Citations and Co-citations as a Source of Keyword Extraction in Digital Libraries

Máté VANGEL*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
mate.vangel@gmail.com

In digital libraries there are several types of information, which can be suitable as an input source for extracting relevant words. In our work we propose a method for extracting keywords and determining their relevancy for research articles in digital libraries using citations and co-citations.

Citations highlight different aspects of research articles [1], which authors of citations consider as relevant. Based on this behavior citations can have similar character as abstracts, but they include more specific information. The positive effect of citation analysis on the relevancy of extracted keywords has been already evaluated in [2]. Authors tried several keyword extraction methods on the text of research articles combined with citations' contexts. The proper size of the citations contexts is also evaluated in [3]. While setting the size of the citation contexts to at least one sentence can be beneficial already, using 100 words before and 100 words after the reference string is the best choice. We use this exact value in our method while we parse citation contexts.

Our proposed method analyses three separate sources of text during the keyword extraction for each research article. These sources are: text of the research article, citations' contexts and co-citations.

Our method of keyword extraction is innovative due to the usage of co-citations. The third source of text is constructed from the text of co-cited articles with weight of co-citation at least two, which means that articles are co-cited at least two times. After these sources are prepared, we compute the relevancy of each word in each source using statistical method *tf-idf*. This step is followed by weight normalization in each source and the computation of the final relevancy of each word. For this calculation we use a simple linear combination.

The evaluation of the proposed method will take place in a domain of digital libraries in a web-based bookmarking system *Annota*[1], where we will calculate the

* Supervisor: Róbert Móro, Institute of Informatics and Software Engineering
[1] http://annota.fiit.stuba.sk

Spring 2015 PeWe Workshop, April 11, 2015, pp. 83–84.

precision of the proposed method based on the users' explicit feedback. For the draft of the evaluative graphical user interface see Figure 1.

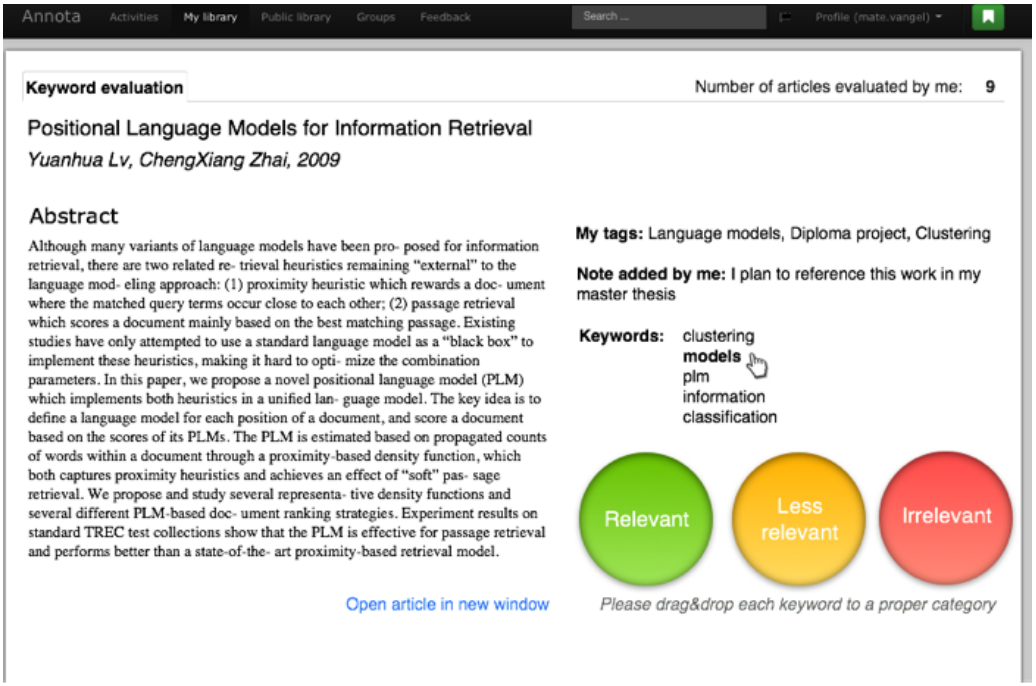

Figure 1. Draft of the evaluative graphical user interface.

Amended version was published in Proc. of the 11th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2015), STU Bratislava, 50-51.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] Elkiss, A., Shen, S., Fader, A., States, D., Radev, D.: Blind Men and Elephants: What Do Citation Summaries Tell Us About a Research Article? *Journal of the American Society for Information Science and Technology*, 2008, vol. 59, no. 2003, pp. 51–62.

[2] Qazvinian, V., Radev, D., Özgür, A.: Citation summarization through keyphrase extraction. In: *COLING '10: Proc. of the 23rd International Conference on Computational Linguistics*, (2010), pp. 895–903.

[3] Ritchie, A., Robertson, S., Teufel, S.: Comparing citation contexts for information retrieval. In: *CIKM '08: Proc. of the 17th ACM Conference on Information and Knowledge Management*, ACM Press, (2008), pp. 213-222.