ANNOTATIONS AS USER INTEREST INDICATORS FOR RELATED DOCUMENT RETRIEVAL

Jakub ŠEVCECH, Mária BIELIKOVÁ

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
{sevcech,bielik}@fiit.stuba.sk

Abstract. When we read printed documents, we often insert various types of notes, marks and highlights into the document. Multiple services allows us to create similar annotations into electronic documents such as web pages or PDF documents as well. The main reason people insert annotations into the document is to highlight important section, to write down some thought or summary of the document. We proposed a method to use these annotations as indicators of user's interest in related document retrieval. The proposed method creates query for document search, in form of list of keywords, to retrieve related document to currently studied document. The method uses text to graph transformation and most important word extraction using spreading activation algorithm for query terms selection. We evaluated proposed method by comparing its related document retrieval precision to other commonly used tf-idf based method. We showed that proposed method improves related document retrieval precision and that annotations can be used as user interest indicators to find document's most important fragments.

1. Introduction

Many software tools such as various bookmarking services or document browsers allows us to insert various types of annotations into electronic documents in a similar way we are accustomed to write into printed documents. These annotations take on form of highlights, comments, freeform notes, various arrows, circles and tags. These are valuable source of information used in supporting document organization, intra and inter document navigation [3] or document summarization [4].

Another possible employment of annotations in information processing is the document search. There are several possible approaches for exploitation of annotations in

search. One is to use annotations while indexing document in a similar way anchor texts are used [3]. The second application is in ranking document quality using bookmarks and annotations as document quality indicators [5]. The other 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 [6], where tags attached to search results are used to expand initial query similarly to pseudo-relevance feedback based query expansion.

In this work, we used user created annotations as indicators of user's interest in selected documents or particular fragments of the document. We used annotations to find the most important parts of the document from user's point of view and to create query from the document content and attached annotations to search for related documents.

2. Method for related document retrieval

We proposed a method [1, 2] for query construction for related document retrieval using content of currently studied document and annotations, user attached to the document. The method uses document text to graph transformation and most important word extraction, using spreading activation, to extract words used as a query for related document retrieval. To insert initial activation to the graph of words created from source document, we use annotations attached to specific parts of the document as well as annotations attached to the document as a whole.

The proposed method was implemented as a component of bookmarking service Annota[1] (Figure 1), which allows users to insert various types of annotations into web pages or PDF documents displayed in the web browser. We determined parameters of proposed method and evaluated the method using simulation populated by data from annotations created by users of bookmarking service Annota.

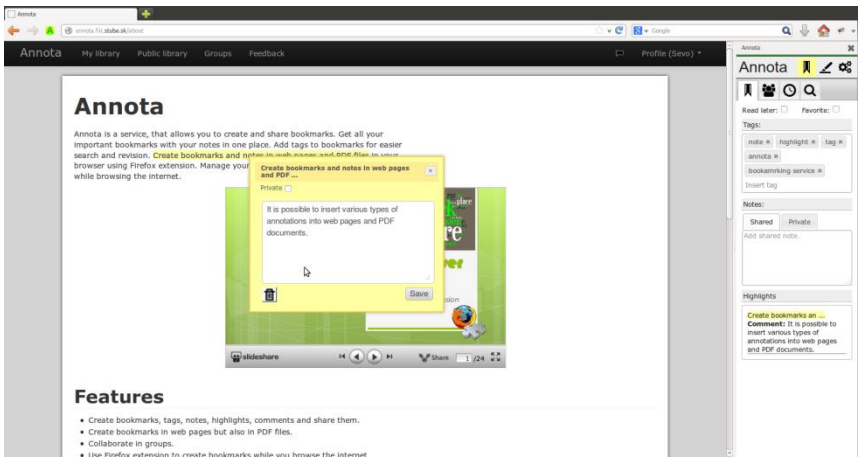

Figure 1. Web page annotated using bookmarking service Annota.

[1] Annota - http://annota.fiit.stuba.sk

We used simulation to optimize parameters of proposed method for related document retrieval as well as to compare this method to other commonly used method for related document retrieval. We compared proposed method with tf-idf based method used in ElasticSearch[2] search engine. We extended compared method to take into account annotations attached to documents in query construction process by attaching annotation content to analyzed document content. We performed simulation to optimize weights of attached annotations for the tf-idf based method. We performed three series of experiments to compare precision of method based on text-to-graph transformation and method based on tf-idf without using annotations and when using various quantities of annotations in query construction process. Results of performed experiments are summarized in table Table 1.

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%

Table 1. Simulation results for proposed method and tf-idf based method

In performed experiments we proved that proposed method outperforms commonly used method for related document retrieval based on tf-idf when using annotations as well as without using annotation in query construction process. As results in table Table 1 indicate the proposed method outperforms compared tf-idf based method in related document retrieval 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 tf-idf based method for every performed experiment to determine if we obtained statistically significant differences in mean precision for compared methods. 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. However we did not achieved significant difference in retrieval precision for compared methods when no annotations were used in query construction process.

The proposed method outperforms commonly used method in related document retrieval precision of queries extracted from document content. By comparing retrieval precision of queries constructed using annotations with queries constructed without annotations in query construction process, we showed that annotations can be used as interest indicators in important document fragments extraction and in query construction process. In connection with achieved results and other cited works we can claim, that annotations can be successfully used as sources of additional information in various tasks in the field of information retrieval.

[2] ElasticSearch, http://www.elasticsearch.org/

Acknowledgement: This work was partially supported by the Scientific Grant Agency of the Ministry of Education of Slovak Republic, grants VG1/0675/11, VG1/0971/11 and by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

to other papers publishing the results that are summarized here

[1] Ševcech, J., Bieliková, M.: Query Construction for Related Document Search Based on User Annotations. In: *Proc. of the Federated Conference on Computer Science and Information Systems 2013*, FedCSIS, (2013), pp. 279–286.

[2] Ševcech, J., Móro, R., Holub, M., Bieliková, M.: User annotations as a context for related document search on the web and digital libraries. In: *Informatica*, (2014), vol. 38, no. 1, pp. 21-30.

Other references

[3] Zhang, X., Yang, L., Wu, X., 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.

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

[5] Yanbe, Y., Jatowt, A., Nakamura, S., Tanaka, K.: 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.

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