname}@stuba.sk

Abstract—We often use various services for creating bookmarks, tags, highlights and other types of annotations while surfing the Web or just reading electronic documents. These annotations represent additional information on particular information source. We proposed a method for query construction to search for related documents to currently studied document. We use the document content where we concentrate on user created annotations as indicators of user's interest in particular parts of the document. Our method for query construction is based on spreading activation in a graph created from the document content. We evaluated proposed method within a service called Annota, which allows users to insert various types of annotations into web pages and PDF documents displayed in the web browser. We analyzed properties of various types of annotations inserted by users of Annota into documents. Based on these properties, we also performed a simulation to determine optimal parameters and compare proposed method against commonly used tf-idf based method.

I. INTRODUCTION

WE OFTEN use various services for creating bookmarks, tags, highlights and other types of annotations while surfing the Web or when reading electronic documents. We use these annotations as means to store our thoughts or to organize personal collections of documents using methods such as tag-cloud. Many services supporting document bookmarking and manual annotation of documents provides us the possibility to create various types of annotation by simulating the process of annotation creation in printed documents. These services do not provide us with new types of annotations in addition to annotations we have been already creating in printed documents. They rather provide us new possibilities for annotation utilization. There is active research in the field of utilization of annotation [1], for example in support of navigation between documents. In the work presented in [2] the authors use the term social document to represent document enhanced by the user generated content such as annotations. They used the user generated content similarly to anchor texts while indexing documents. This representation of documents proved to provide improved performance in content-based mining applications on the Web such as search engines, recommendation systems etc.

User created annotations can be considered a form of user's context he creates while reading documents and trav-

eling in digital space [3]. Great many applications use annotations as means for navigation between documents and for organizing content. For example, in [4] the authors describe an organization of learning materials and collaboration of students while learning using 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 deployed in a large enterprise is presented in [5]. 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 content. Tags are used for organizing bookmarks in services such as Diigo¹ or Delicious², but they are also used to organize notes³, in various blogs and many other applications. Tags and other types of annotations are means document visitors can use to create custom navigation. They can categorize or describe resources and by this way create navigation that fits their needs without relying on navigation provided by document author.

User created annotations can be used not only to support navigation, but there are many other possible applications. Tags are used for folksonomy construction [6], annotations can play a great role for example in content enrichment and content quality improvement such as in an education system presented in [4]. 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 [7].

Currently, there are many services allowing users to annotate the documents. Annotations are used to support the navigation in users' collection of documents, they allow users to create their own organization of documents via tags and they help in search for documents. All of these applications motivate users to create annotations by a prospect of future improvement in inter or intra document navigation. Users benefit created annotations only after there is enough annotated documents, or when returning to once annotated document.

¹ Diigo, <http://www.diigo.com/>

² Delicious, <https://delicious.com/>

³ Evernote, <https://www.evernote.com/>

Problem with this approach is that there is lack of immediate reward for the annotation creation.

In this paper we propose a method for query construction from currently studied document and attached annotations. 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 in time the user is reading the documents and it is used to search for further documents related to the currently studied document. The reward for user creating annotations is thus provided in time of annotation creation.

II. RELATED WORK

One of possible employment of annotations in information processing is the document search. There are two possible approaches for exploitation of annotations in search. One is to use annotations while indexing documents by expanding documents in a similar way anchor texts are used [2] or by ranking document quality using bookmarks and annotations as document quality indicators [8].

The second possible application of annotations in document search is in query expansion or query construction process. An example of annotations used for query expansion is presented in [9], where tags attached to search results are used to expand initial query similarly to pseudo-relevance feedback based query expansion. Multiple methods for query expansion in folksonomies are presented in [10]. 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 [11]. The authors asked users to read a set of documents and to create annotations into documents using a tablet. They used these annotations as queries in related document search. They used different weights for different types of annotations in query construction and 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. Whereas query expansion requires that users create an initial query, query composition using annotations does not require additional activity of readers instead it reuses annotations created with other purposes such as better understanding of the document.

In experiment presented in [11] authors let users to create annotations into documents for evaluation of their applicability in related document search. More often in search for related documents the content of source document is used to create queries. In [12] authors used the most important phrases from the source document as queries for document retrieval. Extracting most important phrases is similar to document summarization. However they used extracted phrases as queries in related documents retrieval.

Another work concerning search for related documents is described in [13]. The authors use related document search as a mean for 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 its application in document recommendation. In work presented in [14] they used list of documents similar to users visited documents to recommend related documents. To compute document similarity they used document representation based on word vector extracted from its content and similarity metric based on cosine similarity.

Searching for related documents can be useful also in the domain of plagiarism detection. In [15] query construction from the source document is used for retrieval of documents the suspicious document may be plagiarized from. In the query construction process the most frequent words from the document are used.

Document term frequency for query construction from document content is used also in popular content-based search engines ElasticSearch⁴ and Apache Solr⁵. They provide special type of query interface called "more like this" query, which processes source text and returns list of similar documents. Internally, the search engine extracts the most important words using tf-idf metric from source text and it uses the most important words as a query for related documents search. By comparison to previous described method, tf-idf based method uses additionally to in-document term frequency also information about terms from the collection related documents are searched in.

In multiple works authors showed that annotations represent important source of information for document retrieval. Methods for query construction for document retrieval are however using only document content and information about document collection in query construction process. They are not using user created annotations as user's interest indicators when creating query for document retrieval. In our work we proposed and evaluated 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 user is mostly interested in. Using user created annotations our method creates a keyword query for related document search taking into account the user interests. Annotations are used in time of their creation and they provide immediate motivation in form of related document search.

III. METHOD FOR QUERY CONSTRUCTION

Currently the most common form of query used when searching for documents on the Web is the list of keywords. To retrieve words from the document to be used as query for related document search it is possible to use multiple different approaches. It is possible to extract most frequent terms, use tf-idf metric or various ATR algorithms [16] to extract keywords and so on. 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.

⁴ ElasticSearch, <http://www.elasticsearch.org/>

⁵ Apache Solr, <http://lucene.apache.org/solr/>

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 in the document collection. We believe that not only the number of word occurrences but also the structure of the source text is important in a search query construction for related document retrieval. Especially, if we suppose that while reading the document users are most commonly interested in only a portion of the document, the portion where they attach annotations.

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 is composed 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 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 [17]. Using various graph algorithms such as community detection, various node and edge weightings or spreading activation we can extract properties such as related words, most important terms, topics etc. We use this graph to extract words that can form queries to retrieve similar documents using spreading activation algorithm.

Text to graph transformation

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

```
words=text.split.remove stopwords.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 [17], where they used two passages through the text with maximal distance set to two words and

five words. By using these setting, the words with greater distance were connected and close words have more common edges at the same time.

All created edges have the same weight but by using two passages through the text, more edges are created between close words than between farther words. For the purpose of speeding up the spreading activation in the next step, we connected multiple edges between the same nodes and we set weight of the resulting edge as number of connected edges.

Query word extraction

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

It is possible to use this algorithm for related nodes search but also for other application such as keyword extraction [18]. We use this algorithm to find the most important words in the graph created from the text. The initial activation is introduced 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. When user created annotations are used to insert initial activation, user's interest are reflected in the most important words extracted after spreading activation.

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

- 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. Annotations 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. This assumption may be violated in the case where the document and the associated comments are in different languages. Therefore, in performed experiments we translated the content of every annotation using Google Translate service.

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 into the query. In our case the activation is inserted into the graph through annotations attached to document by its reader. As we use annotations as user's interest indicators, the activation is spreading from document parts, user is most

interested in and words with highest activation level are reflecting user's interests.

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 number of iteration 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.

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 whole document's text.

IV. CREATION OF THE WEB PAGE ANNOTATION

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 content with regard to documents which may change over time. In [19] multiple criteria, which must meet the robust method for locating annotations into documents, are defined. Some of the criteria are:

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

At the same time in this work they 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 [20]. It is tolerant to changes in the document content and when using approximate matching of strings it is also to some extent tolerant to changes in annotation context as well.

We developed a service called Annota⁶ [21], which allows users to attach annotations to arbitrary web pages or PDF documents displayed in a web browser. Annota service is realized as a browser extension through which user can create various types of annotations such as:

- tags,
- highlights,

- comments attached to text selections and
- notes attached to the document as a whole.

The service is focused on supporting visitors of digital libraries as we collect metadata on articles from selected digital libraries (ACM DL, SpringerLink, IEEE Xplore). We realized the possibility to insert annotations into arbitrary web pages and articles in digital libraries, by bookmarking and sharing documents and annotations in user groups.

The Annota service allows users to organize documents by tags or folders. It is possible to search in document's texts in user's library or library of bookmarked documents of all users. An example of web page annotated using Annota is displayed at Figure 1. The figure shows a sidebar, where it is possible to bookmark displayed page, insert tags, edit note and share bookmark with groups 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 user studying a document. The user has following possibilities of particular activities:

- Bookmarking documents.
- Highlighting parts of the text and creating other types of annotations.
- Sharing bookmarked document via group sharing.
- Collaborative annotation of documents.

The browser extension allows users to create annotations that link to document as a whole (tags, note) or to particular parts of the document (highlight). 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 robust to changes in annotated document.

To attach annotations to document parts we use redundant representation of annotation location to support linking annotations into changing documents. To locate 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 content except changes in selected text and most changes before annotation location. With usage of approximate matching this method is to some extent tolerant to changes in selected text as well.

We analyzed behavior of users of Annota while annotating documents using developed 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 time period. They used Annota on day-to-day basis to bookmark interesting documents, to summarize them, to write down their thoughts about the document content and to highlight important parts of the document. We studied multiple parameters of created annotations and notes 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. All observed parameters were following logarithmic or geo-

⁶ Annota, <http://annota.fiit.stuba.sk/>

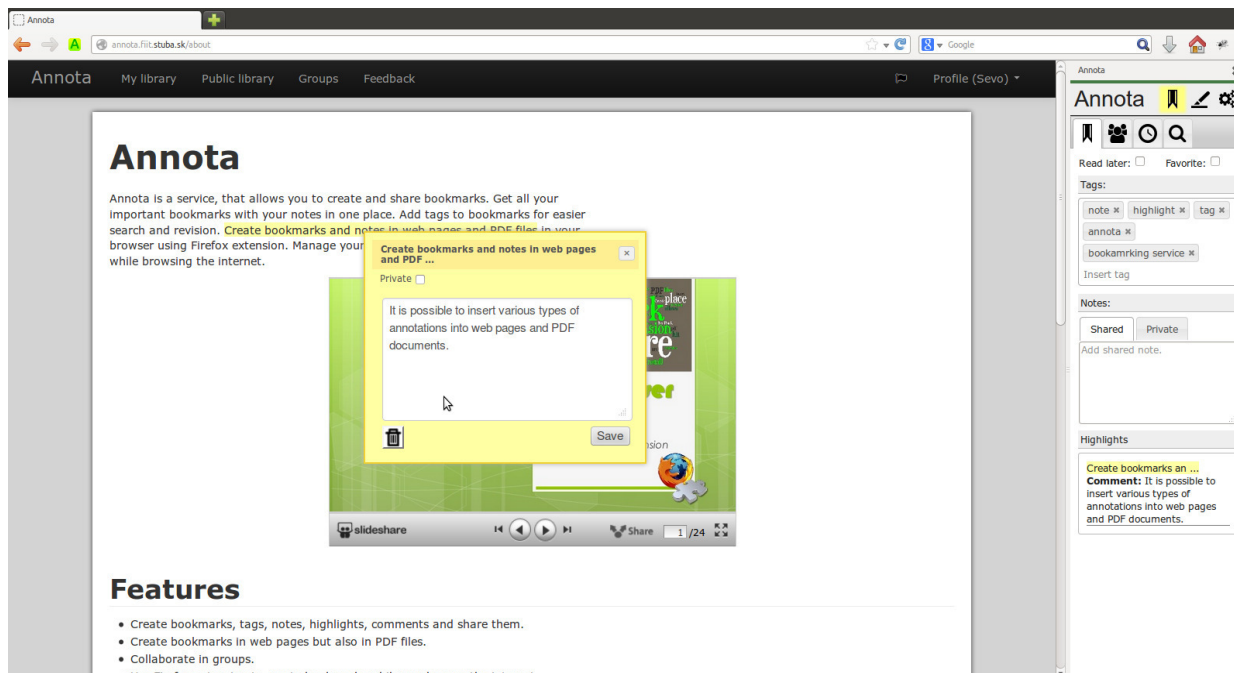


Figure 1 Web page annotated using bookmarking service Annota

metric distributions. Figure 2 represents an example of derived distribution for number of highlighted texts per document that follows logarithmic distribution.

V. EVALUATION

Using various attributes of annotations and their probabilistic distributions described in previous section, 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.

The simulation was performed on the dataset we created by extracting documents from Wikipedia. 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. 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 describing each of these meanings.

By using abstracts of pages describing different meanings of synonyms 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 |

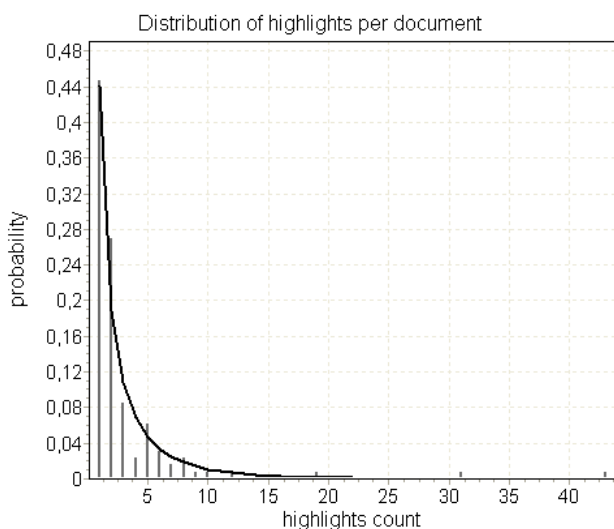


Figure 2 Logarithmic distribution of highlighted texts number per document

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 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 composing abstracts which simulated one topic user is most interested in. Into the selected abstract we generated various types of annotations, both annotations highlighting parts of the document and annotations inserting additional content. Annotations highlighting parts of the document were randomly distributed over the whole abstract. To simulate content of annotations extending content of the document (note, 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 based on text to graph transformation and spreading activation. Created query was used for related documents search in the index of all downloaded documents. When evaluating precision of search for related documents, we considered document to be relevant if it was from the same category as the page of annotated abstract.

We performed a simulation with several combinations of parameters and we implemented hill climbing algorithm to optimize parameter combination for the highest 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.generate(abstract)
    graph.activate(annot, weights)
    graph.spread_activation
    query = graph.top_nodes
    results = Elasticsearch(query)
    cat = abstract.page.categories
    relevant = results.with_category(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 proposed method with method based on tf-idf when using annotations in the query construction process, we extended rather straightforwardly the tf-idf based method to use annotations in query word extraction process. This method uses document word frequency to find 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.

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. These experiments aimed to determine precision of compared methods when no annotations are available and when we have complete information about user's interest.

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 comparison of both methods without using annotations and using generated annotations in query construction process proved that annotations are increasing 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 for query word extraction.

We performed a Student's t-test on 5% level of significance for pairs of proposed method and tf-idf based method for every performed experiment to determine if we obtained statistically significant differences in mean precision for compared methods. As the computed p-value was less than 0.01% for every performed experiment, we can reject null hypothesis that the mean precisions of compared methods are equal. We obtained significant differences in mean precision for proposed method and tf-idf method for experiments using annotations in query construction process as well as for experiment with whole document fragments annotated.

To compare a real increase of precision of related document retrieval using annotations in query construction process and without using annotations, we performed a qualitative user study, where in sequence 8 volunteers were asked to annotate documents of their choice stored in the Annota service. After they annotated these documents, we generated two queries using proposed method, one using annotations and one without using annotations in query construction process. We retrieved two lists of documents using

these queries and we presented them to volunteers in random order. Volunteers were then asked to select documents describing topic related to the topic of source document from displayed lists and to select better from two presented lists.

The volunteers annotated 11 unique documents. In 9 cases they selected for more relevant the list created by method using annotations. In one case method using annotations created query in Slovak and document search returned no 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. By method taking into account annotations we obtained 34 relevant documents in total and with method not using annotations only 15.

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 experiment, the volunteer doubled the number of attached annotations. In the second retrieved list of documents, the number of relevant documents retrieved increased and included one exact match with the topic user was most interested in while annotating source document. With increasing number of annotations attached to document the precision of related document retrieval is increasing.

When using annotations to create a query, the proposed method obtained better results than in the case when annotations were not used in the query construction process. When using annotations, created query retrieves more documents that describe the same topic as the source documents and more documents that describe related topics.

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 for creating annotations such as notes, comments and highlights is to summarize studied documents, describe documents, highlight most important sections, to 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. The distribution of created in-text annotation was uniform over the whole text. In this study interviewed volunteers confirmed our assumption that using annotations users

are indicating those parts of the document they are most interested in.

VI. CONCLUSION AND FUTURE WORK

Annotations represent important source of information on interesting or important parts of documents. Its importance increases by possibilities of manipulating documents on the Web by the way we want to do commonly with paper documents. We studied user behavior while annotating documents on the Web and proposed a method for query construction from document content and attached annotations. In the process of query construction we considered document content and its structure by using text to graph transformation and query terms extraction using spreading activation in created graph. We used user created annotations as user's interest indicators to insert initial activation into graph created from document content.

We have developed a bookmarking service called Annota and a browser extension allowing users to insert various types of annotations into web pages and PDF documents displayed in web browser. The simulation based on probabilistic distributions of various parameters of annotations created by users of Annota proved, that annotations used 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 query construction. It is thus possible to use it even when no annotations are attached into the document with comparable precision as commonly used method when extracting words fit into query for related document retrieval from document content. The spreading activation based method outperformed compared method when document attached annotations were used in query construction process. The proposed method does not use information from other documents, only information from source document content and attached annotations. It is thus search engine independent and can be used to create queries for any search engine accepting queries in form of a list of keywords.

Performed user study showed that users insert annotations into document sections they are most interested in and they are use annotations to summarize documents, highlight most important parts of documents and to store their thoughts.

We evaluated proposed method for increasing related document retrieval precision of created query when using user created annotations in query construction process.

We plan to use annotations created not only by single user but also annotations created by other users when creating query for related document search. We see the potential in use of social relations such as group membership in weighting of annotations created by other users in query construction process.

In the described work, we were using annotations attached into document along with document content and we have not used user's annotations attached to other documents. By using annotations from other documents, we plan to model user's interests. Such annotation based user model can be used for further improvement of query construction process.

Moreover there are several possible enhancements related to document search process from the point of view of search engine. We plan to use annotations to enrich document content while creating index of annotated documents and we will compare performance of search in annotation enriched index against related document search in index created only from document content.

ACKNOWLEDGEMENT

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

REFERENCES

- [1] M. Agosti, N. Ferro, "A formal model of annotations of digital content." *ACM Trans. Inf. Syst.*, Nov. 2007, vol. 26, no. 1.
- [2] X. Zhang, L. Yang, X. Wu, et al., "sDoc: exploring social wisdom for document enhancement in web mining." In *Proc. of the 18th ACM conf. on Inf. and knowledge management*, ACM, 2009, pp. 395-404.
- [3] P. Návrat, "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.
- [4] M. Šimko, M. Barla, V. Mihál, M. Unčík, M. Bieliková, "Supporting Collaborative Web-based Education via Annotations." In *World Conf. on Educational Multimedia*, AACE, 2011, pp.2576-2585.
- [5] D. Millen, M. Yang, S. Whittaker, J. Feinberg, "Social bookmarking and exploratory search." *ECSCW 2007*, Springer, London, 2007, pp. 21-40.
- [6] C. Cattuto, C. Schmitz, A. Baldassarri, et al. "Network properties of folksonomies." *AI Comm.*, 2007, vol. 20, no. 4, pp. 245-262.
- [7] R. Moro, M. Bieliková, "Personalized Text Summarization Based on Important Terms Identification." *23rd Int. Workshop on Database and Expert Systems Applications*, IEEE, 2012, pp. 131-135.
- [8] Y. Yanbe, A. Jatowt, S. Nakamura, K. Tanaka, "Can social bookmarking enhance search in the web?" In *Proc. of the 7th ACM/IEEE-CS joint conf. on Digital libraries*, ACM, 2007, pp. 107-116.
- [9] C. Biancalana, A. Micarelli, "Social tagging in query expansion: A new way for personalized web search." *Computational Science and Engineering*, 2009. vol. 4. IEEE, pp. 1060-1065.
- [10] R. Abbasi, "Query expansion in folksonomies." In *Semantic Multimedia*, Springer Berlin Heidelberg, 2011, pp. 1-16.
- [11] G. Golovchinsky, M. N. Price, B. N. Schilit, "From reading to retrieval: freeform ink annotations as queries." *SIGCHI Bulletin*. ACM Press, 1999, pp. 19-25.
- [12] Y. Yang, N. Bansal, W. Dukka, et al., "Query by document." In *Proc. of the Second ACM International Conf. on Web Search and Data Mining*, ACM, 2009, pp. 34-43.
- [13] T. Strohman, W. B. Croft, D. Jensen, "Recommending Citations for Academic Papers." In *Proc. of the 30th annual int. SIGIR conf. on Research and development in inf. retrieval*, ACM, 2007, pp. 5-6.
- [14] M. Kompan, M. Bieliková, "Content-based News Recommendation." In *E-Commerce and Web Technologies, Lecture Notes in Business Information Processing*, vol. 61, part 2, Springer, pp.61-72.
- [15] A. R. Pereira, N. Ziviani, "Retrieving similar documents from the web." *Journal of Web Engineering*, vol. 2, no. 4, 2003, pp. 247-261.
- [16] Z. Zhang, J. Iria, C. A. Brewster, F. Ciravegna, "A comparative evaluation of term recognition algorithms." In *Proc. of 6th Int. Conf. on Language Resources and Evaluation*, Marrakech Morocco, 2008.
- [17] D. Paranyushkin, "Visualization of Text's Polysingularity Using Network Analysis." *Prototype Letters*, 2011, vol. 2, no. 3, pp. 256-278.
- [18] G. K. Palshikar, "Keyword extraction from a single document using centrality measures." *Pattern Recognition and Machine Intelligence*, Springer Berlin Heidelberg, 2007, pp. 503-510.
- [19] T. A. Phelps, R. Wilensky, "Robust intra-document locations." *Computer Networks*, 2000, vol. 33, no. 1, pp. 105-118.
- [20] P. Ciccarese, M. Ocana, L. J. Garcia Castro, S. Das, T. Clark, "An open annotation ontology for science on Web 3.0." *J Biomed Semantics*, 2011, vol. 2, no. 2.
- [21] J. Ševcech, M. Bieliková, R. Burger, M. Barla, "Logging activity of researchers in digital library enhanced by annotations." In *Proc. of 7th Workshop on Int. and Knowledge oriented Tech.*, 2012, pp. 197-200. (in Slovak)