Determining the Relevancy of Important Words in a Digital Library using the Citation Sentences

Máté VANGEL*

Slovak University of Technology in Bratislava

Faculty of Informatics and Information Technologies

Ilkovičova 2, 842 16 Bratislava, Slovakia

mate.vangel@gmail.com

The number of articles and publications in digital libraries is enormous and is persistently increasing, therefore keeping everything organized in the world of digital libraries has become impossible without automation of some of the processes. The main concept is realized by methods of text mining, which is used to extract and link relevant information (objects) from texts. Text mining can be used to fulfil various kinds of tasks like domain modelling, automatic text summarization or navigation in a cloud of keywords.

Keyword extraction is usually done using the text of the specified document, but in digital libraries there are some other possible options for extracting relevant words. One of these possibilities is to use the available information related to the article, which is called metadata.

There are many kinds of metadata in digital libraries, for instance keywords provided by the authors, tags, year of publishing, category in which the article is located and tags associated by users. Citations can also be considered as an important source of keywords, because they can characterize the article. They can also highlight different, but relevant aspects of the analysed article, which can be relevant for other researchers.

There are various possible solutions for keyword extraction, however, they are mainly using one source of available information. Usually this is the main text of an article. We aim to extract keywords with the help of document metadata in our work, and we consider citations as the main source (input) for keyword extraction.

Citations can be used for keyword extraction in many ways. Using citations in keyword extraction can give distinct results compared to extraction from the abstract of the article [2], which is also a very popular and efficient input for extracting and evaluating keywords.

*Supervisor: Róbert Móro, Institute of Informatics and Software Engineering

Spring 2014 PeWe Workshop, March 21, 2014, pp. 43–44.

We plan to analyse citations from different aspects. One of the most important aspect is the citation context, which means the environment of the cited text [1]. Citation context can be important from two distinct views:

- Where is the cited text located in the article
- What is surrounding the cited text

By using the first view we plan to analyse the structure of the article, because articles in digital libraries share the same structure, i.e. at the beginning of the article there is always the title, then name of the author(s), publication year, abstract, text of the article and the conclusion at the end. Information located in the abstract and the conclusion usually reflect the main facts of the article, so citations from these parts are often more relevant than from other parts.

By using the second view we want to discover the surroundings of the cited text. In its surroundings we will search for relevant words and also for other citations. If we will find a keyword or other citations in its vicinity, then we will consider the analysed citation more relevant. An important factor will be the distance of the found keywords and other citations from the original [3].

Other aspects which can be analysed are the authors of citations, the article in which is the analysed article cited and the popularity of the citation (how many times was the given text cited).

In order to identify the most relevant keywords, we can use the spreading activation method on a citation network. In digital libraries most of the article references are directly connected to the referenced document, so we can recursively find and analyse all referenced articles and apply our method for keyword extraction from citations.

We plan to evaluate our method of keyword extraction in a web-based bookmarkig system Annota[1] on the dataset of articles from ACM Digital Library[2].

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

References

[1] Qazvinian, V., Radev, D.R.: Identifying non-explicit citing sentences for citation-based summarization. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics - ACL '10*, (2010), pp. 555–564.

[2] Liu, S., Chen, C.: The differences between latent topics in abstracts and citation contexts of citing papers. *Journal of the American Society for Information Science and Technology*, vol. 64, no. 3, (2013) pp.627–639.

[3] Elkiss, A. et al.: Blind men and elephants : What do citation summaries tell us about a research article. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 1, (2008), pp. 51–62.

[1]http://annota.fiit.stuba.sk
[2]http://dl.acm.org