Automatic Web Content Enrichment Using Parallel Web Browsing

Michal RAČKO*

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

Creating links between resources on the Internet is now an acute problem due to the large amount of various content that it includes. In the past content could only be created by the authors of pages, but now at the time of Web 2.0 Internet users can already add content themselves. This is main cause od improper structure of such content and weak or no links between similar sources.

Creation of a clear and long-term sustainable structure of the Web is important because of easier navigation when browsing or searching. Meta-information and semantics of each page allows more accurate searching and thus provide search engines with ability of creating a complete picture of the sites area. Based on the information about content it is also possible to create links between similar resources without need of physical links between them.

Therefore it is necessary resolve problem mentioned above, but most of the existing solutions are inadequate or methods that have been chosen to solve this issue are not very suitable. Therefore it is necessary to propose a method able to provide us with additional information about web resources and thus allow easier creation of links between those sites. Such a solution should have the least impact on users and the way they work with the resources on the Internet and not try to change their approach.

Currently, there is large number of web browsers allowing tabbed browsing. The most common ones are Internet Explorer, Chrome and Firefox. All of them in latest versions support this kind of browsing [2]. Some of them allow persistently maintain the selected tabs, or renew their state even after you close the application.

Various researches found that user uses tabs in large number of ways. Among them are temporary list, parallel search results tab to return to the previous page, and so on., while the majority of these ways of usage were not planned [2].

For the purpose of creating links between Web resources and purpose of their markup is necessary to provide a domain model for specific domain, due to differences between them. It is then possible to create model for selected domain, which represents

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

Spring 2013 PeWe Workshop, April 5, 2013, pp. 85–86.

the semantic description of the relevant sources, relations between them and their evaluation. Such an explicitly defined model we can pass to an algorithm or a program that is able to work with it.

Domain modelling is indirectly connected to creation of a user model itself. It can be created by defining user characteristic, demographic information, and then placed in one of the predefined groups. There are two basic user models [1]:

- Stereotype
- Overlay

Stereotype model describes and categorizes individual users to the default groups based on characteristics (demographics, knowledge level) into groups of similar users. Each group can be created with varying degrees of tolerance of similarity that is necessary for us.

Overlay model refers to the degree to which the selected user has the properties that describe the domain model. For each characteristic, the user is assigned a value that expresses degree to which is selected feature true for him. These models can be dynamically changed by user actions and therefore is more dynamic than the stereotype model.

Obtaining relevant content on the web today is quite a difficult task, since the pages are structured in different ways and contain a lot of useless information unrelated to its content, such as advertisements. Therefore, it is necessary while extracting relevant content to get rid of these parts [3] that are unnecessary for the content analysis.

The aim of my diploma thesis is the relationship discovery between the sites frequently visited by users using multiple tabs. What are the relations between them, or whether they can be linked together based on way they were created while using user model for particular domain. These links may not be dependent on the physical interconnection of sites, but based on the similarities in content, which may be too unstructured. It is also possible to take into account how the individual pages in tabs were created or used and then adjust the strength of the relationship between resources.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Brusilovsky, P. (1996). Methods and Techniques of Adaptive Hypermedia. User Model. User-Adapt. Interact., 6(2-3):87–129.

[2] Dubroy, P., & Balakrishnan, R. (2010). A study of tabbed browsing among mozilla firefox users. Proceedings of the 28th international conference on Human factors in computing systems - CHI ’10, 673.

[3] Jing Li and C. I. Ezeife. 2006. Cleaning web pages for effective web content mining. In Proceedings of the 17th international conference on Database and Expert Systems Applications (DEXA'06), Stéphane Bressan, Josef Küng, and Roland Wagner (Eds.). Springer-Verlag, Berlin, Heidelberg, 560-571.