Web Usage Driven Adaptation of the Semantic Web

Alexander Mikroyannidis, Babis Theodoulidis

School of Informatics, University of Manchester,
Sackville Street, Manchester M60 1QD, United Kingdom
{A.Mikroyannidis, babis.theodoulidis}@manchester.ac.uk

Abstract. The notion of the Semantic Web has emerged as a solution to the problem of organizing the immense information provided by the World Wide Web. However, this information has to be constantly updated and reorganized in order to better serve the changing needs of the web users. A static Semantic Web can therefore be of little use in the environment of the ever-transforming World Wide Web. In the context of the present work, we propose a framework for web usage driven adaptation of the Semantic Web. Based on the usage of the web, we perform evolution of its topology and ontology. This procedure aims to facilitate the way the user interacts with the web, resulting in an increase in the usability of the web through the refinement of its physical and semantic structure.

1 Introduction

Being a large and dynamic information source, both structurally complex and ever growing, the World Wide Web poses great difficulties to its full exploitation. The Semantic Web addresses this problem by "giving information a well-defined meaning, better enabling computers and people to work in cooperation" [1]. This is implemented by expressing the web data in forms that are machine-understandable and machine-processable, in order to be more efficiently maintained by software agents.

Nevertheless, a significant issue, which is usually overlooked, is the usability of the Semantic Web. The way with which a user browses the web is heavily dependent on his needs, knowledge and interests. These needs and interests have to be addressed by the Semantic Web, in order for enhancement to be achieved in the user's interaction with the web. Moreover, since these preferences are altered through time, the Semantic Web must have the ability to satisfy them through a constant adaptation process.

In [8, 9] we introduced a framework for self-adaptive web sites. The present paper extends this work by addressing the adaptation of the Semantic Web, based on web usage data. A framework that employs web usage mining as well as text mining methodologies is presented. The proposed framework adapts the web in order to assist the users in their browsing tasks. Both the physical and semantic structure of the web are targeted. The web site ontology is semi-automatically built and evolves through the adaptation procedure.

The remainder of this paper is organized as follows: Section 2 describes the approaches that have been followed by researchers in the area of web adaptation. Section 3 introduces the theoretical principles upon which our framework was built. Section 4 presents an architecture that implements the framework. Section 5 discusses the results of the proposed approach on the usability of the web. Finally, the paper is concluded and some plans for future work are provided.

2 Related Work

Providing users with assistance in their web navigation can help keep them in a web site, or even attract more visitors. This has always been a popular subject, especially in the e-commerce domain. Several systems have been developed towards this direction. WebWatcher [7] suggests links that may interest a user, based on the online behaviour of other users. Each user is asked, upon entering the site, what kind of information he is seeking. Before he departs, he is asked whether he has found what he was looking for. His navigation paths are used to deduce suggestions for future visitors that seek the same content. These suggestions are visualized by highlighting existing hyperlinks.

The Avanti project [6] tries to predict the user's final objective as well as his next step. A model for the user is built, based partly on information the user provides about him. His interests are also extracted from his navigation paths. Visitors are provided with direct links to pages that are probably the ones they are looking for. In addition, hyperlinks that lead to pages of potential interest to each visitor are highlighted.

A drawback of both the WebWatcher and the Avanti system is that they require the active participation of the users in the adaptation process, by asking them to provide information about themselves. On the other hand, the Footprints [13] system relies entirely on the navigation paths of the users. The system does not perform user identification. All navigation paths are recorded and the most frequent ones are presented to the visitor, in the form of maps or trails. Html pages also display next to each link the percentage of people who have followed it. Nevertheless, as in the WebWatcher and the Avanti systems, no adaptation of the site's structure is performed.

Perkowitz et al [12] have presented a conceptual framework for adaptive web sites. They have focused on the semi-automatic creation of index pages, based on discovering clusters of pages. They assume that if a large number of visitors frequently visit a set of pages, this provides strong evidence that these pages are related. They have developed two cluster mining algorithms, PageGather and IndexFinder. The first one relies on a statistical approach to discover candidate link sets, while the second is a conceptual cluster mining algorithm, as it finds link sets that are conceptually coherent. They have also performed experiments on three web sites by placing the automatically generated pages online and observing the user response.

However, the majority of the existing approaches in web adaptation lack in a crucial factor: they do not address the semantic aspect of the web. The ontological perspective is overlooked and the researchers' attention is drawn mainly by the site topology. Even though the improvement of the site topology is unquestionably signifi-

cant, we should not disregard the fact that users browse a site mainly for its content. Consequently, the content classification structure should also be adaptive through the evolution of the site ontology. The innovative concept of the Semantic Web is a most suitable region for applying such adaptation methodologies, targeting to the direct benefit of the end users.

3 Framework

The proposed framework defines the adaptation process as absolutely *transparent* to the user, requiring no active participation from him. In addition, the adaptations of our framework perform *web transformation*, instead of focusing on personalization tasks. Mobasher et al [10] define personalization as "any action that tailors the web experience to a particular user, or a set of users". On the other hand, according to Perkowitz et al [11], transformation is "improving the site's structure based on interactions with *all* visitors". The advantage of this approach is that it does not require user identification, which cannot be safely performed from usage data, unless the user contributes in an explicit or implicit way [5]. Nevertheless, most users are reluctant to give away personal information. Moreover, through transformation, transparency is achieved, as the adaptation procedure relies completely on the data gathered in the access logs.

Coenen et al [2] distinguish between *tactical* and *strategic* adaptations in their framework for self-adaptive web sites. They call tactical the adaptations that can be performed in real time, without the webmaster's approval, since they do not affect the overall site structure. On the other hand, strategic adaptations are the ones that "go against or conflict with the original beliefs of the site, and consequently have an important influence on the original site-structure". Coenen et al suggest that such modifications should be performed offline, with the approval of the webmaster.

The role of the webmaster is considered fundamental in the present framework. Human designers often dedicate a large effort in developing a site. By no means, the adaptation process should undo their work. The framework puts the webmaster in charge of the adaptation procedure, by requiring from him to approve the adaptations. In addition, we propose an adaptation engine that will learn from the webmaster's responses. Instead of predefining which modifications are strategic and which tactical, the adaptation system should gradually learn to classify the adaptations, by studying the webmaster's approvals and rejections of proposed adaptations. Adaptations that are classified by the system as tactical should be applied automatically, without the webmaster's interference. In this way, the site will adapt not only to the end user's preferences, but to the webmaster's as well.

In order to improve the reorganization of the information provided by a web site, we have exploited the semantic aspect of the web. Apart from the topology of the web site, the framework also addresses the evolution of the site ontology. A web site ontology is strongly related to the topology of the site. It is comprised of the thematic categories covered by the site's pages. These categories are the concepts of the ontology. Each web page, depending on its content, is an instance of one or more concepts of the ontology. The concepts can be organized in a hierarchy, representing an "is a"

relationship. This means that a class is a subclass of another class if every instance of the second class is also an instance of the first.

Figure 1 illustrates the web site ontology of the University of Manchester School of Informatics (http://www.co.umist.ac.uk). The ontology has been built considering the organization of the thematic categories as this is defined in the current topology of the site. The hierarchy's top level contains seven classes: School, Undergraduate Programmes, Postgraduate Taught Programmes, Postgraduate Research, Research, News and Intranet. These are the main thematic categories of the site. These categories are then expanded to more specific concepts, which are represented by subclasses. All the concepts are instantiated in the web pages of the site.

Fig. 1. The School of Informatics web site ontology

It must be stressed that the web site ontology is quite different from the domain ontology [4]. The latter describes relationships between the concepts of a domain, whereas the first is based on the organization of the information found in a web site. The ontology of a domain is usually more complex than the ontology of a web site related to the same domain. However, the maintenance of a web site ontology requires considerable effort and has to be performed on a regular basis, since the content of a web site is constantly updated.

The adaptation of the ontology can include the discovery of new associations between its concepts. Moreover, a concept can be found to have more than one superclasses or a web page to be an instance of more than one concepts. Finally, a web page may possibly need to be categorized under a different concept than its current one.

4 Architecture

Figure 2 presents an architecture implementing the theoretical principles of the proposed framework. As it can be seen, the inputs of the adaptation process consist of the server's access logs, the site topology and ontology. The whole procedure aims at the evolution of the topology and ontology of the web site.

The adaptation starts with a preprocessing stage, during which the data stored in the raw access logs are cleaned and visiting sessions are identified. The sessions are then mined with the use of Frequent Itemset Mining algorithms in order to produce *pagesets*. We call pagesets the sets of pages that are frequently accessed together throughout the same session.

The extracted pagesets are classified in relation to certain features of their pages. More specifically, two classification criteria have been used: linkage state and content. The first criterion refers to the connection that the pages of each pageset have, according to the site structure. The key factor is whether the pages contained in a pageset, are directly linked to each other or not. Pagesets of unlinked pages might suggest the insertion of shortcut links between these pages, in order to achieve shorter navigation paths. From the pagesets of linked pages, changes in the appearance of existing links can be extracted. For example, if an index page and some of its links comprise one or more pagesets, then by highlighting these links in the index page, first time visitors would be able to navigate the site easier.

The second classification criterion refers to the content of the pages contained in each pageset. The pages of the pagesets are classified in order to discover new associations between the concepts of the site ontology. More particularly, if a pageset includes pages belonging to concepts that were not previously linked, the ontology should then be modified to reflect the relevance these concepts have, according to the preferences of the users.

Based on the linkage state and content classification, a report containing proposals for the improvement of the site is generated. This report contains proposals for the insertion of shortcut links from source pages to target pages that are frequently accessed together but are currently not linked. It also contains proposals for the change

of the appearance of popular hyperlinks. Furthermore, the report contains proposals for the evolution of the site ontology.

After the proposed modifications have been revised by the webmaster, they can be applied to the web site. The site topology is then refined through the insertion of new shortcut links, as well as changes in the appearance of the existing ones. The ontology is also refined in a number of ways.

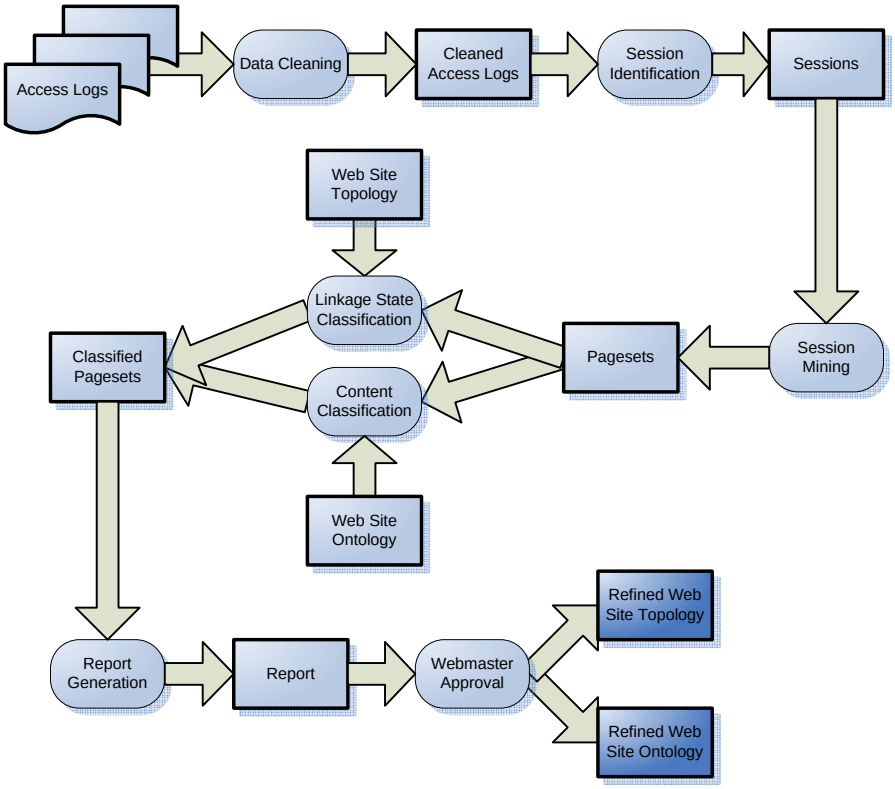

Fig. 2. Web site adaptation architecture

5 Results and Discussion

We have applied our methodology on the University of Manchester School of Informatics web site. The topology of the web site was refined through the insertion of new shortcut links between pages that were not previously linked together, as well as through the highlighting of popular existing links. In addition, the web site ontology was modified in several ways, based on the outcomes retrieved from the classified pagesets.

More specifically, the adaptation system produced two sets of reports: shortcut links reports and highlighted links reports. Figure 3 shows an extract from a report containing proposals for insertion of shortcut links. From a source page, shortcut links to target pages are suggested. The target pages have been found to be frequently visited after the source page. However, the source page is not linked to the target pages, thus forcing the users to follow alternative paths in order to reach them. Shorter navigation paths can be therefore achieved if the source page is linked to the target pages. This is the purpose of this type of report. For instance, some pages that contain additional resources on certain courses are frequently accessed by users after accessing the "Information on Modules" page, which contains a list of all the department's modules. Consequently, as it can be seen in Figure 3, shortcut links are proposed that lead to these pages.

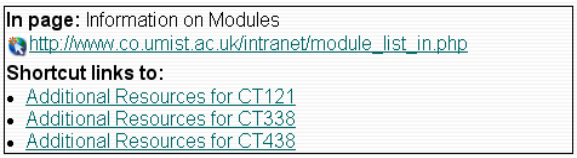

Fig. 3. Extract from a shortcut links report

The highlighted links report is comprised of suggestions for emphasizing popular hyperlinks. This can be quite useful, especially in pages that contain large amounts of hyperlinks, such as index pages. In such cases, the user can gain valuable time if prompted with the most popular choices. Figure 4 shows an example of a highlighted links report. Certain links, based on their popularity have been proposed to be suggested to the user who visits the "Postgraduate Research Programmes" page.

Fig. 4. Extract from a highlighted links report

Figure 5 shows an example of a modified web page, according to our system's suggestions. It is the "Information on Modules" page, which is very popular in the users' preferences. The page has been modified to facilitate the navigation of the users during the first semester. Shortcut links to popular courses of the first semester have been inserted in the left side of the page, under the title "Quick links". Moreover, popular links that already existed, such as the hyperlink leading to the page of the "Personal and Professional Development" course, have been highlighted.

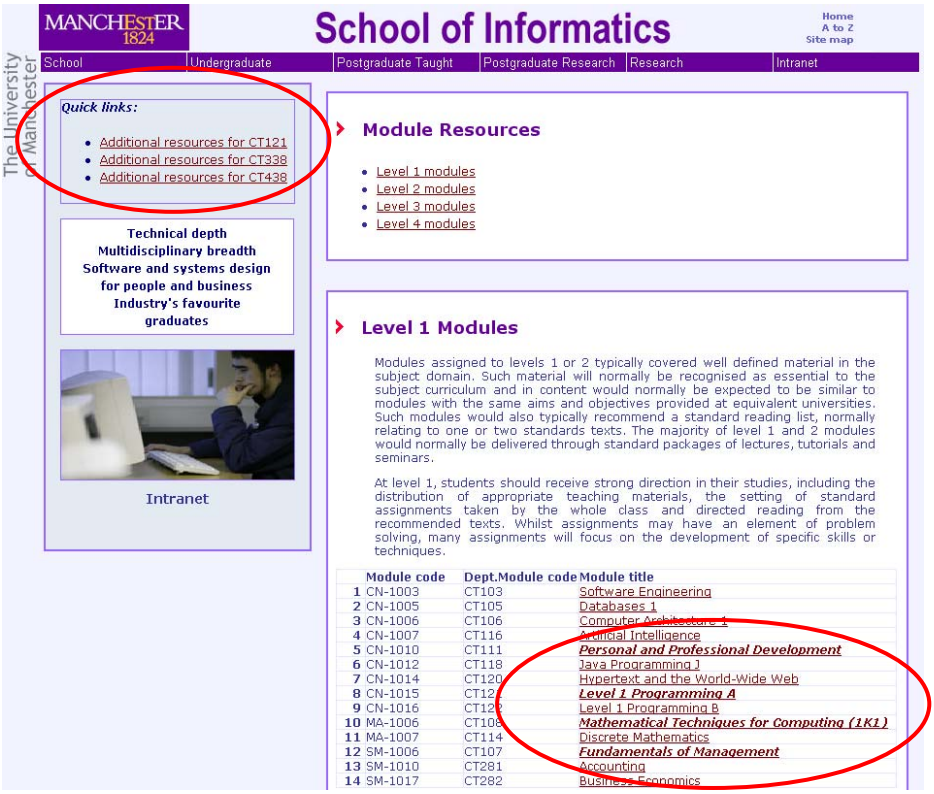

Fig. 5. Example of modified web page

The web site ontology was modified in several ways, based on the outcomes retrieved from the classified pagesets. The resulting ontology, after the application of adaptations suggested by our system, is shown in Figure 6. Based on these adaptations, the content organization of the site was altered to better satisfy the needs of its visitors. For the content classification of the web pages belonging to the pagesets, a classifier implementing the Support Vector Machines categorization algorithm [3] was used.

First of all, new associations were discovered between concepts. These associations reflect the interests of the users, as documents belonging to these concepts are frequently accessed together. New associations were inserted between the following concepts:

- “Research” and “Students”
- “Research” and “Staff”
- “School” and “Students”
- “School” and “Staff”
- “Students” and “Staff”

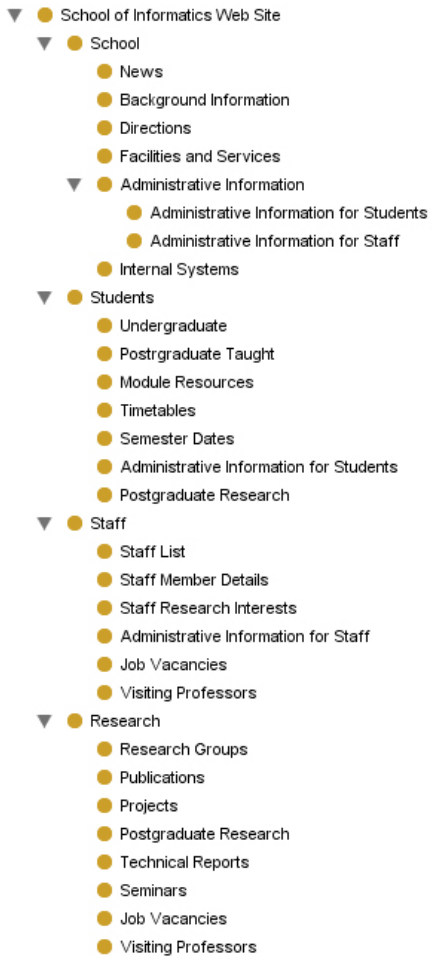

Fig. 6. Refined web site ontology of the School of Informatics

Reorganization of the concepts' hierarchy was also performed. Further improvements included the creation of new categories, the removal of existing categories, as well as changes to the levels of hierarchy that the concepts belong to. For instance, the "Staff" concept was previously a subconcept of the "School" concept, which resided in the highest level of the ontology. It should be noted that the "Staff" concept has as instances all the web pages that carry information about the staff members of the school. However, the high frequency with which this concept appeared in the pagesets implies the significance that it has in the interests of the users. It would be thus appropriate to transfer this concept to the top level of the ontology, as shown in Figure 6. Based on the performed classification, the undergraduate and postgraduate programmes were grouped under the more general concept "Students". The "School" concept was also extended to include more subconcepts.

The ontology of the site was extended to include multiple instances of concepts or multiple subconcepts. The categorization of the web pages that was carried out, suggested that several pages belong to more than one concept. Moreover, in some cases, web pages and the corresponding concepts were categorized under different concepts than they previously were in the existing ontology. The site ontology should be therefore updated in order to reflect this fact. For example, the "Job vacancies" web page, which corresponds to the "Job Vacancies" concept, was found to be an instance of both the "Staff" and "Research" concepts. The information contained in this page regards mainly research job posts and is also highly related to the "Staff" concept. This page was previously categorized only under the "School" concept. In the updated ontology (Figure 6), the "Job Vacancies" concept has been placed both under the "Staff" and "Research" concepts. The same modification has been applied to the concepts "Visiting Professors", "Administrative Information for Students", etc.

Finally, useful conclusions were deduced about the usage of the web site. Particularly, the thematic category that was the first in the preferences of the users was, as expected, the "Students" concept. This concept contains all pages that support the school's modules, both undergraduate and postgraduate. This is not surprising, since most of the traffic is generated by the students. Second in the users' interests comes the "Staff" concept. The "Research" concept is third, followed by the "School" category. These results can be used to enhance the performance of the server, for example by the use of additional servers that will host the popular resources, or to promote the problematic concepts by making them more easily accessible.

6 Conclusions and Future Work

The present work investigated a web usage driven approach on the adaptation of the Semantic Web. A framework was introduced that enables adaptation of the web topology and ontology to the needs and interests of web users. In addition, an architecture based on the principles of the framework was presented. The proposed adaptation process exploits the access data of the users, together with the semantic aspect of the web, in order to facilitate web browsing.

A real web site was used as a case study, in order to study the impact that the proposed framework can have on the usability of the web. The topology and ontology of the site were refined in several ways. Apart from changes in specific web pages, enhancements of the whole formation of the site were derived. Furthermore, useful knowledge was acquired, regarding the overall usage of the site. The sections that mostly interest the users were identified, leading to further improvements in their usability. Moreover, the regions of the site that need more promotion were revealed.

The current framework regards each web site as a separate unit. In future work, we plan to extend this approach, by performing simultaneous adaptation of multiple web sites. This task requires consideration of the relationships between the topologies and ontologies of different web sites. This extension is necessary in order to view the World Wide Web as an integral whole, towards the development of the Adaptive Semantic Web.

References

[1] Berners-Lee, T., Hendler, J., and Lassila, O. *The semantic web*. Scientific American, 2001. **279**(5): p.34-43.

[2] Coenen, F., Swinnen, G., Vanhoof, K., and Wets, G. *A Framework for Self Adaptive Websites: Tactical versus Strategic Changes.* In *Proc. of WEBKDD'2000 Web Mining for E-Commerce - Challenges and Opportunities, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2000. Boston.

[3] Cortes, C. and Vapnik, V. *Support Vector Networks*. Machine Learning, 1995. **20**(3): p.273-297.

[4] Daconta, M., Obrst, L., and Smith, K. *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management.* Wiley, 2003.

[5] Eirinaki, M. and Vazirgiannis, M. *Web Mining for Web Personalization*. ACM Transactions on Internet Technology, 2003. **3**(1): p.1-27.

[6] Fink, J., Kobsa, A., and Nill, A. *User-Oriented Adaptivity and Adaptability in the AVANTI project.* In *Proc. of Designing for the Web: Empirical*. 1996.

[7] Joachims, T., Freitag, D., and Mitchell, T. *WebWatcher: A Tour Guide for the World Wide Web.* In *Proc. of International Joint Conference on Artificial Intelligence*. 1997. Nagoya, Japan, p.770-775.

[8] Mikroyannidis, A. *Development of a framework for self-adaptive web sites*. School of Informatics, University of Manchester, MPhil Thesis, 2004.

[9] Mikroyannidis, A. and Theodoulidis, B. *A Theoretical Framework and an Implementation Architecture for Self Adaptive Web Sites.* In *Proc. of IEEE/WIC/ACM International Conference on Web Intelligence (WI'04).* 2004. Beijing, China, p.558-561.

[10] Mobasher, B., Cooley, R., and Srivastava, J. *Automatic Personalization Based on Web Usage Mining*. Communications of the ACM, 2000. **43**(8): p.142-151.

[11] Perkowitz, M. and Etzioni, O. *Adaptive Web sites*. Communications of the ACM, 2000. **43**(8): p.152-158.

[12] Perkowitz, M. and Etzioni, O. *Towards adaptive Web sites: Conceptual framework and case study*. Artificial Intelligence, 2000. **118**(1-2): p.245-275.

[13] Wexelblat, A. and Maes, P. *Footprints: History-Rich Tools for Information Foraging.* In *Proc. of Proceedings of Human Factors in Computing Systems (CHI).* 1999. Pittsburgh, Pennsylvania, United States, p.270-277.