Automatic Web Content Enrichment Using Parallel Web Browsing

Michal RAČKO*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xracko@stuba.sk

In the domain of technology enhanced learning, it is important to know relevant information about learning objects and relationships in it. Adaptive learning systems are expanding the area of personalizing the learning needs of individual users. Assuming that the user behaviour is the same while following the same goal e.g. searching for additional information for a given topic, we propose a method for automatic web content enrichment based on their actions in the domain of open Web.

Currently, there are a large number of web browsers allowing tabbed browsing. All of them in their latest versions support this kind of browsing [1]. Some of them allow persistently maintaining selected tabs, or renew tab state even after the user closes the application. Various researchers found that the users use tabs in a ways like temporary lists, parallel search results, etc. Majority of these ways of usage was not explicitly planned.

The aim of this project is the relationship discovery between sites frequently visited by users using multiple tabs. What are the relations between them, or whether they can be linked together based on the way the users have accessed them. These links may not depend on the hyperlinks between sites, but based on the user browsing behaviour. The proposed method is evaluated in ALEF adaptive learning system where the aim is to make easier or automatize adding external resources to learning objects.

We can idealize a user browsing session as a sequence of actions performed during browsing [2]. Processing the tabbing actions among browser tabs can pose a problem since each tab is seen as a separate dimension. Therefore we have to propose a method that will flatten those dimensions into one sequence of user actions. Then we identify sequences that led to adding external resources to adaptive learning system. When analysing parallel browsing behaviour data, we consider different ways of switching among tabs. Each of these actions is used irregularly. Research has shown [3] that re-visitation rate of tabs is much larger than the number of sites visited. Another important information when examining user browsing behaviour is tracking

* Supervisor: Martin Labaj, Institute of Informatics and Software Engineering

Spring 2014 PeWe Workshop, March 21, 2014, pp. 107–108.

how much time he/she spent browsing the content of the page [2] and thus we can reconstruct the actual page viewing.

When browsing the web, users also perform unwanted actions in addition to the ordinary actions (reading the content, switching between tabs), which we have to remove in pre-processing, such as actions of fast switching between tabs. Then we classify pages into categories: adaptive system, digital library, search engine and other. Web browser usage data is a continuous record, so we have to be able to identify the user sessions, that can be determined based on the number of existing tabs, and when the number reaches zero, it means the end of session.

In the last phase of pre-processing, we divide the continuous record to sessions and then divide these sessions into loops. A loop is defined as the smallest sequence of actions, which starts and ends in the same learning object. Between the initial and final action, there must be at least one other switch action between pages. The loops that are not closed by the end of a session are discarded. The resulting loops contain potentially relevant external resources to the learning object that began the loop.

We adjust the weighting of web pages with a function defining ratio between the active time spent on the site and its significance. The category represents 70% of final page weight, which is assigned from 0 to 1. Browsing time weight multiplier is logarithmically dependent on the time spent on the page, where the upper limit is empirically defined to be 20 minutes. Final formula for weight calculation is (1).

$$w = 0.7 \times w_c + 0.3 \times \log(t) \tag{1}$$

w_c is category importance and t is browsing time in seconds. Method output is a set of pairs learning object – external resource. Best method for discovery of loops potentially containing relevant external resource to a learning object was Naive Bayes that has 93.24% recall rate for chosen class. Trained classification method is used for further loop classification.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 173-178.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Dubroy P., Balakrishnan R. 2010. A study of tabbed browsing among mozilla firefox users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, ACM Press, pp. 673-682 (2010)

[2] Labaj, M., Bieliková, M. Modeling parallel web browsing behavior for web-based educational systems. In *2012 IEEE 10th Int. Conf. on Emerging eLearning Technologies and Applications (ICETA 2012)*, IEEE, pp.229-234 (2012)

[3] Zhang H., Zhao S. Measuring web page revisitation in tabbed browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, pp. 1831-1834 (2011)