WEB SITE ONTOLOGY EVOLUTION THROUGH WEB SITE ADAPTATION

A. Mikroyannidis, B. Theodoulidis
School of Informatics, University of Manchester, Sackville Street Manchester, M60 1QD United Kingdom

Key words to describe the work: Web Site Ontology, Adaptive Web Sites, Web Mining, Text Classification

Key Results: Semi-automatic construction and maintenance of web site ontology.

How does the work advance the state-of-the-art?: A new approach to web site adaptation that results in a novel methodology in web site ontology evolution.

Motivation (problems addressed): The lack of web sites that can dynamically adapt their semantic structure to their users' needs.

Introduction

The web user often faces information overload, which prevents him from reaching the content that interests him. A solution to this problem can be provided by a web site ontology that evolves based on the user's needs. In this paper, a methodology for the semi-automatic construction and maintenance of a web site ontology is presented. The results of the application of this methodology on a real web site are also presented and discussed.

Web Site Ontology

The ontology of a web site 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. Alternatively, a subclass of a class represents a concept that is a "kind of" the concept that the superclass represents.

Web Site Adaptation

We have introduced in [1, 2] a framework for self-adaptive web sites. This framework aims to the adaptation of the physical as well as the semantic structure of a web site, based on the site's usage data. In addition, an architecture that implements this framework has been introduced, which employs web usage mining as well as text mining methodologies for the offline adaptation of a web site.

Figure 1 illustrates the proposed architecture. It 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's topology. The second classification criterion is based on the content of the pages contained in each pageset.

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

Web Site Ontology Evolution

We applied our methodology on the web site of the Informatics School of the University of Manchester. Figure 2 shows the ontology that was developed for this site. The ontology was semi-automatically constructed with the use of a classification tool based on the Support Vector Machines (SVM) categorization algorithm [3]. First, we defined the thematic categories, according to the topology of the site, and then trained the algorithm on these categories with sample web pages. Afterwards, the rest of the site's pages were automatically categorized under the predefined categories.

The ontology was modified in several ways, based on the outcomes retrieved from the classified pagesets. First of all, new associations were discovered between concepts. These associations reflect the users' preferences, as documents

belonging to these concepts are frequently accessed together.

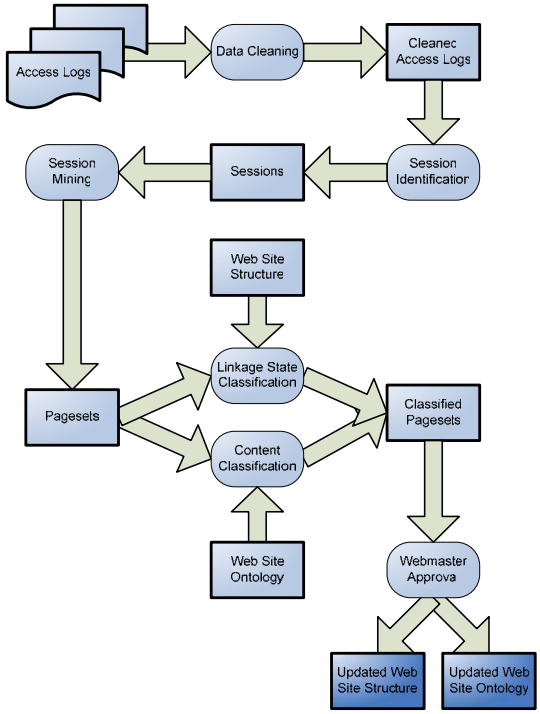

Figure 1. Web site adaptation architecture

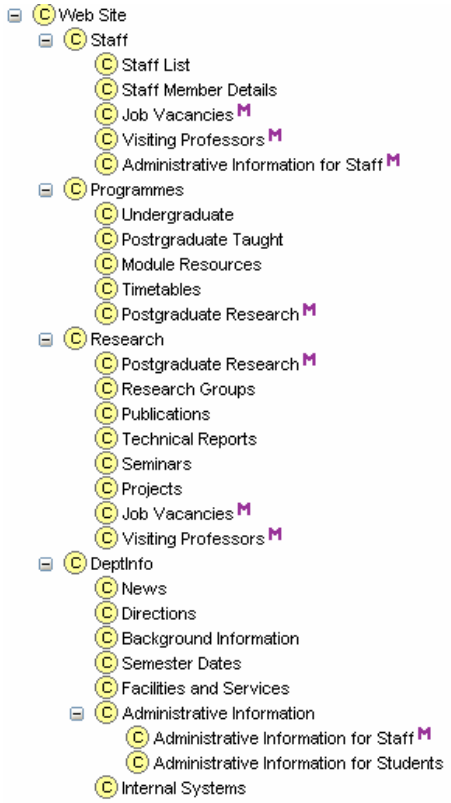

Figure 2. The case study's web site ontology

Reorganization of the concepts' hierarchy was also performed. Further improvements included the creation of new categories, the removal of others, as well as changes in the levels of hierarchy that the concepts belong to. For instance, the Staff concept was transferred to the highest level of the ontology. The high frequency with which this concept appeared in our pagesets implies the significance that the Staff concept has in the users' preferences. Based on the classification performed by our system, the undergraduate and postgraduate programmes were grouped under the more general concept Programmes. The Departmental Information concept was also extended to include more subconcepts.

The ontology of the site was extended to include multiple instances of concepts and 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 site's topology.

Finally, useful conclusions were deduced about the usage of the web site. Particularly, the most and least popular thematic categories according to the users' preferences were discovered. These results can be used to enhance the performance of the server, or to promote the problematic concepts by making them more easily accessible.

Conclusions

Web site adaptation is valuable, considering the volume of information provided in the web. We propose the evolution of a site's ontology, using web mining and text classification techniques. The modifications performed on the site's ontology are reflected on its topology, in order to produce a site that best satisfies the needs of its visitors.

References

1. Mikroyannidis, A., *Development of a framework for self-adaptive web sites*, School of Informatics, University of Manchester, MPhil Thesis, 2004.
2. Mikroyannidis, A. and Theodoulidis, B., "A Theoretical Framework and an Implementation Architecture for Self Adaptive Web Sites", in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), Beijing, China, 2004, pp. 558-561.
3. Vapnik, V., *The Nature of Statistical Learning Theory*, New York, Springer, 1995.