Extracting Keywords from Educational Content

Jozef HARINEK*

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

When considering social educational systems, educational content has one advantage compared to other types of documents. With the emergence of e-learning 2.0 [2], it often has user annotations connected with it. These annotations are created by users who want to help themselves when interacting with the document. By processing these user created annotations assigned to the documents we can improve results of base Automatic Term Recognition (ATR) algorithms.

Base ATR algorithms give us results which are not as good as they could be. It means that there is still possibility to improve these results. The user created annotations provide us with useful and personalized semantic information about the documents.

We propose a method for relevant domain terms (RDT) extraction based on user generated annotations processing. We consider three basic annotation types (tag, comment, and highlight). We compute the final term weight by combining relevant domain terms weights obtained from the individual annotation types and those obtained from the text.

Our method consists of the following steps:

1. Document and annotations pre-processing
2. RDT extraction from text and annotations
3. Combining the results from both sources

The final weight of the term is computed according to the following formula:

$$w_{final}(t,d) = (1-p) * w_{ATR}(t,d) + p * w_{annot}(t,d) \qquad (1)$$

where the final weight is computed as a combination of term weight computed only from the text of the document and the term weight acquired from annotations for that particular term.

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 99-100.

Document pre-processing consists of extracting plain text from the document, lemmatization and stop-words removal. The annotations pre-processing is similar, but before lemmatization and stop-words removal we have to prepare "extended document of annotations".

This extended document of annotations consists of all the annotations connected with the particular document and we also take into account user proficiency level. The user ranking method which was used for the first experiments is a simple one which we are planning to substitute in future work. It divides users into four groups, according to the number of annotations they added. It is based on assumption that the more user interacts with the documents the higher his level of knowledge is. Based on the group which the user falls into we add his or her annotations one to four times to the extended document of annotations.

We have experimented with our proposed method. Our dataset consisted of 180 learning objects (documents), about 170 annotations per document and 1000 users. The dataset was from learning system ALEF [4] from Principles of Software Engineering course.

Our experimental results show that the annotations help in RDT extraction results improvement. The first results showed that the most promising annotation types are tags (19.8 % improvement) and highlights (12.5 % improvement). Our final method yields improvement of 22.6 % in RDT extraction.

In further experiments we also want to take into account content of the comments. Our plans are to filter out irrelevant parts of the document by finding comments with such content. Next plan is to substitute the user evaluation method with a better one, based on HITS [3] or PageRank [1] algorithm.

Extended version was published in Proc. of the 9th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2013), STU Bratislava, 7-12.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine*. *Computer Networks*. doi:10.1016/j.comnet.2012.10.007

[2] Downes, S. (2005*). E-learning 2.0. eLearn magazine*. Issue 10. ACM, p. 1.

[3] J. Kleinberg. *Authoritative sources in a hyperlinked environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.

[4] Šimko,M., Barla, M., Bieliková, M. *ALEF" A Framework for Adaptive Web Web-based Learning 2.0. In Reynolds*, N., Turcsányi-Szabó, M. (Eds.): KCKS 2010, IFIP Advances in Information and Communication Technology, Volume 324. Springer, pp. 367–378.