Web Content Annotation in Educational Web-Based Application

Jakub ŠEVCECH*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
sevo_jakub@yahoo.fr

Today we are facing the problem of information overload. It is thus important to provide information to web page readers in such a way, that it can be used most effectively. Annotations attached to documents are frequently used as a way for organising and providing additional information.

In our work we propose a method for automatic extending the information content of web pages by adding annotations to keywords in the text of the pages. The method is designed to be able to insert annotations into the text written in Slovak. The process of creating annotations for a web page consists of several steps:

1. elimination of redundant parts and selection of text to be annotated,
2. extraction of candidate words for assignment of annotations,
3. search for information to the annotations, and
4. visualization and personalization of the annotations.

Before creating the annotations, it is necessary to analyze the document and find the words to which it is appropriate to assign the annotations. As a first step it is necessary to remove redundant parts of the web page as various navigation elements, advertisement banners, etc.

The second step of the analysis of document is the search for candidate words for assignment of annotations. For this step, it is possible to use various services for keyword extraction [1]. Since these services provide best results when English text is processed, method for creating annotations uses text of the web page translated into English. To assign annotations to words in the original text, we propose a method for mapping of equivalent words between the Slovak text and its translation into English.

A method for mapping equivalent words between text translations requires an extensive bilingual dictionary. The size of the dictionary required for this task may be prohibitively large; hence we use a different approach. We employ a much smaller dictionary and match different word forms using Levenshtein distance. We calculate

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

Spring 2011 PeWe Workshop, April 1, 2011, pp. 65–66.

the Levenshtein distance of the words in the dictionary and the words in the text. If this distance is lower than a selected threshold, we consider these words to be equivalent to each other.

Information for the annotations is obtained through publicly available services for information retrieval. Using these services, the proposed method does not depend on any particular domain. Annotations created using different services can take different forms, depending on the services used to fill them. We employ services providing definitions of keywords and links to web pages concerning these keywords, but other services that provide multimedia information can be used as well.

Annotation personalization is performed on the basis of implicit feedback, gathered from user clicks on links in the content of the annotations. Whereas these links are presented in a list, user is affected by its order when deciding which link to follow. Consequently, we do not assign same weights to clicks on different positions in the list, but we treat them as indications that the clicked link is better than the other [3]. We use different strategies to partially order the links using the knowledge which links were clicked and which not. We then build a graph from this partial ordering and we use adapted PageRank algorithm to find the ratings of links for annotation purposes. The links in an annotation are displayed sorted by their rating.

In the future work we will evaluate these methods on education system ALEF [2], on the course of software engineering.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028-025STU-4/2010.

References

[1] Barla, M., Bieliková, M.: Ordinary Web Pages as a Source for Metadata Acquisition for Open Corpus User Modeling. In *Proc. of IADIS WWW/Internet 2010*. IADIS Press, pp. 227–233, 2010.

[2] Šimko, M., Barla, M., Bieliková, M.: ALEF: A Framework for Adaptive Web-based Learning 2.0. In *Proc. of IFIP Advances in Information and Communication Technology*, *Key Competencies in the Knowledge Society*, Springer, Vol. 324, pp. 367–378, 2010.

[3] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In *Proc. of the 28th annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, USA, ACM, pp. 154–161, 2005.