Web Navigation Based on Annotations

Jakub ŠEVCECH*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
sevo.jakub@gmail.com

We often use various services for creating bookmarks, tags, highlights and other types of annotations while surfing the Internet or when reading electronic documents. We use these annotations to highlight important parts of documents and to mark our thoughts in the margin of the document. User created annotations are commonly used to support navigation, for text summarization [2], search [1] etc. We proposed a method for searching related documents to currently studied document using annotations created by the document reader. We perceive annotations as indicators of user's interest in particular parts of the document.

The proposed method uses text to graph transformation to preserve document content and its structure and spreading activation algorithm to identify most important words in the text of studied document. The text to graph transformation step transforms words of document to graph nodes and creates edges using word neighbourhood in the source document. We use annotations highlighting parts of the document as well as annotations attaching additional information to insert initial activation into the document graph. Annotations highlighting parts of the document insert activation to nodes representing highlighted words and annotations inserting additional content extend document graph by new nodes and edges and introduce activation to these nodes. The initial activation is spreading from nodes with attached annotations and concentrates in most important words. The proposed method extracts words, which are important for annotated parts of the document, but it also extracts globally important words that are important for the document as a whole. The portion of locally and globally important words can be controlled by number of iteration of the algorithm. We use these words as query in retrieval of related documents. We determined the right number of iterations and the right amount of activation introduced by various types of annotations into the graph using simulation based on user created annotations.

To evaluate proposed method we created a service called Annota (annota.fiit.stuba.sk), which allows users to insert various types of annotations into web pages and PDF documents displayed in the web browser. We analysed properties of various types of annotations inserted by users of Annota into documents and we

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2013PeWe Workshop, April 5, 2013, pp. 39–40.

derived probabilistic distributions of annotation attributes such as note length, number of highlights per user and per document or probability of comment to be attached to text selection. Figure 1presents an example of derived distribution of number of highlighted texts per document that follows logarithmic distribution.

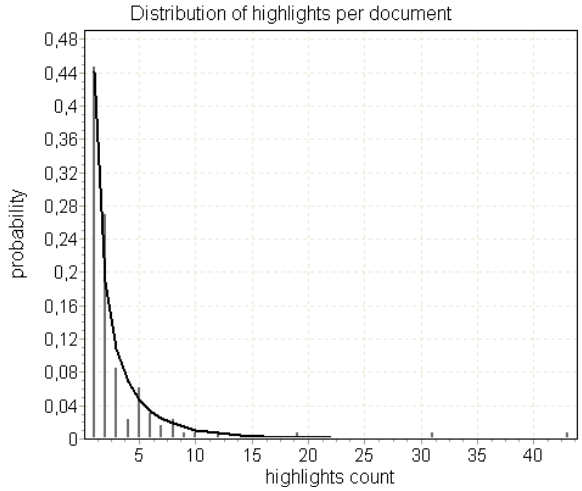

Figure 1.Logarithmic distribution of highlighted texts number per document.

Using these distributions we created a simulation, to find optimal weights for various types of annotations and number of iterations of proposed method, where we optimized query construction for document search precision. We used dataset created from Wikipedia pages to create index of documents. We created query from source document and annotations generated using parameter distributions. We compared relevancy of documents retrieved when searching within created index using this query and when searching using query created by TF-IDF based method.

The proposed method outperforms compared method when generated annotations were used as user's interest indicators and it received better results when no annotations were used and when whole document sections were annotated. Proposed method outperformed compared method even though it is not using information about other documents from collection unlike the TF-IDF based method.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 143-148.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

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

[2] Moro, R., Bieliková, M.: Personalized Text Summarization Based on Important Terms Identification. 23rdInt. Workshop on Database and Expert Systems Applications, IEEE, (2012), pp. 131–135.