

Introducing Semantics in Web Personalization: The Role of Ontologies

Magdalini Eirinaki, Dimitrios Mavroeidis, George Tsatsaronis,
and Michalis Vazirgiannis

Athens University of Economics and Business, Dept. of Informatics,
Athens, Greece
{eirinaki, dmavr, gbt, mvazirg}@aueb.gr

Abstract. Web personalization is the process of customizing a web site to the needs of each specific user or set of users. Personalization of a web site may be performed by the provision of recommendations to the users, highlighting/adding links, creation of index pages, etc. The web personalization systems are mainly based on the exploitation of the navigational patterns of the web site's visitors. When a personalization system relies solely on usage-based results, however, valuable information conceptually related to what is finally recommended may be missed. The exploitation of the web pages' semantics can considerably improve the results of web usage mining and personalization, since it provides a more abstract yet uniform and both machine and human understandable way of processing and analyzing the usage data. The underlying idea is to integrate usage data with content semantics, expressed in ontology terms, in order to produce semantically enhanced navigational patterns that can subsequently be used for producing valuable recommendations. In this paper we propose a semantic web personalization system, focusing on word sense disambiguation techniques which can be applied in order to semantically annotate the web site's content.

1 Introduction

During the past few years the World Wide Web has emerged to become the biggest and most popular way of communication and information dissemination. Every day, the Web grows by roughly a million electronic pages, adding to the hundreds of millions pages already on-line. WWW serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content, software and personal weblogs (blogs). Because of its rapid and chaotic growth, the resulting network of information lacks of organization and structure. Users often feel disoriented and get lost in that information overload that continues to expand. On the other hand, the e-business sector is rapidly evolving and the need for Web market places that anticipate the needs of the customers is more than ever evident. Therefore, an ultimate need nowadays is that of predicting the user needs in order to improve the usability and user retention of a Web site.

In brief, web personalization can be defined as any action that adapts the information or services provided by a web site to an individual user, or a set of users, based

on knowledge acquired by their *navigational behavior*, recorded in the web site's logs. This information is often combined with the *content* and the *structure* of the web site as well as the *user's interests/preferences*, if they are available. Using the four aforementioned sources of information as input to pattern discovery techniques, the system tailors the provided content to the needs of each visitor of the web site. The personalization process can result in the dynamic generation of recommendations, the creation of index pages, the highlighting of existing hyperlinks, the publishing of targeted advertisements or emails, etc. In this paper we focus on personalization systems that aim at providing personalized recommendations to the web site's visitors.

The problem of providing recommendations to the visitors of a web site has received a significant amount of attention in the related literature. Most of the earlier research efforts in Web personalization correspond to the evolution of extensive research in Web usage mining [3, 9, 41]. Pure usage-based personalization, however, presents certain shortcomings, for instance when there is not enough usage data available in order to extract patterns related to certain navigational actions, or when the web site's content changes and new pages are added but are not yet included in the web logs.

Motivated by the fact that the users' navigation is extremely semantically-driven, in other words the users' visits usually aim at finding information concerning a particular subject, we claim that the underlying content semantics should be a dominant factor in the process of web personalization. There have been a number of research studies that integrate the web site's content in order to enhance the web personalization process [18, 22, 30, 37]. Most of these efforts characterize web content by extracting features from the web pages. Usually these features are keywords subsequently used to retrieve similarly characterized content. The similarity between documents is usually based on exact matching between these terms. In this way, however, only a binary matching between documents is achieved, whereas no actual *semantic* similarity is taken into consideration.

The need for a more abstract representation that will enable a uniform and more flexible document matching process imposes the use of semantic web structures, such as ontologies¹ [6, 19]. By mapping the keywords to the concepts of an ontology, or topic hierarchy, the problem of binary matching can be surpassed through the use of the hierarchical relationships and/or the *semantic similarities* among the ontology terms, and therefore, the documents.

Several research studies proposed frameworks that express the users' navigational behavior in terms of an ontology and integrate this knowledge in semantic web sites [36], Markov model-based recommendation systems [2], or collaborative filtering systems [11, 33]. Overall, all the aforementioned approaches are based on the same intuition: enhance the web personalization process with content semantics, expressed using the terms of a domain-ontology. The extracted web content features are mapped to ontology terms and this abstraction enables the generalizations/specializations of the derived patterns and/or user profiles. In all proposed models, however, the ontology-term mapping process is performed manually or semi-automatically (needing the manual labeling of the training data set). Some approaches are based on collaborative filtering systems, which assume that some kind of user ratings are available, or on

¹ In this work we focus on the hierarchical part of an ontology. Therefore, in the rest of this work we use the terms *concept hierarchy*, *taxonomy* and *ontology* interchangeably.

semantic web sites, which assume that an existing underlying semantic annotation of the web content is available a priori. Finally, none of the aforementioned approaches fully exploits the underlying semantic similarities of terms belonging to an ontology, apart from the straightforward “is-a” or “parent-child” hierarchical relationships.

Since ontologies resemble the semantic networks underlying the word thesauri, the process of keyword mapping to ontology concepts can be related to thesaurus-based Word Sense Disambiguation (WSD). The analogy stems from the fact that both thesauri and ontologies contain a vast amount of semantic background information concerning the concepts they contain. The semantic information is usually expressed through semantic relations, such as “is-a” and “has-part” relations. Thesaurus-based WSD algorithms aim at exploiting such semantic relations for successfully mapping words to thesaurus concepts. Although the effectiveness of such methods for the semantic representation of documents had been an issue of controversy, recent thesaurus-based WSD algorithms have been shown to consistently improve the performance of classification and clustering tasks [7, 20, 29, 44].

In this paper we present a Semantic Web Personalization framework (further referred to as SEWeP) that integrates usage data with content semantics expressed in ontology terms in order to effectively generate useful recommendations. This framework is mainly based on the work presented in [14, 15, 38]. Similar to previously proposed approaches, the proposed personalization framework uses ontology terms to annotate the web content and the users’ navigational patterns. The key departure from earlier approaches, however, is that the proposed personalization framework employs fully automatic ontology mapping WSD-based techniques [19, 29], by exploiting the underlying semantic similarities between ontology terms.

In the Section that follows we present work related to thesaurus-based WSD algorithms and web personalization systems which use ontologies. We then discuss several measures for computing similarity between ontology terms in Section 3. In Section 4 we present in detail the proposed Semantic Web Personalization framework and we conclude in Section 5.

2 Related Work

In this Section we present a short review on thesaurus-based WSD algorithms. We also review the research studies which integrate content data in the web personalization process, focusing on those that employ ontologies in order to represent the web documents.

2.1 Word Sense Disambiguation for Ontologies

In this subsection we present a short review on WSD approaches that are based on utilizing semantic relations in a word thesauri. Since ontologies resemble the semantic networks underlying a word thesaurus, these methods can be naturally extended for mapping keywords to ontology concepts. In the subsequent paragraph, we use WSD terminology with regards to a given word w . The “sense” of a word w is the concept of the thesaurus assigned to w . The “context” of w , refers to its surrounding words in the text it belongs to, and depending on the method its definition can vary and may

include from a small window of surrounding words, like in the method of Lesk [27], to all the words occurring in the same text as w , like in the method proposed in [34].

Several WSD approaches take advantage of the fact that a thesaurus offers important vertical (is-a, has-part) and horizontal (synonym, antonym, coordinate terms) semantic relations. Sussna [43] has proposed an unsupervised WSD approach where the distance among the candidate senses of a noun, as well as the senses of the words in its context are taken into account, using a sliding window of noun words occurring in the text. The correct sense that disambiguates each noun is found through minimizing the distance between possible assignments of senses of neighboring nouns. In order to compute this distance, the author considers a semantic distance measure which utilizes the semantic relations expressing the hypernym/hyponym, meronym/holonym and synonym/antonym nature. In the work of Agirre and Rigau [1] the hypernym/hyponym relation is used again to form a measure of conceptual distance between senses, by measuring the shortest path between the possible senses of a noun to be disambiguated and the senses of its context words. Rigau et al. [40] combined previous WSD approaches and utilized the hypernym/hyponym and the domain semantic relation, along with other heuristics that make use of measuring word co-occurrence in senses' definitions and constructing semantic vectors, to form a new unsupervised WSD approach. Leacock et al. [25] have also used the hypernym/hyponym, the synonym and the coordinate terms semantic relationship (the latter expresses senses sharing the same immediate hypernym) existing in WordNet [16] to form the training material of their WSD algorithm. Mihalcea et al. [32] have used synonyms and hypernoms/hyponyms as well to generate the semantically connected senses for the words to be disambiguated. Their disambiguation takes place in an iterative manner, generating a set for the already disambiguated words and a set for ambiguous word, while utilizing possible semantic connections between the two sets. Montoyo et al. [31] also use hypernoms and hyponyms, along with their glosses, in order to combine knowledge-based and corpus-based WSD methods. Moreover, in [5, 17, 42] lexical chaining is used for word sense disambiguation, which is a process of connecting semantically related words (thus making use of hypernym/hyponym and other semantic relations) in order to create a set of chains that represent different threads of cohesion through a given text. Lexical chaining has also been validated in the area of text summarization [5, 42]. Finally, in [35], they use a variety of knowledge sources to automatically generate semantic graphs, which are essentially alternative conceptualizations for the lexical items to be disambiguated. For building these graphs they used various types of semantic relations, like meronymy/holonymy, hypernymy/hyponymy and synonymy.

In contrast to the approaches described above, the WSD algorithm proposed in [29] has been validated experimentally both in "pure" WSD (using WSD benchmark datasets), and in the document classification task. The fact that our approach has been shown to improve classification accuracy, constitutes a strong indication that it can be used effectively for enhancing semantics in document representation.

2.2 Using Content Semantics for Web Personalization

Several frameworks based on the claim that the incorporation of information related to the web site's content enhances the web mining and web personalization process

have been proposed prior [30, 37] or subsequent [18, 22, 23] to our original work [14, 15]. In this subsection we overview in detail the ones that are more similar to ours, in terms of using a domain-ontology to represent the web site's content for enhancing the web personalization results.

Dai and Mobasher [11] proposed a web personalization framework that characterizes the usage profiles of a collaborative filtering system using ontologies. These profiles are transformed to "domain-level" aggregate profiles by representing each page with a set of related ontology objects. In this work, the mapping of content features to ontology terms is assumed to be performed either manually, or using supervised learning methods. The defined ontology includes classes and their instances therefore the aggregation is performed by grouping together different instances that belong to the same class. The recommendations generated by the proposed collaborative system are in turn derived by binary matching the current user visit expressed as ontology instances to the derived domain-level aggregate profiles, and no semantic relations beyond hyperonymy/hyponymy are employed.

The idea of semantically enhancing the web logs using ontology concepts is independently described by Oberle et.al. [36]. This framework is based on a semantic web site built on an underlying ontology. This site contains both static and dynamic pages being generated out of the ontology. The authors present a general framework where data mining can then be performed on these semantic web logs to extract knowledge about groups of users, users' preferences, and rules. Since the proposed framework is built on a semantic web knowledge portal, the web content is inherently semantically-annotated exploiting the portal's inherent RDF annotations. The authors discuss how this framework can be extended using generalizations/specializations of the ontology terms, as well as for supporting the web personalization process, yet they mainly focus on web mining.

Acharyya and Ghosh [2] also propose a general personalization framework based on the conceptual modeling of the users' navigational behavior. The proposed methodology involves mapping each visited page to a topic or concept, imposing a tree hierarchy (taxonomy) on these topics, and then estimating the parameters of a semi-Markov process defined on this tree based on the observed user paths. In this Markov models-based work, the semantic characterization of the context is performed manually. Moreover, no semantic similarity measure is exploited for enhancing the prediction process, except for generalizations/specializations of the ontology terms.

Middleton et. al [33] explore the use of ontologies in the user profiling process within collaborative filtering systems. This work focuses on recommending academic research papers to academic staff of a University. The authors represent the acquired user profiles using terms of a research paper ontology (is-a hierarchy). Research papers are also classified using ontological classes. In this hybrid recommender system which is based on collaborative and content-based recommendation techniques, the content is characterized with ontology terms, using document classifiers (therefore a manual labeling of the training set is needed) and the ontology is again used for making generalizations/specializations of the user profiles.

Finally, Kearney and Anand [23] use an ontology to calculate the impact of different ontology concepts on the users navigational behavior (selection of items). In this work, they suggest that these impact values can be used to more accurately determine distance between different users as well as between user preferences and other items

on the web site, two basic operations carried out in content and collaborative filtering based recommendations. The similarity measure they employ is very similar to the Wu & Palmer similarity measure presented here. This work focuses on the way these ontological profiles are created, rather than evaluating their impact in the recommendation process, which remains opens for future work.

3 Similarity of Ontology Terms

As already mentioned, the proposed semantic web personalization framework exploits the expressive power of content semantics, that are represented by ontology terms. Using such a representation, the similarity between documents is deduced to the distance between terms that are part of a hierarchy. The need for such a similarity measure is encountered throughout the personalization process, namely during content characterization, keyword translation, document clustering and recommendations' generation.

There is an extensive bibliography addressing the issue of defining semantic distances and similarity measures based on semantic relations. A popular similarity measure for ontology concepts is proposed by Resnik [39]. The similarity between two ontology concepts is based on the "depth" of their least common ancestor, where the "depth" is measured using the information content. Formally the similarity measure is defined as: $RSsim(a,b) = \max_{c \in Supp(a,b)} IC(c)$, where $IC(c) = -\log P(c)$ is the information content of concept c and $Supp(a,b)$ is a set containing all ancestors (in the hierarchical structure) of a and b .

Jiang and Conrath [21] define a distance measure based on the path of two concepts to their least common ancestor. Their distance measure does not depend solely on the edge counting, since the information content is used for weighting the edges. Formally the Jiang and Conrath distance measure is defined as: $JCdis(a,b) = IC(a) + IC(b) - 2IC(lca(a,b))$, where $IC(c)$ is the information content of concept c and $lca(a,b)$ is the least common ancestor of a and b .

Leacock and Chodorow [24] define a similarity measure that is based on the shortest path that connects two concepts normalized by the maximum depth of the ontology. Their similarity measure is defined as: $LCsim(a,b) = -\log \frac{path_length(a,b)}{2D}$,

where $path_length$ is the length of the shortest path that connects the two concepts in the ontology and D denotes the maximum depth of the ontology.

Finally, Lin [28] has proposed a similarity measure, based on the Wu and Palmer similarity measure [48]. More precisely, they incorporated the information content in order to extend the flexibility of the similarity measure beyond edge counting, The Lin similarity measure is defined as: $LinSim(a,b) = \frac{2IC(lca(a,b))}{IC(a) + IC(b)}$, where $lca(a,b)$

defines the least common ancestor of concepts a and b .

The ontology similarity and distance measures described above are defined for pairs of concepts that reside in an ontology and cannot be directly used for evaluating the similarity between documents (a document contains a set of concepts). However, they can be used for evaluating the similarity between two documents either in the

context of distance measures for sets of objects, or in the context of the Generalized Vector Space Model (GVSM model) [29, 37].

In our approach, we adopt the Wu & Palmer similarity measure [48] for calculating the distance between terms that belong to a tree (hierarchy). Moreover, we use its generalization, proposed by Halkidi et.al. [19] to compute the similarity between sets of terms that belong to a concept hierarchy. Furthermore we utilize a recently proposed similarity measure for sets of ontology concepts that is based on the GVSM model proposed in [29]. We should stress that the choice of the similarity measure is orthogonal to the rest system functionality, as long as it serves for calculating the distance between hierarchically organized terms. The definitions of the three similarity measures are given in what follows. A more detailed description and theoretical proof can be found in the related publications.

3.1 Wu&Palmer Similarity Measure

Given a tree, and two nodes a, b of this tree, their similarity is computed as follows:

$$WPsim(a,b) = \frac{depth(a) + depth(b)}{2 * depth(c)} \quad (1)$$

where the node c is their deepest (in terms of tree depth) common ancestor.

3.2 THESUS Similarity Measure

Given an ontology T and two sets of weighted terms $\mathcal{A}=\{(w_i, k_i)\}$ and $\mathcal{B}=\{(v_i, h_i)\}$, with $w_i, v_i \in T$, their similarity is defined as:

$$THESim(\mathcal{A},\mathcal{B}) = \frac{1}{2} \left[\left(\frac{1}{K} \sum_{i=1}^{|\mathcal{A}|} \max_{j \in [1,|\mathcal{B}|]} (\lambda_{i,j} \times WPsim(k_i, h_j)) \right) + \left(\frac{1}{H} \sum_{i=1}^{|\mathcal{B}|} \max_{j \in [1,|\mathcal{A}|]} (\mu_{i,j} \times WPsim(h_i, k_j)) \right) \right] \quad (2)$$

where $\lambda_{i,j} = \frac{w_i + v_j}{2 \times \max(w_i, v_j)}$ and $K = \sum_{i=1}^{|\mathcal{A}|} \lambda_{i,x(i)}$, with

$$x(i) = x \mid \lambda_{i,x} \times WPsim(k_i, h_x) = \max_{j \in [1,|\mathcal{B}|]} (\lambda_{i,x} \times WPsim(k_i, h_j))$$

The theoretical and experimental justification of the effectiveness of the aforementioned similarity measure is included in [19].

3.3 GVSM Based Similarity Measure

The similarity between two documents d_1 and d_2 that contain ontology concepts is defined as

$$GVSMsim(d_1, d_2) = d_1 D D^T d_2^T \quad (3)$$

where the rows of matrix D contain the vector representations of the ontology concepts. For constructing the vector representations, we initially consider an index of all the ontology concepts. Then the vector representation of each concept has non-zero elements only at the dimensions that correspond to the concept's ancestors. For illustrative purposes, consider two ontology concepts $c_1=insurance_company$ and

$c_2=bank$, where both concepts have two ancestors in the ontology hierarchy, $c_3=financial_insitution$ and $c_4=institution$. Then, if we consider that the concept hierarchy contains only these four concepts, the index of all the ontology concepts will be (c_1, c_2, c_3, c_4) , and the vector representation of c_1 will be $(1,0,1,1)$, while the vector representation of c_2 will be $(0,1,1,1)$. In [29] we justify theoretically (Propositions 1, 2) and experimentally the effectiveness of the proposed similarity measure.

4 Ontology-Based Semantic Web Personalization

The users' navigation in a web site is usually content-oriented. The users often search for information or services concerning a particular topic. Therefore, the underlying content semantics should be a dominant factor in the process of web personalization. In this Section we present a web personalization framework that integrates content semantics with the users' navigational patterns, using ontologies to represent both the content and the usage of the web site. This framework is mainly based on the SEWeP personalization system, presented in [14, 15, 38]. To the best of our knowledge, it is the only web personalization framework where the content characterization process is performed using WSD-based methods [19, 29], fully exploiting the underlying semantic similarities of ontology terms.

4.1 SEWeP System Architecture

SEWeP uses a combination of web mining techniques to personalize a web site. In short, the web site's content is processed and characterized by a set of ontology terms (categories). The visitors' navigational behavior is also updated with this semantic knowledge to create an enhanced version of web logs, C-logs, as well as semantic document clusters. C-Logs are in turn mined to produce both a set of URI and category-based association rules. Finally, the recommendation engine uses these rules, along with the semantic document clusters in order to provide the final, semantically enhanced set of recommendations to the end user.

As illustrated in Figure 1, SEWeP consists of the following components:

- *Content Characterization*. This module takes as input the content of the web site as well as a domain-specific ontology and outputs the semantically annotated content to the modules that are responsible for creating the C-Logs and the semantic document clusters.
- *Semantic Document Clustering*. The semantically annotated pages created by the previous component are grouped into thematic clusters. This categorization is achieved by clustering the web documents based on the semantic similarity between the ontology terms that characterize them.
- *C-Logs Creation & Mining*. This module takes as input the web site's logs as well as the semantically annotated web site content. It outputs both URI and category-based frequent itemsets and association rules which are subsequently matched to the current user's visit by the recommendation engine.
- *Recommendation Engine*. This module takes as input the current user's path and matches it with the semantically annotated navigational patterns produced in

the previous phases. The recommendation engine generates three different recommendation sets, namely *original*, *semantic* and *category-based* ones, depending on the input patterns used.

The creation of the ontology as well as the semantic similarity measures used as input in the aforementioned web personalization process are orthogonal to the proposed framework. We assume that the ontology is descriptive of the web site’s domain and is provided/created by a domain expert. We elaborated on several similarity measures for ontology terms in Section 3. In what follows we briefly describe the key components of the proposed architecture. For more details on the respective algorithms and system implementation the reader may refer to [14, 15, 38].

4.2 Content Characterization

A fundamental component of the SEWeP architecture is the automatic content characterization process. SEWeP is the only web personalization framework enabling the automatic annotation of web content with ontology terms without needing any human labeling or prior training of the system. The keywords’ extraction is based both on the content of the web pages, as well as their connectivity features. What is more, SEWeP enables the annotation of multilingual content, since it incorporates a context-sensitive translation component which can be applied prior to the ontology mapping process. In the subsections that follow we briefly describe the aforementioned processes, namely the keyword extraction and translation as well as the semantic characterization modules.

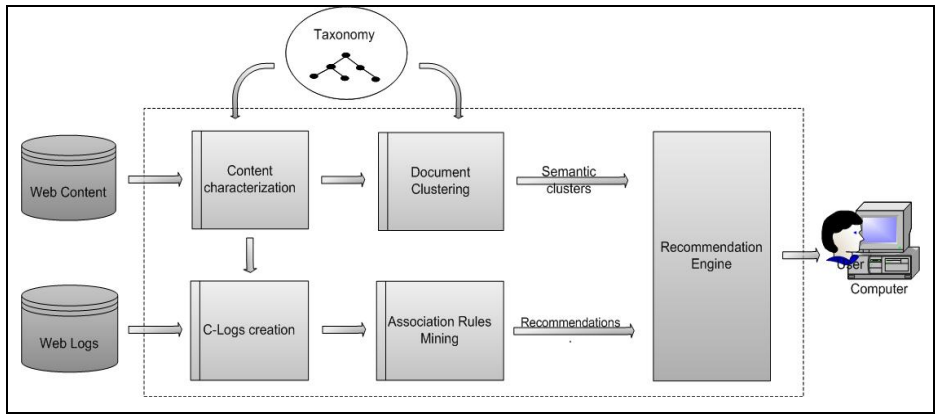


Fig. 1. SEWeP architecture

4.2.1 Keyword Extraction

There exists a wealth of methods for representing web documents. The most straightforward approach is to perform text mining in the document itself following standard IR techniques. This approach, however, 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 [8, 10]. Therefore, in many approaches

information contained in the links that point to the document and the text near them - defined as “anchor-window” - is used for characterizing a web document [10, 46, 45]. This approach is based on the hypothesis that the text around the link to a page is descriptive of the page’s contents and overcomes the problems of the content-based approach, since it takes into consideration the way others characterize a specific web page. In our work, we adopt and extend this approach, by also taking into consideration the content of the pages that are pointed by the page under examination, based on the assumption that in most Web pages the authors include links to topics that are of importance in the page’s context.

In the proposed framework, the keywords that characterize a web page p are extracted using:

1. The raw term frequency of p .
2. The raw term frequency of a selected fraction (anchor-window) of the web pages that point to p .
3. The raw term frequency of the web pages that are pointed by p .

This hybrid content & structure –based keyword extraction process is motivated by the fact that the text around the links pointing to the web page, as well as the content of the web pages pointed by the web page under consideration are descriptive of the page’s contents.

At the end of this phase, each document d is characterized by a weighted set of keywords $d = \{(k_i, w_i)\}$, where w_i is the weight representing the summed (over the combination of methods) word frequency of keyword k_i . Before proceeding with mapping the extracted keywords to related ontology terms, however, all non-English keywords should be translated. In our approach, we determine the most suitable synonym using a context-sensitive automatic translation method. Assuming that the set of keywords will be descriptive of the web page’s content, we derive the best synonym set by comparing their semantics. This translation method is applicable for any language, provided that a dictionary and its inflection rules are available. In our system implementation we applied it for the Greek language. More details on this algorithm can be found in [14, 26].

4.2.2 Semantic Characterization

In order to assist the remainder of the personalization process (C-logs creation, semantic document clustering, semantic recommendations) the n most frequent (translated) keywords that were extracted in the previous phase, are mapped to the terms $T = \{c_1, \dots, c_k\}$ of a domain ontology (in our approach we need the concept hierarchy part of the ontology). This mapping is performed using a thesaurus, like Wordnet [16]. If the keyword belongs to the ontology, then it is included as it is. Otherwise, the system finds the “closest” (i.e. most similar) term (*category*) to the keyword using the unsupervised WSD algorithm proposed in [29]. This algorithm adopts the intuition that context terms (adjacent term in text) are semantically close to each other and that this is reflected by their pathwise distance on the hierarchical structure of the ontology. Based on this intuition the WSD algorithm maps a set of terms to the ontology concepts that minimize the pathwise distances on the ontology hierarchical structure. Thus, the objective of our WSD algorithm is to find the set of senses (among the candidate sets of senses) that is more “compact” in the ontology structure. The compactness measure utilized for selecting the

appropriate set of senses is based on the concept of the Steiner Tree (minimum-weight Tree that connects a set of vertices in a Graph).

If more than one keywords are mapped to the same category c_i , the relevance r_i assigned to it is computed using the following formula:

$$r_i = \frac{\sum_{k_j \rightarrow c_i} (w_j \cdot s_j)}{\sum_{k_j \rightarrow c_i} w_j} \quad (4)$$

where w_j is the weight assigned to keyword k_j for document d and s_j the similarity with which k_j is mapped to c_i . At the end of this process, each document d is represented as a set $d = \{(c_i, r_i)\}$, where $r_i \in [0,1]$ since $s_j \in [0,1]$. If only one keyword k_j is mapped to a category c_i , then the respective relevance r_i equals the keyword's weight w_j .

4.3 C-Logs Creation and Mining

C-Logs are in essence an enhanced form of the web logs. The C-Logs creation process involves the correlation of each web logs' record with the ontology terms that represent the respective URI. C-logs may be further processed in the same way as web logs, through the use of statistical and data mining techniques, such as association rules, clustering or sequential pattern discovery.

The web mining algorithms currently supported by SEWeP is frequent itemsets' and association rules' discovery. Both algorithms are based on a variation of the Apriori algorithm [4], used to extract patterns that represent the visitors' navigational behavior in terms of pages often visited together. The input to the algorithm is the recorded users' sessions expressed both in URI and category level. The output is a set of URI and category-based frequent itemsets or association rules respectively. Since no explicit user/session identification data are available, we assume that a session is defined by all the pageview visits made by the same IP, having less than a maximum threshold time gap between consecutive hits.

4.4 Document Clustering

After the content characterization process, all web documents are semantically annotated with terms belonging to a concept hierarchy. This knowledge is materialized by grouping together documents that are characterized by semantically "close" terms, i.e. neighboring categories in the hierarchy. This categorization is achieved by clustering the web documents based on the similarity among the ontology terms that characterize each one of them. The generated clusters capture semantic relationships that may not be obvious at first sight, for example documents that are not "structurally" close (i.e. under the same root path).

For this purpose we use the THESUS similarity measure, as defined earlier, with a modification of the density-based algorithm DBSCAN [13] for clustering the documents. After the document clustering, each cluster is labeled by the most descriptive categories of the documents it contains, i.e. the categories that characterize more than $t\%$ of the documents. Modification details and the algorithm itself are described in [19, 46]. The semantic document clusters are used in turn in order to expand the recommendation set with semantically similar web pages, as we describe in the subsequent Section.

4.5 Recommendation Engine

As already mentioned, after the document characterization and clustering processes have been completed, each document d is represented by a set of weighted terms (categories) that are part of the concept hierarchy: $d = \{(c_i, r_i)\}$, $c_i \in T$, $r_i \in [0, 1]$ (T is the concept hierarchy, r_i is c_i 's weight). This knowledge can be transformed into three different types of recommendations, depending on the rules that are used as input (association rules between URIs or between categories) and the involvement of semantic document clusters: *original*, *semantic*, and *category-based recommendations*.

Original recommendations are the “straightforward” way of producing recommendations, simply relying in the usage data of a web site. They are produced when, for each incoming user, a sliding window of her past n visits is matched to the *URI-based* association rules in the database, and the m most similar ones are selected. The system recommends the URIs included in the rules, but not visited by the user so far.

The intuition behind *semantic recommendations* is that, useful knowledge semantically similar to the one originally proposed to the users, is omitted for several reasons (updated content, not enough usage data etc.) Those recommendations are in the same format as the *original* ones but the web personalization process is 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. In short, they are produced when, for each incoming user, a sliding window of her past n visits is matched to the *URI-based* association rules in the database, and the single most similar one is selected. The system finds the URIs included in the rule but not yet visited by the user (let A) and recommends the m most similar documents that are in the same semantic cluster as A .

Finally, the intuition behind *category-based* recommendations is the same as the one of *semantic* recommendations: incorporate content and usage data in the recommendation process. This notion, however, is further expanded; users' navigational behavior is now expressed using a more abstract, yet semantically meaningful way. Both the navigational patterns' knowledge database and the current user's profile are expressed by categories. Therefore, pattern matching to the current user's navigational behavior is no longer exact since it utilizes the semantic relationships between the categories, as expressed by their topology in the domain-specific concept hierarchy. The final set of recommendations is produced when, for each incoming user, a sliding window of the user's past n visits is matched to the category-based association rules in the database, and the most similar is selected. The system finds the most relevant document cluster (using similarity between category terms) and recommends the documents that are not yet visited by the user.

In what follows, we briefly describe the *semantic* and *category-based recommendations*' algorithms. The description of the generation of original recommendations is omitted, since it is a straightforward application of the Apriori [4] algorithm to the sessionized web logs. The respective algorithms, as well as experimental evaluation of the proposed framework can be found in [12, 14, 15].

4.5.1 Semantic Recommendations

We use the Apriori algorithm to discover frequent itemsets and/or association rules from the C-Logs. We consider that each distinct user session represents a different

transaction. We will use $S = \{I_m\}$, to denote the final set of frequent itemsets/association rules, where $I_m = \{uri_i\}$, $uri_i \in CL$.

The recommendation method takes as input the user's current visit, expressed a set of URIs: $CV = \{uri_j\}$, $uri_j \in WS$, (WS is the set of URIs in the visited web site. Note that some of these may not be included in CL). The method finds the itemset in S that is most similar to CV , and recommends the documents (labeled by related categories) belonging to the most similar document cluster $Cl_m \in Cl$ (Cl is the set of document clusters). In order to find the similarity between URIs, we perform binary matching. In other words, the more common URIs in CV and S , the more similar they are.

4.5.2 Category-Based Recommendations

We use an adaptation of the Apriori algorithm to discover frequent itemsets and/or association rules including categories. We consider that each distinct user session represents a different transaction. Instead of using as input the distinct URIs visited, we replace them with the respective categories. We keep the most important ones, based on their frequency (since the same category may characterize more than one documents). We then apply the Apriori algorithm using categories as items. We will use $C = \{I_k\}$, to denote the final set of frequent itemsets/association rules, where $I_k = \{(c_i, r_i)\}$, $r_i \in T$, $r_i \in [0,1]$ (r_i reflects the frequency of c_i).

The recommendation method takes as input the user's current visit, expressed in weighted category terms: $CV = \{(c_j, f_j)\}$, $c_j \in T$, $f_j \in [0,1]$ (f_j is frequency of c_j in current visit - normalized). The method finds the itemset in C that is most similar to CV , creates a generalization of it and recommends the documents (labeled by related categories) belonging to the most similar document cluster $Cl_n \in Cl$ (Cl is the set of document clusters). To find the similarity between categories we use the Wu & Palmer metric, whereas in order to find similarity between sets of categories, we use the THESUS metric, as defined in Section 3.

5 Conclusions

The exploitation of the pages' semantics hidden in user paths can considerably improve the results of web personalization, since it provides a more abstract yet uniform and both machine and human understandable way of processing and analyzing the data. In this paper, we present a semantic web personalization framework, which enhances the recommendation process with content semantics. We focus on word sense disambiguation techniques which can be used in order to semantically annotate the web site's content with ontology terms. The framework exploits the inherent semantic similarities between the ontology terms in order to group web documents together and semantically expand the recommendation set.

A more detailed analysis, possible limitations and an extensive experimental evaluation of the several components of the proposed framework can be found in [14, 15, 29]. The experimental results are more than promising. Our plans for future work include the evaluation of the proposed integrated SEWeP-GVSM framework. We also plan to integrate different semantic similarity measures in our architecture.

References

1. E. Agirre, G. Rigau, *A proposal for word sense disambiguation using conceptual distance*, In Proc. of Recent Advances in NLP (RANLP), 1995, pp. 258–264
2. S. Acharyya, J. Ghosh, *Context-Sensitive Modeling of Web Surfing Behaviour Using Concept Trees*, in Proc. of the 5th WEBKDD Workshop, Washington, August 2003
3. M. Albanese, A. Picariello, C. Sansone, L. Sansone, *A Web Personalization System based on Web Usage Mining Techniques*, in Proc. of WWW2004, May 2004, New York, USA
4. R. Agrawal, R. Srikant, *Fast Algorithms for Mining Association Rules*, in Proc. of 20th VLDB Conference, 1994
5. R. Barzilay, M. Elhadad, *Using lexical chains for text summarization*, in Proc. of the Intelligent Scalable Text Summarization Workshop (ISTS 1997), ACL, 1997.
6. B. Berendt, A. Hotho, G. Stumme, *Towards Semantic Web Mining*, in Proc. of 1st International Semantic Web Conference (ISWC 2002)
7. S. Bloehdorn, A. Hotho: *Boosting for text classification with semantic features*. In: Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Mining for and from the Semantic Web Workshop. (2004) 70–87
8. S. Brin, L. Page, *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks, 30(1-7): 107-117, 1998, Proc. of the 7th International World Wide Web Conference (WWW7)
9. R. Baraglia, F. Silvestri, *An Online Recommender System for Large Web Sites*, in Proc. of ACM/IEEE Web Intelligence Conference (WI'04), China, September 2004
10. 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
11. H. Dai, B. Mobasher, *Using Ontologies to Discover Domain-Level Web Usage Profiles*, in Proc. of the 2nd Workshop on Semantic Web Mining, Helsinki, Finland, 2002
12. M. Eirinaki, *New Approaches to Web Personalization*, PhD Thesis, Athens University of Economics and Business, Dept. of Informatics, 2006
13. 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
14. M. Eirinaki, C. Lampos, S. Pavlakis, M. Vazirgiannis, *Web Personalization Integrating Content Semantics and Navigational Patterns*, in Proc. of the 6th ACM International Workshop on Web Information and Data Management (WIDM'04), November 2004, Washington DC
15. M. Eirinaki, M. Vazirgiannis, I. Varlamis, *SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process*, in Proc. of the 9th ACM International Conference on Knowledge Discovery and Data Mining (KDD '03), Washington DC, August 2003
16. C. Fellbaum, ed., *WordNet, An Electronic Lexical Database*. The MIT Press, 1998
17. M. Galley, K. McKeown, *Improving Word Sense Disambiguation in Lexical Chaining*, in Proc. of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), August 2003, Acapulco, Mexico.
18. J. Guo, V. Keselj, Q. Gao, *Integrating Web Content Clustering into Web Log Association Rule Mining*. In Proc. of Canadian AI Conference 2005
19. M. Halkidi, B. Nguyen, I. Varlamis, M. Vazirgiannis, *THESUS: Organizing Web Documents into Thematic Subsets using an Ontology*, VLDB journal, vol.12, No.4, 320-332, Nov. 2003

20. A. Hotho, S. Staab, G. Stumme: *Ontologies Improve Text Document Clustering*. Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA: 541-544
21. J. Jiang, D. Conrath, *Semantic similarity based on corpus statistics and lexical taxonomy*, in Proc. of the International Conference on Research in Computational Linguistics, 1997
22. X. Jin, Y. Zhou, B. Mobasher, *A Maximum Entropy Web Recommendation System: Combining Collaborative and Content Features*, in Proc. of the 11th ACM International Conference on Knowledge Discovery and Data Mining (KDD '05), Chicago, August 2005
23. P. Kearney, S. S. Anand, *Employing a Domain Ontology to gain insights into user behaviour*, in Proc. of the 3rd Workshop on Intelligent Techniques for Web Personalization (ITWP 2005), Endinburgh, Scotland, August 2005
24. C. Leacock, M. Chodorow. *Combining Local Context and WordNet Similarity for Word Sense Identification*. In Fellbaum 1998, pp. 265--283.
25. C. Leacock, M. Chodorow, G. A. Miller. 1998. *Using Corpus Statistics and WordNet Relations for Sense Identification*. In Computational Linguistics, 24:1 pp. 147-165.
26. C. Lampos, M. Eirinaki, D. Jevtuchova, M. Vazirgiannis, *Archiving the Greek Web*, in Proc. of the 4th International Web Archiving Workshop (IWA04), September 2004, Bath, UK
27. M. E. Lesk, *Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone*, in Proc. of the SIGDOC Conference, June 1986, Toronto.
28. D. Lin, *An information-theoretic definition of similarity*, in Proc. of the 15th International Conference on Machine Learning (ICML), 1998, pp. 296-304
29. D. Mavroeidis, G. Tsatsaronis, M. Vazirgiannis, M. Theobald, G. Weikum, *Word Sense Disambiguation for Exploiting Hierarchical Thesauri in Text Classification*, in Proc. of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, (PKDD'05), Porto, Portugal, 2005
30. 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 Ecommerce and Web Technologies (ECWeb2000), Greenwich, UK, September 2000
31. A. Montoyo, A. Suarez, G. Rigau, M. Palomar, *Combining Knowledge- and Corpus-based Word-Sense-Disambiguation Methods*, Journal of Artificial Intelligence Research, 23, pp.299-330.
32. R. Mihalcea, D. I. Moldovan, *A Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation*, International Journal on Artificial Intelligence Tools, 2001, 10(1-2), pp. 5-21.
33. S.E. Middleton, N.R. Shadbolt, D.C. De Roure, *Ontological User Profiling in Recommender Systems*, ACM Transactions on Information Systems (TOIS), Jan. 2004/ Vol.22, No. 1, 54-88
34. R. Mihalcea, P. Tarau, E. Figa, *Pagerank on semantic networks, with application to word sense disambiguation*, in Proc. of the 20th International Conference on Computational Linguistics (COLING 2004), August 2004, Switzerland.
35. R. Navigli, P. Velardi, *Structural Semantic Interconnection: a knowledge-based approach to Word Sense Disambiguation*, IEEE Transactions on Pattern Analysis and Machine Intelligence (TRAMI), 27(7), p. 1075-1086.
36. D.Oberle, B.Berendt, A.Hotho, J.Gonzalez, *Conceptual User Tracking*, in Proc. of the 1st Atlantic Web Intelligence Conf. (AWIC), 2003
37. M. Perkowit, O. Etzioni, *Towards Adaptive Web Sites: Conceptual Framework and Case Study*, in Artificial Intelligence 118[1-2] (2000), pp. 245-275

38. S. Paulakis, C. Lampos, M. Eirinaki, M. Vazirgiannis, *SEWeP: A Web Mining System supporting Semantic Personalization*, in Proc. of the ECML/PKDD 2004 Conference, Pisa, Italy, September 2004
39. P. Resnik, *Using information content to evaluate semantic similarity in a taxonomy*, in Proc. of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995
40. G. Rigau, J. Atserias, E. Agirre. 1997. *Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*, in Proc. of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain.
41. B. Sarwar, G. Karypis, J. Konstan, J. Riedl, *Item-based Collaborative Filtering Recommendation Algorithms*, in Proc. of WWW10, May 2001, Hong Kong
42. G. Silber, K. McCoy, *Efficiently computed lexical chains as an intermediate representation for automatic text summarization*, Computational Linguistics, 29(1), 2003.
43. M. Sussna, *Word sense disambiguation for free-text indexing using a massive semantic network*. In: Proc. of the 2nd International Conference on Information and Knowledge Management (CIKM). (1993) 67–74
44. Theobald, M., Schenkel, R., Weikum, G.: *Exploiting structure, annotation, and ontological knowledge for automatic classification of xml data*. In: International Workshop on Web and Databases (WebDB). (2003) 1–6
45. H. Utard, J. Furnkranz, *Link-Local Features for Hypertext Classification*, in Proc. of the European Web Mining Forum (EWMF 2005), Porto, Portugal, October 2005
46. I. Varlamis, M. Vazirgiannis, M. Halkidi, B. Nguyen, *THESUS, A Closer View on Web Content Management Enhanced with Link Semantics*, in IEEE Transactions on Knowledge and Data Engineering Journal, June 2004 (Vol. 16, No. 6), pp. 585-600.
47. S.K.M. Wong, W. Ziarko, P.C.N. Wong, *Generalized vector space model in information retrieval*, in Proc. of the 8th Intl. ACM SIGIR Conference (SIGIR'85), 1985, pp. 18–25
48. Z. Wu, M. Palmer, *Verb Semantics and Lexical Selection*, 32nd Annual Meetings of the Associations for Computational Linguistics, 1994, pp. 133-138