SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process

M. Eirinaki, M. Vazirgiannis, I. Varlamis

Athens University of Economics and Business
Department of Informatics
Patision 76, Athens, 10434, GREECE
(30210) 8203519
{eirinaki, mvazirg, varlamis}@aueb.gr

ABSTRACT

Web personalization is the process of customizing a Web site to the needs of each specific user or set of users, taking advantage of the knowledge acquired through the analysis of the user's navigational behavior. Integrating usage data with content, structure or user profile data enhances the results of the personalization process. In this paper, we present SEWeP, a system that makes use of both the usage logs and the semantics of a Web site's content in order to personalize it. Web content is semantically annotated using a conceptual hierarchy (taxonomy). We introduce C-logs, an extended form of Web usage logs that encapsulates knowledge derived from the link semantics. C-logs are used as input to the Web usage mining process, resulting in a broader yet semantically focused set of recommendations.

Categories

H.2.8 [Database Management]: Database Applications - Data Mining; H.3.5 [Information Storage and Retrieval]: Information Services - Web-based services

Keywords

Web personalization, Web Mining, Concept Hierarchies, Semantic Annotation of Web Content

1. INTRODUCTION - MOTIVATION

The continuous growth in the size and use of the World Wide Web imposes new methods of design and development of on-line information services. The need for predicting the users' needs in order to improve the usability and user retention of a Web site is more than evident and can be addressed by personalizing it. *Web personalization* is defined as any action that adapts the information or services provided by a Web site to the needs of a user or a set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the Web site. As mentioned in [26], *"The objective of a Web personalization*

system is to provide users with the information they want or need, without expecting from them to ask for it explicitly".

Most of the research efforts in Web personalization correspond to the evolution of extensive research in Web usage mining. The majority of the proposed architectures focus on the use of usage data [19, 20, 22, 23, 24, 30, 31, 34, 41] and only a few efforts also incorporate knowledge associated with the content [2, 3, 4, 13, 25, 32] or the structure [10, 32] of the Web site, or even its registration-based users' profiles. As noted in [25], usage-based personalization can be problematic either when there is not enough usage data in order to extract patterns related to certain categories, or when the site content changes and new pages are added but are not yet included in the web log. The incorporation of information related to the content and/or the structure of the Web site provides a way of overcoming such problems, thus improving the whole personalization process. In this work, we associate Web usage and content knowledge, by enhancing the information in the Web usage logs with semantics derived from the content of the Web site's pages. The enhanced Web logs, called C-Logs are then used as input to the Web mining process, resulting in the creation of a broader set of recommendations.

An important issue to be dealt with is the characterization of Web content. Research on the area of searching and querying the Web has been very active the past few years, and several methods for extracting keywords that characterize Web content have been proposed [5, 8, 18]. We have to stress at this point that Web content processing enhances it with semantic annotations. Then, in the context of Web personalization, data mining algorithms are applied to extract relevant patterns. This imposes the need for using a limited vocabulary in order to characterize content in a uniform way. Uniformity is achieved when the vocabulary used is a concept hierarchy (taxonomy). Web content is then annotated using categories of this taxonomy. Every document falls under one or more taxonomy categories, and this classification enables a personalization system to recommend documents not only based on exact keyword matching but on semantic similarity as well.

Chakrabarti et al. [7] have demonstrated that "taxonomies provide a means for designing enhanced searching, browsing and filtering systems", focusing on text databases. We extend this rationale, using taxonomies to semantically annotate Web content. This notion is further supported by the results of Srikant et. al. [37]. In this work, the problem of mining generalized association rules is introduced. They prove that the process of finding associations between a set of items belonging to a large database of transactions is improved when those items are mapped to a taxonomy. Association rule mining across different levels of this taxonomy is valuable since uninteresting or redundant rules are

pruned and rules that would be omitted due to low support are included in the form of "parent", more general, rules. This arguments in favor of our decision for using a taxonomy in the Web content characterization process, since association rule mining is the main method used in our personalization system named SEWeP (standing for <u>S</u>emantic <u>E</u>nhancement for <u>We</u>b <u>P</u>ersonalization) for extracting navigational patterns and recommendations.

In this paper we propose an architecture that makes use of both the usage and the content data of a Web site in order to personalize it. The innovation of our work lies in the introduction of C-logs (concept logs) and their use as input to the Web usage mining process. C-logs are a conceptual abstraction of the original Web usage logs based on the Web site's semantics. Another innovative feature of the proposed architecture is that it integrates a combination of IR techniques, used to characterize Web content, with the use of a domain-specific taxonomy, in order to semantically annotate this content. The keywords that are extracted using these techniques are mapped to the categories of the taxonomy, resulting in a uniform and consistent vocabulary. The semantically annotated Web documents are further clustered and ranked in order to be used as recommendations. In the proposed system the Web usage mining process is performed using as input the C-logs. Since these logs encapsulate knowledge derived from the site semantics, the results of the usage mining process are further augmented. Alternatively, the usual Web personalization process, which is based on the Web-site's logs, can be enhanced by taking into account the semantic proximity of the content. In this way, the system's suggestions are enriched with content bearing similar semantics.

The rest of this paper is organized as follows: In Section 2 an overview of the research efforts that had a big impact in the area of Web usage mining and Web personalization is given. This synopsis focuses on the systems that make use of the site semantics except for the site usage in the Web personalization process, especially ones that hierarchically categorize Web content. In Section 3, SEWeP's architecture is presented in more detail. We describe the keyword extraction-category mapping process and introduce C-logs (concept logs), an abstraction of the Web usage logs that encapsulate knowledge derived from the site semantics. Additionally, we provide a running example that illustrates how the Web site's content is employed to enhance the results of the Web personalization process. This enhancement is further supported by the results of an unbiased test, presented in Section 4. Future enhancements of the system and conclusions are presented in Section 5.

2. RELATED WORK

Lately, a lot of research projects concentrate on Web usage mining and Web personalization areas. Most of the efforts focus on extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behaviour, so that decisions concerning site restructuring or modification may then be made by humans [2, 3, 4, 6, 12, 34, 35, 36, 38, 42]. In several cases, a recommendation engine helps the user navigate through a site [19, 20, 21, 22, 23, 41]. Some of the most integrated systems provide much more functionality, introducing the notion of adaptive Web sites and providing means of dynamically changing a site's structure [29, 30, 31]. Finally, as already mentioned, only a few research projects use content or

structure data in addition to usage data for Web personalization [10, 13, 24, 25, 32]. An extensive overview of the most important research efforts in the Web mining and personalization domain can be found in [15]. In this work we focus on the systems that combine usage and content knowledge in order to dynamically modify a Web site. Additionally, we examine the use of taxonomies in related research areas.

Perkowitz et al. [29, 30, 31] were the first to refer to the notion of adaptive Web sites, defining them to be "sites that semi-automatically improve their organization and presentation by learning from visitor access patterns" [28]. The system they proposed semi-automatically modifies a Web site, creating index pages containing collections of links to related but unlinked pages. In their most recent work [32], they move from the statistical cluster-mining algorithm PageGather to IndexFinder, which fuses statistical and logical information to synthesize index pages. In this latter work, they formalize the problem of index page synthesis as a conceptual clustering problem and try to discover coherent and cohesive link sets which can be represented to a human Webmaster as candidate index pages. The difference from the previous approach based on PageGather is that information is also derived from the site's structure and content. Therefore, IndexFinder combines the statistical patterns gleaned form the log file with logical descriptions of the contents of each Web page in order to create index pages.

WebPersonalizer proposed by Mobasher et al. [22, 23], provides a framework for mining Web log files to discover knowledge for the provision of recommendations to current users based on their browsing similarities with previous users, relying on anonymous usage data. This framework was extended in a more recent work [24, 25] to incorporate content profiles into the recommendation process as a way to enhance the effectiveness of personalization actions. Usage profiles and content profiles are represented as weighted collections of page view records. The content profiles represent different ways in which pages with partly similar content may be grouped together. The overall goal is to create a uniform representation for both content and usage profiles in order to integrate them more easily. The system is divided into two modules, the offline, which is comprised of data preparation and specific Web mining tasks, and the online component, which is a real-time recommendation engine. The recommendation engine matches each user's activity against these profiles and provides to him a list of recommended hypertext links. In their most recent work [13], they present a general framework where domain ontologies are used for automatically characterizing these profiles.

Berendt et al. introduced "service based" concept hierarchies in [4], for analysing the search behaviour of visitors, i.e. "how they navigate rather than what they retrieve". This idea is further analysed in [2, 3], where concept hierarchies as the basic method of grouping Web pages together. STRATDYN, is the add-on module that extends WUM's ([34, 35, 36]) capabilities by identifying the differences between navigation patterns, and exploiting the site's semantics in the visualization of the results. The accessed pages or paths are abstracted, since Web pages are treated as instances of a higher-level concept, based on page content, or by the kind of service requested. This work focuses mainly on the creation of navigational patterns rather than recommendations and the use of conceptual hierarchies is, as already mentioned, service (and not usage)-based.

Concept hierarchies are used from Parent et al. [27] in a different recommendation process. They present ARCH, an agent for assisting users in the query formulation process. The initial search query of the user is semi-automatically modified based on their interaction with a keyword-based concept hierarchy. The system takes as input the initial set of keywords the user includes in their query and displays the most appropriate portions of the hierarchy. The user then selects the most relevant categories and deselects the irrelevant ones. Therefore, a new query is formulated. This query vector is matched to the user's profile vector, which is created using heuristics based on their past browsing behavior, and the final query vector is formulated. Finally, the most relevant documents (which are represented by term-vectors and pre-classified under the hierarchy categories' nodes), are returned to the end user.

3. SEWeP ARCHITECTURE

In this section we present the architecture of SEWeP, a Web personalization system that integrates site semantics and a taxonomy with usage data. The need for such a system and the way it enhances the results of the Web personalization process is depicted using a running example.

3.1 Motivating Example

In order to demonstrate the need for broadening the recommendation set by integrating site semantics in the web personalization process, we introduce an example based on the experiments we performed on usage and content data collected from the Web site of our research group and use it throughout this section to illustrate the way our system addresses this problem.

For our experiments we used log files recorded in the Web server of our Web site, http://www.db-net.aueb.gr. This is the site of DB-NET research group in the Athens University of Economics and Business. The site's contents cover a variety of academic-related topics, such as lectures notes and tutorials, course descriptions, various research areas, research projects and personal home pages. We used as input data set the 90 log files recorded during a period of 3 months (1/3/02-31/5/02). A typical daily log of this Web server includes more than 1500 hits, therefore the total number of distinct hits is in the order of 10^5. In our experiments, during the data preparation phase we first eliminated all the hits that where redirected, or caused (client or server) error (i.e. hits with status code 4xx, 5xx). We also removed the records that corresponded to non-textual accesses (i.e. images, scripts, multimedia files etc). After this procedure, the "cleaned" log files that were used as input to the rest of the process included up to 10^4 hits to 131 distinct Web pages. Table 1 lists a fraction of the Web site's URIs, along with a description of their content, as defined by an expert.

We first apply association rules mining to the Web logs in order to extract interesting patterns. These patterns will consist the recommendation basis every time a user navigates through the site. Let's assume that a user is currently browsing our Web site, and the higher-ranked rules that correspond to his behavior (the rules are in the form A→B meaning that a user U that visited page A also visited page B) are:

R1: /courses.html & /courses/filesdb/index.html →
/courses/filesdb/b-trees2002.htm

R2: /research.html & /people/michalis.html →
/michalis/phds_new.html

Therefore, assuming that the user's navigational behavior matched the left-hand side of the rules, the system will recommend the following set of ULRs:

/courses/filesdb/b-trees2002.htm

/michalis/phds_new.html

Table 1: URIs of http://www.db-net.aueb.gr [1]

URI	Description
/courses.html	General information about courses offered
/courses/datamining/index.html	Data Mining course
/courses/filesdb/index.html	Files and Databases course index page
/courses/filesdb/btree.htm	Tutorial on B-Trees
/courses/filesdb/b-trees2002.htm	Coursework on B-Trees
/courses/filesdb/source/index.html	B-trees source code
/demosdm.htm	Data Mining demos
/index.htm	DB-NET Home Page
/michalis/phds_new.html	Announced PhD positions
/michalis/publications.html	List of publications of M.Vazirgiannis
/michalis/res_plan.html	Research Work and Plan of M. Vazirgiannis
/people/michalis.html	Home page of professor M. Vazirgiannis
/projects.htm	DB-NET Projects
/research.htm	DB-NET Research Interests

Based on the contents of the URIs included in Table 1, we notice that there exist more than one Web pages that might be of interest for the user, according to his navigational behavior, such as */courses/filesdb/btree.html* and *michalis/publications.html.* However, these URIs are not included in the recommendation list. This usually occurs when a Web page is new, therefore not yet included (or included with low frequency) in the web logs, or when it appears in rules with low confidence in the recommendation basis.

3.2 System Architecture

Motivated by the previous example, we observe that if a personalization system relies solely on usage-based results, then valuable information conceptually related to what is finally recommended may be missed. To tackle this problem we designed and developed a Web personalization system that is based on semantic enhancement of the Web usage logs and the related Web content. We call this system SEWeP (Semantic Enhancement for Web Personalization). The block diagram representing SeWeP's

---

[1] Since the data collection time, the web site has changed regarding its content, structure and presentation. However, a copy of the old version is still kept and may be found under http://www.db-net.aueb.gr/Olddbnet/.

architecture appears in Figure 1. The innovative feature of this system is the creation of C-logs (concept-logs) from the original Web logs and their use for extraction of usage patterns. C-Logs is an extended form of the Web server logs. Each record of the Web server logs is enhanced with keywords (from a taxonomy) representing the semantics of the respective URI. Data mining algorithms are then applied to this enriched version of Web logs, resulting in a set of recommendations that include thematic categories, except for recommendations including URIs. The Web documents are clustered based on the taxonomy categories; therefore the recommended categories are further expanded to contain the documents that fall under them.

The extraction of the keywords that characterize a Web page is performed using a combination of IR techniques. These keywords are mapped to the categories of a predefined domain-specific taxonomy through the use of a thesaurus. C-logs are processed in the same way as the Web server logs, through the use of statistical and data mining techniques, such as association rules, clustering or sequential pattern discovery. The outcome of this phase is a set of rules/patterns consisting of categories as well as URIs. Additionally, the semantically annotated Web site content is processed in order to be organized in coherent clusters. These clusters are then used in order to expand the set of recommendations provided to the end user. The functional architecture of SEWeP is described in more detail in the subsequent sections.
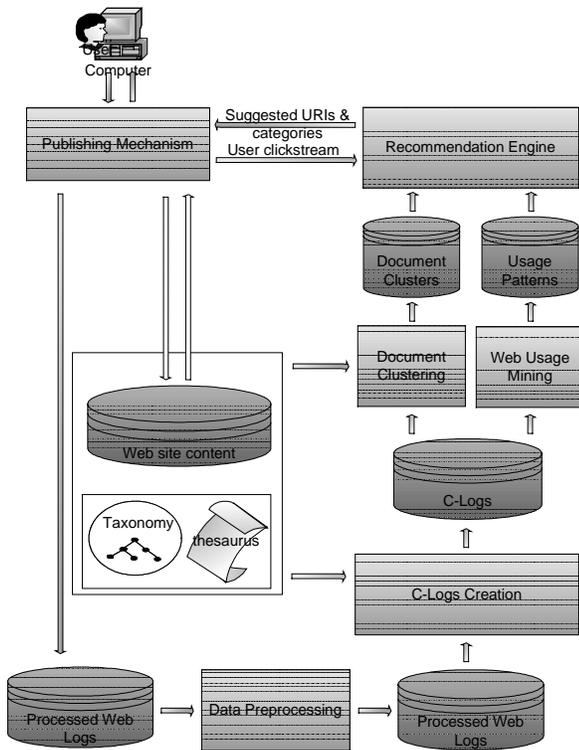

Figure 1: System architecture

3.3 C-logs Creation

The Web server logs are first preprocessed. After data preparation, session identification is performed. Since no cookies or registration data are available, the sessionizing is performed using heuristics based on IP and time thresholds, assuming that consecutive accesses from the same host during a time interval come from the same user. An extensive overview of the methods that may be employed in the data preprocessing phase is given in [11]. The preprocessed data are used as input for the C-logs' creation process.

The C-Logs creation process involves two distinct sub-processes: content classification and log transformation. The content classification process is performed once for every content object of the Web site (in most cases this is a Web page, however in the case of portals, a Web page may consist of several content objects)[2] and is repeated only when the content is altered or new content is added in the Web site. The log transformation process is performed whenever a Web log should be transformed to C-Log format. The whole process is described in Figure 2.
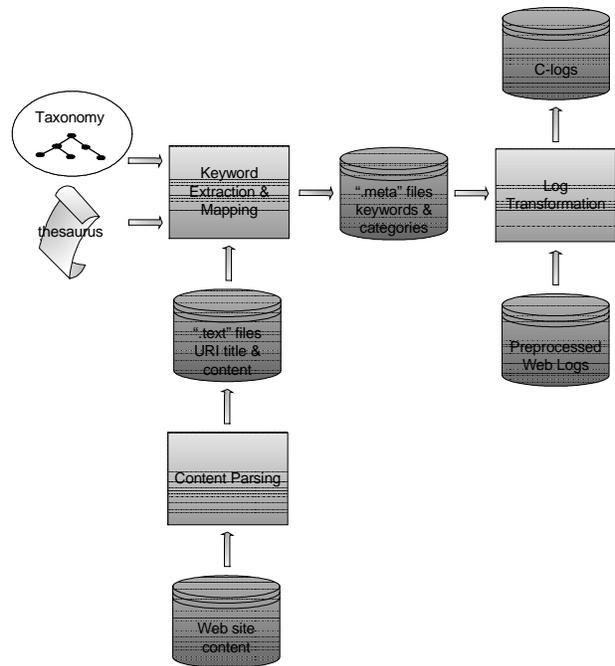

Figure 2: C-logs creation

3.3.1 Content Classification
3.3.1.1 Keyword Extraction

In order to extract keywords characterizing each Web page, the Web content is isolated from structure data (HTML and XML tags, meta-tags etc.). Therefore, in the first phase each Web page is parsed in order to separate its content for further processing. We should point out that SEWeP handles multiple document formats available in Web pages (i.e. .html, .doc, php, ppt, pdf, flash etc). At the end of this phase, a folder is created containing files corresponding to every URI of the Web site.

---

[2] For simplicity, we will refer to Web pages or URIs and not content objects in the rest of this paper.

There exists a wealth of methods for representing Web documents, most of which have been proposed for supporting similarity search in the Web. All of them are based on the extraction of keywords that characterize every document. The most straightforward approach is to perform text mining in the document itself, following standard IR techniques. However, this approach proves insufficient for the Web content, since it relies solely on the information included in the document ignoring semantics arising from the connectivity features of the Web [5, 9]. It is difficult to extract keywords from Web documents that contain images, programs etc. Additionally, many Web pages do not include words that are the most descriptive ones for their content (for example rarely a portal Web site includes the word "portal" in its home page). Therefore, in many approaches information contained in the links that point to the document and the text near them is used for characterizing a Web document. Chakrabarti et al. define this as the "anchor-window" [8]. The assumption made is that the text around the link to a page is descriptive of its contents. This approach overcomes the problems of the content-based approach, since it takes into consideration the way others characterize a specific Web page.

A further consideration to be made is the term weighting phase, when the extracted keywords are given weights in order to use the most important ones. Term weighting, extensively used in the vector space model for document clustering, is carried out using several methods, such as raw term frequency, or algorithms belonging to the Tf*Idf family [33]. Raw term frequency is based on the term statistics within a document and is the simpler way of assigning weights to terms. Tf*Idf is a method used for collections of documents, i.e. documents that have similar content. In the case of a Web site however, this assumption is not always true since a Web site may contain documents that refer to different thematic categories (especially in the case of Web portals).

For our system implementation, we use a combination of the aforementioned techniques, and a method based on page links. More specifically, the keywords that characterize a Web page p are extracted using:

1. raw term frequency of p

2. raw term frequency of the Web pages that are pointed by p (outlinks)

3. raw term frequency of a selected fraction (anchor-window) of the most important Web pages that point to p (inlinks)

The first and third methods are the ones described before. However, based on the assumption that in most Web pages the authors include links (to other pages) for topics that are of importance/interest in the page's context, we also use the second method for extracting a set of keywords.

As already mentioned, the text of every Web document is isolated from structure data and stored in a file named after the relevant URI. Therefore, in order to extract the most frequent terms for the first method, we first removed all the non-significant words (document indexing) using an appropriate stop-words list. This list includes very common words, such as pronouns, articles, adjectives, adverbs, and prepositions, both in English and in Greek, since the Web site contains content written in both languages. We keep the n most frequent words from the remaining ones. A similar method is used for every Web page that is pointed by page p. The pages are visited, indexed and the most frequent words are extracted. These words are then aggregated and the n most frequent are also chosen as representatives.

As far as the third method is concerned, after experimentation we decided to use the first 20 Web pages pointing to p^3. A decision that had to be made concerned the anchor-window's width. Chakrabarti et al. use a window of 50 bytes on either side of the link [8]. Haveliwala et al. [18] after experimenting with different window widths have reported that large fixed anchor-windows give the best results. They carry on with their experiments using a window of 32 words on each side of the link, (approximately 150 bytes). In our experiments, we used a window of 100 bytes, adopting the outcomes of [17]. Again, the n most frequent words where finally selected, resulting in a total (maximum) number of $3n$ keywords that characterize every Web page. For more details on mapping between keywords and categories see [17].

At this point we should mention that we had to address the bilingualism problem, since the Web site contains pages written both in Greek and English. As a result, the extracted keywords form a mixed set of words in both languages. Therefore, an intermediate translation step is needed. It should be stressed that all words in Greek language can be inflected. Therefore after the keyword extraction process they should be first transformed to the nominative, to be subsequently translated to English. For this purpose a Greek-English dictionary is used in combination with stemming algorithms for the Greek language. The details of this process are out of the scope of this paper, therefore won't be discussed further here. For more details concerning the translation process, please refer to [14].

3.3.1.2 Keyword-Category Mapping

After the aforementioned process, the most frequent keywords that are extracted using these three methods are selected as representatives of its content. However, for a number of reasons that were explained earlier (uniformity, clustering, similarity search), the keywords that were extracted in the previous stage should be mapped to the concepts of the taxonomy. This mapping is performed by using a thesaurus and a domain-specific taxonomy. If the keyword belongs to the taxonomy, then it is included as it is. Otherwise, the system finds the "closest" category word to the keyword through the mechanisms provided by the thesaurus.

In our system implementation, we defined the domain-specific taxonomy in XML. We should stress here that the selection of the taxomomy influences the outcomes of the mapping process. For this purpose, it should be semantically relevant to the content to be processed. There exist few publicly available taxonomies, either general, or area-specific. Since no appropriate domain-specific taxonomy could be found, the one used in our experiments was created manually by a domain-expert (the Web administrator). A fraction of this taxonomy is included in Figure 3. We use WordNet [1, 39] as thesaurus^4. In order to find the

---

[3] For this purpose we used Google's backwards link service (http://www.google.com/advanced_search)

[4] WordNet is a lexical database containing English nouns, verbs, adjectives and adverbs organized into synonym sets, each representing one underlying lexical concept. It provides

closest term in the taxonomy to represent each keyword, we compute the Wu & Palmer similarity [40] between all senses of the keyword and all senses of all the taxonomy terms. We select the category that gives the maximum Wu & Palmer similarity for each keyword.

At the end of this stage, each URI is characterized by a set of categories that are part of this taxonomy. These categories and the keywords of each content URI are internally stored in a relevant meta-file. This procedure is performed offline once, and should only be repeated if the content of a Web page changes or if new Web pages are added in the Web site.

Based on the example presented before, the categories characterizing the URIs included in Table 1 are presented in detail in Table 2.

Table 2: Categories characterizing www.db-net.aueb.gr Web pages

URI	Categories
/courses.html	database, lecture, data, knowledge, room, record, datamining
/courses/datamining/index.html	data, mining, editor, clustering, classification, link
/courses/filesdb/index.html	room time, hashing, lab
/courses/filesdb/btree.htm	node, leaf, b-tree, key, deletion, algorithm, example, insertion, tree
/courses/filesdb/b-trees2002.htm	b-tree, record, insertion, deletion
/courses/filesdb/source/index.html	data, b-tree, example, function
/demosdm.htm	data, association, mining, rules, research, page
/index.htm	research, environment, data department, athens, thesis, aueb,
/michalis/phds_new.html	system, mining, multimedia, interactive, phd
/michalis/publications.html	vazirgiannis, multimedia, journal, interactive, publication, research, proceedings
/michalis/res_plan.html	multimedia, spatiotemporal, vazirgiannis, interactive, tool
/people/michalis.html	athens, university, greece, research, system, qualification, vazirgiannis
/projects.htm	database, data, system, project, research, medicine, dbglobe
/research.htm	data, clustering, classification, database, research, vrshop

3.3.2 Log Transformation

Since there exists a meta-file for every URI of the Web server, the Web log records may be enriched with semantics. Every record is enriched with two extra fields, including every URI's relevant keywords and categories. The final form of the C-logs will therefore resemble to that of the Web logs and may be further processed in the same way as Web logs.

---

mechanisms for finding synonyms, hypernyms, hyponyms, etc. for a given word.

```
<db-netonto>
    …
    <aueb>
      <department>
        <informatics>
          <db-net>
            …
            <research>
              …
              <course>
                …
                <database>
                  …
                  <algorithm>
                    …
                    <quicksort/>
                    <b-tree/>
                    <children/>
                    <search/>
                    <deletion/>
                    <insertion/>
                  </algorithm>
                </database>
                <datamining/>
              </course>
              <project>
               …
              </project>
              <mining>
                …
                <knowledge/>
                <clustering/>
                <uncertainty/>
                <classification/>
                <association/>
              </mining>
              <multimedia>
                …
              </multimedia>
              <spatiotemporal/>
            </research>
          </db-net>
        </informatics>
      </department>
    </aueb>
</db-netonto>
```

Figure 3: Fraction of the DB-NET Web site taxonomy

3.4 Document Clustering

The content of the Web site is now semantically annotated with terms belonging to a taxonomy. Our purpose is to expand the recommendation set suggested to the end user, taking into account these content-bearing semantics. In order to accomplish this, the content should be classified. This is achieved by clustering the documents based on the similarity between the category terms that characterize them. This stands for documents that are not "structurally" close (i.e. under the same path) as well, therefore the clusters that are created capture semantic relationships that may not be obvious at first sight. The clustering algorithm that we used is an extension of DBSCAN [16], a density based algorithm used for categorical data. This algorithm doesn't need an a priori specification of the number of clusters (unlike K-Means), and is relatively efficient (compared to COBWEB). The original algorithm is used for clustering points of a metric space, so we used a modified version that employs a similarity measure in the clustering process, based on semantic proximity between sets of terms of an ontology. Again, we used a generalization of the Wu & Palmer similarity measure [40]. For more information regarding the clustering process, please refer to [17].

After applying this clustering algorithm to the documents contained in the DB-NET Web site, the URIs included in Table 2 are classified into three clusters (note that the whole set of the 131 URIs was clustered into 12 clusters, however we only include here the ones that contain the URIs included in Table 2):

C1: {/courses.html,
/courses/filesdb/index.html,
/courses/filesdb/btree.htm,
/courses/filesdb/b-trees2002.htm,
/courses/filesdb/source/index.html}

C2: { /demosdm.htm,
/projects.htm,
/research.htm}

C3: {/people/michalis.html,
/michalis/phds_new.html,
/michalis/publications.html,
/michalis/res_plan.html}

3.5 Recommendation Engine

The C-logs are used as input to data mining algorithms. The output consists of patterns representing the users' navigational behavior in the form of clusters or association rules. As in the case of Web usage mining and personalization using Web logs, this set of patterns is then used as the recommendation basis for each user or group of users. In SEWeP, instead of simply extracting a set of rules including URIs, we output a set of rules including categories that are recommended to the user. The left-hand sides of both sets of rules can be expanded to contain URIs that belong to the same thematic cluster. This approach takes into account the content of the Web site except for its usage and broadens the candidate recommendation set, since instead of the set of URIs that are directly derived, the system also recommends a set of URIs that are also characterized by the thematic categories that seem to be of interest for the user. This process broadens the recommendation set including URIs that wouldn't be proposed otherwise. As mentioned before, a relevant page can be exempted from the recommendation set if for example it weren't visited before.

Let's assume that the recommendation set created after
(a) association rule mining[5] in the Web logs of http://www.db-net.aueb.gr, and
(b) the matching of the results against the user's current navigational behavior,
includes the following rules:

R1: /courses.html & /courses/filesdb/index.html →
/courses/filesdb/b-trees2002.htm

R2: /courses.html & /research.html →
/projects.html

R3: /research.html & /people/michalis.html →
/michalis/phds_new.html

The URIs in the right hand side of the rules correspond to a set of categories as can be seen in Table 2, i.e. *{b-tree, record, insertion, deletion}*, *{database, data, system, project, research, medicine, dbglobe}* and *{system, mining, multimedia, interactive, phd}* respectively. As it was mentioned before, every document is already assigned to a relevant cluster, based on the similarity between sets of categories characterizing this and the rest of the documents. Therefore, based on the analysis above, the initial recommendation set {*/courses/filesdb/ btree2_files/slide0002.htm,*

/research.htm, /michalis/res_plan.html} is expanded to contain the URIs in clusters *{C1, C2, C3}* respectively.

3.6 Motivating Example Revisited: A Broader Set of Recommendations

The Web personalization process concludes by recommending to the user a set of URIs. This recommendation can be in the form of links that will be dynamically inserted in the Web page the user is visiting.

Therefore, the final candidate recommendation set consists of:

1. The Web pages derived from the initial association rules:

   /courses/filesdb/b-trees2002.htm

   /projects.html

   /michalis/phds_new.html

2. The "similar" Web pages to the ones initially derived, i.e. the ones belonging in the same cluster:

   /courses.html
   /courses/filesdb/index.html
   /courses/filesdb/btree.htm
   /courses/filesdb/b-trees2002.htm
   /courses/filesdb/source/index.html

   /demosdm.htm
   /projects.htm
   /research.htm

   /people/michalis.html
   /michalis/phds_new.html
   /michalis/publications.html
   /michalis/res_plan.html

We should stress at this point that the system chooses and recommends the most similar among the candidate documents (based on the similarity measure).

4. EVALUATION OF THE RESULTS

The fact that the process of semantically annotating Web content using terms derived from a domain-specific taxonomy prior to the recommendation process enhances the results of Web personalization is intuitive.

Experimental setup: We chose the Web logs of the DB-NET web site, including 131 web pages, collected for a 3 months period. The total web log size was 10^4 hits (order of magnitude). We applied the processes of the SEWeP system to the Web logs as described in previous sections extracting thus 500 association rules and for each page we stored 1-15 categories relevant to the page's content. Then we applied clustering and we grouped the web site documents into 12 clusters.

We chose several paths followed by the Web site visitors, analyzed the paths and found the best recommendations using the association rules usage (further called *original* recommendations). We enriched the recommendations' set by adding documents that bear similar semantics (these *additional* recommendations were the result of the semantic characterization and clustering process bescribed in previous sections). Then we created, for each path, a recommendation set consisting of *original* and *additional* recommendations in equal proportions. The recommendations

were mixed so that the users do not distinguish between original and additional ones.

We selected 4 different paths and for each of them we proposed a respective recommendation set. The recommendations were ranked by the users according to their relevance in the range 1-4 (1, the most relevant). We used 9 blind testers to evaluate these groups of recommendations[6]. Therefore, the recommendation sets with lower sums are considered better. For every group of recommendations, we selected the same number of original and additional recommendations and compared their average rankings, after normalizing them to a range [0-1], 0 (1) indicating highest (lowest) relevance. The comparative results for the original and the additional recommendations are presented in Table 3.

In the first two recommendation sets (A, B), the average ranking of the recommendations is equal. This means that users evaluated the additional recommendations to be as valuable as the original ones. This occurred because the recommendation sets in both cases included additional recommendations that are of related content with the original ones, but where not visited very often, therefore where ranked low in the original association rules. In the third group, C, there is a slight advantage of the original recommendations. This happened because the path included visits to top-level pages, a situation the original association rules handle successfully. Nevertheless, in the fourth group, D, there is a significant difference in favor of the additional recommendations. This occurred because one of the additional recommendations, not included in the original because the document was new, therefore not included in the association rules, was very relevant to the users' interests. The overall recommendation relevance indicates that the blind testers found more relevant the additional recommendations (*0.53*) than he original ones (*0.62*).

Table 3: Recommendation sets evaluation based on users' blind testing (0: very relevant, 1: irrelevant)

Recommendation Set	Original recommendations relevance	Additional recommendations relevance
A	0,56	0.56
B	0,5	0,5
C	0,47	0,67
D	0,94	0,39
Total Average Relevance:	**0,62**	**0,53**

The evaluation of the results is based on the following assumption: SEWeP enhances the personalization process if the additional recommendations it provides to the users based on semantic similarity between documents are ranked as high (or even higher) as the ones that would be recommended initially. In such a case, the end user would receive a more cohesive and precise set of recommendations. The results presented in Table 3 verify this assumption.

---

[6] The 4 paths and relevant recommendations are included in Appendix A

5. CONCLUSIONS – FUTURE WORK

In this paper, we presented the architecture of SEWeP, a Web personalization system that integrates the Web usage mining process with site semantics in order to enrich the set of recommendations that are provided to the end user. The innovative feature of this architecture is the introduction of C-logs, an extension of the Web usage logs that encapsulate content semantics. The semantic annotation of the content is performed using a conceptual hierarchy (taxonomy). This categorization enables clustering and further ranking of the Web documents. The application of web usage mining methods to C-logs results in a broader set of recommendations, containing, apart from the set of the original URIs, the semantic categories related to them, and the rest URIs related to those categories. We have implemented the C-Logs creation and the document clustering modules and have demonstrated the process of semantically annotating Web content using a running example. Additionally, we handled the bilingualism problem, occurring because some Web pages were written in Greek. Finally, we performed a set of blind tests and proved that the recommendations derived using SEWeP enhance the personalization process.

We still aim to involve user profile data in the personalization process, by taking into consideration preferences of the users regarding one or more taxonomy categories, in order to further filter the candidate recommendation set. We plan on experimenting with the variables of the document characterization phase, by changing the anchor-window width, adding weights to the terms or using different taxonomy as input. We aim at improving the similarity measure for comparing sets of terms. Finally, we intend to add the association rules mining algorithm in SEWeP and to implement the recommendation engine.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] R. Al-Halami, R. Berwick et. al., C. Fellbaum (editor), WordNet, an electronic lexical database, Bradford Books, May 1998, ISBN 0-262-06197-X

[2] B. Berendt, Web usage mining, site semantics, and the support of navigation, in Proc. of the Web Mining for E-Commerce - Challenges and Opportunities Workshop (WEBKDD'00), Boston, MA, August 2000

[3] B. Berendt, Understanding Web usage at different levels of abstraction: coarsening and visualizing sequences, in Proc. of the Mining Log Data Across All Customer TouchPoints Workshop (WEBKDD'01), San Francisco, CA, August 2001

[4] B. Berednt, M. Spiliopoulou, Analysis of navigation behaviour in web sites integrating multiple information systems, The VLDB Journal (2000) 9, 56-75

[5] S. Brin, L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proc. of WWW7, 1998

[6] A.G. Buchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, and J.G. Hughes, Navigation pattern discovery from Internet data, in Proc. of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99), San Diego, CA, August 1999

[7] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan, Using taxonomy, discriminants, and signatures for navigation in text databases, in Proc. of the 23rd VLDB Conference, Athens, Greece, 1997

[8] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer (1999) Vol.32 No.6

[9] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, J. Kleinberg, Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, in Proc. of WWW7, 1998

[10] F. Coenen, G. Swinnen, K. Vanhoof, G. Wets, A Framework for Self Adaptive Websites: Tactical versus Strategic Changes, in Proc. of Web Mining for E-Commerce – Challenges and Opportunities Workshop (WEBKDD'00), Boston, MA, August 2000

[11] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide Web browsing patterns, Knowledge and Information Systems, February 1999/Vol.1, No. 1

[12] R. Cooley, P.N. Tan, J. Srivastava, WebSIFT: The Web Site Information Filter System, in Proc. of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99), San Diego, CA, August 1999

[13] H. Dai, B. Mobasher, Using Ontologies to Discover Domain-Level Web Usage Profiles, in Proc. of the 2nd Workshop on Semantic Web Mining, at PKDD'02, Helsinki, Finland, August 2002

[14] M. Eirinaki, H. Labos, M. Vazirgiannis, Archiving the Greek Web, Technical Report (2003), http://www.db-net.aueb.gr

[15] M. Eirinaki, M. Vazirgiannis, Web Mining for Web Personalization, ACM Transactions on Internet Technology (TOIT), February 2003/ Vol.3, No.1, 1-27

[16] M. Ester, H.P. Kriegel, J. Sander, M. Wimmer and X. Xu, Incremental Clustering for Mining in a Data Warehousing Environment, in Proc. of the 24th VLDB Conference (1998).

[17] M. Halkidi, B. Nguyen, I. Varlamis, M. Vazirgiannis, THESUS: Organizing Web Document Collections Based on Link Semantics, to appear in VLDB Journal, special issue on Semantic Web (2003)

[18] T.H. Haveliwala, A. Gionis, D. Klein, P. Indyk, Evaluating Strategies for Similarity Search on the Web, in Proc. of WWW11, Hawaii, USA, May 2002

[19] F. Masseglia, P. Poncelet, R. Cicchetti, WebTool: An Integrated Framework for Data Mining, in Proc. of the 9th International Conference on Database and Expert Systems Applications (DEXA'99), Florence, Italy, August 1999, 892-901

[20] F. Masseglia, P. Poncelet, M. Teisseire, Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure, in ACM SIGWEB Letters, October 1999/Vol. 8, No. 3, 13-19

[21] F. Masseglia, P. Poncelet, M. Teisseire, Web Usage Mining: How to Efficiently Manage New Transactions and New Customers, in Proc. of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00), Lyon, France, September 2000

[22] B. Mobasher, R. Cooley, J. Srivastava, Creating Adaptive Web Sites Through Usage-Based Clustering of URLs, in Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), November 1999

[23] B. Mobasher, R. Cooley, J. Srivastava, Automatic Personalization Based on Web Usage Mining, Communications of the ACM, August 2000/Vol. 43, No. 8, 142-151

[24] B. Mobasher, H. Dai, T. Luo, Y. Sung, J. Zhu, Discovery of Aggregate Usage Profiles for Web Personalization, in Proc. of the Web Mining for E-Commerce Workshop (WEBKDD'00), Boston, MA, August 2000

[25] B. Mobasher, H. Dai, T. Luo, Y. Sung, J. Zhu, Integrating Web Usage and Content Mining for More Effective Personalization, in Proc. of the International Conference on E-Commerce and Web Technologies (ECWeb2000), Greenwich, UK, September 2000

[26] M.D. Mulvenna, S.S. Anand, A.G. Buchner, Personalization on the Net using Web Mining, Communications of the ACM, August 2000/Vol. 43, No. 8, 123-125

[27] S. Parent, B. Mobasher, S. Lytinen, An Adaptive Agent for Web Exploration Based on Concept Hierarchies, in Proc. of the 9th International Conference on Human Computer Interaction (HCI), August 2001

[28] M. Perkowitz, O. Etzioni, Adaptive Web Sites: An AI Challenge, in Proc. of the Fifteenth International Joint Conference on Artificial Intelligence, Nagoya, Japan, 1997

[29] M. Perkowitz, O. Etzioni, Adaptive Web Sites: Automatically Synthesizing Web Pages, in Proc. of the 15th National Conference on Artificial Intelligence, Madison, WI, July 1998

[30] M. Perkowitz, O. Etzioni, Adaptive Web Sites: Conceptual Cluster Mining, in Proc. of the 16th International Joint Conference on Artical Intelligence (IJCAI99), Stockholm, Sweden, 1999

[31] M. Perkowitz, O. Etzioni, Towards Adaptive Web Sites: Conceptual Framework and Case Study, in Proc. of WWW8, 1999

[32] M. Perkowitz, O. Etzioni, Adaptive Web Sites, Communications of the ACM, August 2000/Vol. 43, No. 8, 152-158

[33] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management (1998), Vol. 24, 513-523

[34] M. Spiliopoulou, Web Usage Mining for Web Site Evaluation, Communications of the ACM, August 2000/Vol. 43, No. 8, 127-134

[35] M. Spiliopoulou and L.C. Faulstich, WUM: A Web Utilization Miner, In International Workshop on the Web and Databases, Valencia, Spain, March 1998

[36] M. Spiliopoulou, L.C. Faulstich, and K. Wilkler, A data miner analyzing the navigational behaviour of Web users, in Proc. of the Workshop on Machine Learning in User Modelling of the ACAI99, Greece, July 1999

[37] R. Srikant, R. Agrawal, Mining Generalized Association Rules, in Proc. of 21st VLDB Conf., Zurich, Switzerland, September 1995

[38] J. Srivastava, R. Cooley, M. Deshpande, P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, January 2000/Vol. 1, No. 2, 12-23

[39] WordNet, A lexical database for the English language, http://www.cogsci.princeton.edu/~wn/

[40] Z. Wu, M. Palmer: Verb Semantics and Lexical Selection, 32nd Annual Meetings of the Associations for Computational Linguistics (1994), 133-138

[41] T.W. Yan, M. Jacobsen, H. Garcia-Mollina, U. Dayal, From User Access Patterns to Dynamic Hypertext Linking, In Proc. of WWW5, Paris, France, 1996

[42] O.R. Zaiane, M. Xin, J. Han, Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, in Proc. of Advances in Digital Libraries Conference (ADL'98), Santa Barbara, CA, April 1998

APPENDIX A – TEST PATHS AND CANDIDATE RECOMMENDATIONS

Below are included the 4 paths and relevant recommendations presented to the testers as part of the experiment described in Section 4. In order to distinguish between the *original* and the *additional* recommendations for the purpose of this paper, we present the latter ones in italics.

Path A

http://www.db-net.aueb.gr/Olddbnet/index.html ->

http://www.db-net.aueb.gr/Olddbnet/courses.html ->

http://www.db-net.aueb.gr/Olddbnet/courses/filesdb/index.html ->

http://www.db-net.aueb.gr/Olddbnet/courses/filesdb/sql2002.htm

Recommendation Set A

- *http://www.db-net.aueb.gr/Olddbnet/courses/filesdb/btree.htm*

- http://www.db-net.aueb.gr/Olddbnet/courses/filesdb/b-trees2002.htm

- *http://www.db-net.aueb.gr/Olddbnet/courses/filesdb/source/index.html*

- http://www.db-net.aueb.gr/Olddbnet/courses/filesdb/students/tables.htm

Path B
(http://www.db-net.aueb.gr/Olddbnet/courses/filesdb/sql2002.htm ->)

http://www.db-net.aueb.gr/Olddbnet/courses.html ->

http://www.db-net.aueb.gr/Olddbnet/index.html ->

http://www.db-net.aueb.gr/Olddbnet/research.html

Recommendation Set B

- *http://www.db-net.aueb.gr/Olddbnet/courses/postgrdb/asilomar.html*

- http://www.db-net.aueb.gr/Olddbnet/courses/datamining/index.html

- *http://www.db-net.aueb.gr/Olddbnet/demosdm.htm*

- http://www.db-net.aueb.gr/Olddbnet/projects.html

Path C
(http://www.db-net.aueb.gr/Olddbnet/research.html ->)

http://www.db-net.aueb.gr/Olddbnet/index.html ->

http://www.db-net.aueb.gr/Olddbnet/projects.html

Recommendation Set C

- *http://www.db-net.aueb.gr/magda/research.htm*

- http://www.db-net.aueb.gr/mhalk/papers/QUnc_book.html

- *http://www.db-net.aueb.gr/mhalk/Publ_maria.htm*

- http://www.db-net.aueb.gr/Olddbnet/jobs.htm

Path D
(http://www.db-net.aueb.gr/Olddbnet/projects.html ->)

http://www.db-net.aueb.gr/Olddbnet/index.html ->

http://www.db-net.aueb.gr/Olddbnet/people.html ->

http://www.db-net.aueb.gr/Olddbnet/people/michalis.html

Recommendation Set D

- http://www.db-net.aueb.gr/michalis/phds_new.html

- *http://www.db-net.aueb.gr/michalis/publications.html*

- *http://www.db-net.aueb.gr/michalis/res_plan.html*

- http://www.db-net.aueb.gr/michalis/thesis.html