SEWeP: A Web Mining System supporting Semantic Personalization

Stratos Paulakis, Charalampos Lampos, Magdalini Eirinaki, Michalis Vazirgiannis
Athens University of Economics and Business, Department of Informatics
{paulakis, lampos, eirinaki, mvazirg}@aueb.gr

Abstract. We present SEWeP, a Web Personalization prototype system that integrates usage data with content semantics, expressed in taxonomy terms, in order to produce a broader yet semantically focused set of recommendations.

1. Introduction

Web personalization is the process of customizing a Web site to the needs of each specific user or set of users. Most of the research efforts in Web personalization correspond to the evolution of extensive research in Web usage mining. When a personalization system relies solely on usage-based results, however, valuable information conceptually related to what is finally recommended may be missed. To tackle this problem we propose a Web personalization framework based on semantic enhancement of the Web usage logs and the related Web content. We present SEWeP, a Web Personalization prototype system, based on the framework proposed in [3]. This system integrates usage data with content semantics, expressed in taxonomy terms, in order to produce semantically enhanced navigational patterns that can subsequently be used for producing valuable recommendations.

2. SEWeP Framework

The innovation of the SEWeP prototype system is the exploitation of web content semantics throughout the Web mining and personalization process. It utilizes Web content/structure mining methods to assign semantics to Web pages and subsequently feeds this knowledge to Web usage mining algorithms. Web content is semantically annotated using terms of a predefined domain-specific taxonomy (categories) through the use of a thesaurus. These annotations are encapsulated into C-logs, a (virtual) extension of Web usage logs. C-logs are used as input to the Web usage mining process, resulting in a set of rules/patterns consisting of thematic categories in addition to URIs. Furthermore, the semantically annotated Web pages are organized in coherent clusters (using THESUS system [6]) based on the taxonomy. These clusters are then used in order to further expand the set of recommendations provided to the end user. The whole process results in a broader, yet semantically focused set of recommendations. The system architecture is depicted in Figure 1. The main functionalities of the demonstrated system are described below. A more detailed description can be found in [2,3].

Logs Preprocessing: The system provides full functionality for preprocessing any kind of Web logs, by enabling the definition of new log file templates, filters (including/excluding records based on field characteristics), etc. The new logs are stored in new ("clean") log files.

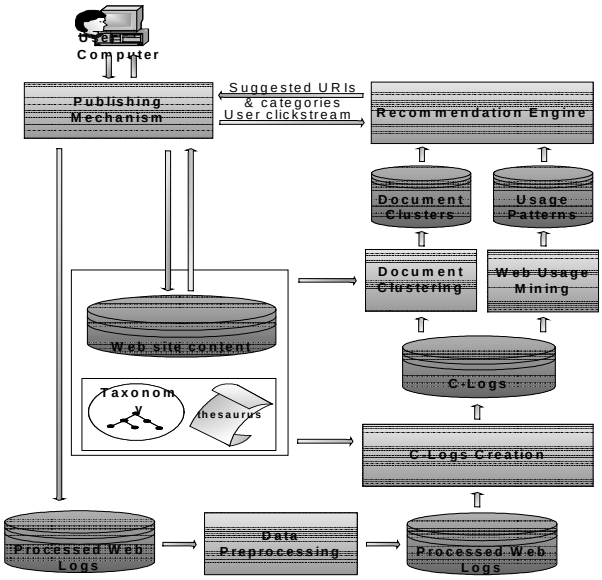

Fig. 1. System architecture

Content Retrieval: The system crawls the Web and downloads the Web site's pages, extracting the plain text from a variety of crawled file formats (html, doc, php, ppt, pdf, flash, etc.) and stores them in appropriate database tables.

Keyword Extraction & Translation: The user selects among different methods for extracting keywords. Prior to the final keywords selection, all non-English keywords are translated using an automated process (the system currently also supports Greek content). All extracted keywords are stored in a database table along with their relevant frequency.

Keyword – Category Mapping: The extracted keywords are mapped to categories of a domain-specific taxonomy. The system finds the "closest" category to the keyword through the mechanisms provided by a thesaurus (WordNet [7]). The weighted categories are stored in XML files and/or in a database table.

Session Management: SEWeP enables anonymous sessionizing based on distinct IPs and a user-defined time limit between sessions. The distinct sessions are stored in XML files and/or database tables. (Figure 2 includes a screenshot of this module)

Semantic Association Rules Mining: SEWeP provides a version of the apriori algorithm [1] for extracting frequent itemsets and/or association rules (confidence and support thresholds set by the user). Apart from URI-based rules, the system also provides functionality for producing category-based rules. The results are stored in text files for further analysis or use by the recommendation engine.

Clustering: SEWeP integrates clustering facilities for organizing the results into meaningful semantic clusters. Currently SEWEP capitalizes on the clustering tools available in the THESUS system [6].

Recommendations: The (semantic) association rules/frequent itemsets created feed a client-side application (servlet) in order to dynamically produce recommendations to the visitor of the personalized site.

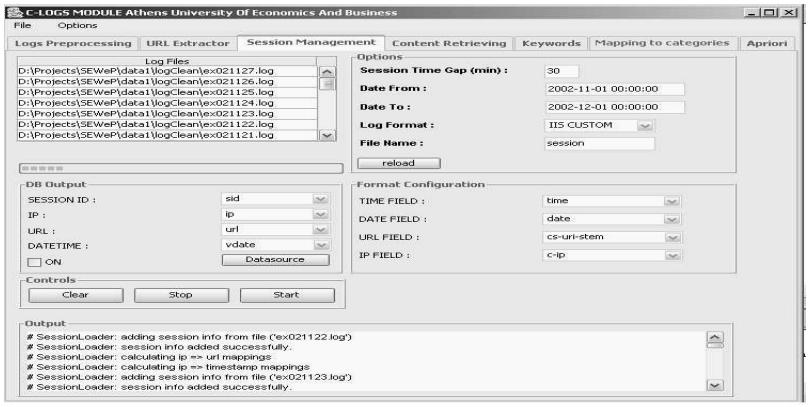

Fig. 2. SEWeP screenshot: Session Management module

3. Implementation Details

The SEWeP system is entirely based on Java (JDK 1.4 or later). For the implementation of SEWeP we utilized the following third party tools & algorithms: PDF Box Java Library, Jacob Java-Com Bridge Library, and swf2html library (for text extraction); Xerces XML Parser; Wordnet v1.7.1 Ontology; JWNL and JWordnet 1.1 java interfaces for interaction with Wordnet; Porter Stemming Algorithm [4] for English; Triantafillidis Greek Grammar [5]; Apache Tomcat and 4.1 and Java Servlets for recommendation engine; JDBC Library for MS SQL Server.

4. Empirical Evaluation

The main advantage of SEWeP is the involvement of semantics in the recommendation process resulting in semantically expanded recommendations. As long as the system's effectiveness is concerned, we have performed a set of user-based experiments (blind tests), evaluating SEWeP's usefulness, i.e. whether the semantic enhancement results in better recommendations [2,3]. The experiments' results verify our intuitive assumption that SEWeP enhances the personalization process, since users evaluate the system's recommendations as of high quality.

References

1. R. Agrawal, R. Srikant, Fast Algorithms for Mining Association Rules, in Proc. of 20th VLDB Conference, 1994
2. M. Eirinaki, M. Vazirgiannis, H. Lampos, S. Pavlakis, Web Personalization Integrating Content Semantics and Navigational Patterns, submitted for revision at WIDM 2004
3. M. Eirinaki, M. Vazirgiannis, I. Varlamis, SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process, in Proc. of the 9th SIGKDD Conf., 2003
4. M. Porter, An algorithm for suffix stripping, Program (1980)/ Vol. 14, No. 3, 130-137
5. M.Triantafillidis, Triantafillidis On-Line, Modern Greek Language Dictionary, http://kastor.komvos.edu.gr/dictionaries/dictonline/DictOnLineTri.htm
6. I. Varlamis, M. Vazirgiannis, M. Halkidi, B. Nguyen. THESUS: Effective Thematic Selection And Organization Of Web Document Collections Based On Link Semantics, to appear in IEEE TKDE Journal
7. WordNet, http://www.cogsci.princeton.edu/~wn/