Knowl Inf Syst (2012) 30:527–541
DOI 10.1007/s10115-011-0386-4

REGULAR PAPER

Improving pattern quality in web usage mining by using semantic information

Pinar Senkul · Suleyman Salin

Received: 19 April 2010 / Revised: 5 January 2011 / Accepted: 6 February 2011 /
Published online: 24 February 2011

Abstract Frequent Web navigation patterns generated by using Web usage mining techniques provide valuable information for several applications such as Web site restructuring and recommendation. In conventional Web usage mining, semantic information of the Web page content does not take part in the pattern generation process. In this work, we investigate the effect of semantic information on the patterns generated for Web usage mining in the form of frequent sequences. To this aim, we developed a technique and a framework for integrating semantic information into Web navigation pattern generation process, where frequent navigational patterns are composed of ontology instances instead of Web page addresses. The quality of the generated patterns is measured through an evaluation mechanism involving Web page recommendation. Experimental results show that more accurate recommendations can be obtained by including semantic information in navigation pattern generation, which indicates the increase in pattern quality.

Keywords Semantics · Ontology · Web usage mining · Sequential association rule mining · Recommendation · Frequent navigation pattern

1 Introduction

Semantic Web aims to make Web content understandable not only by humans but also by computers, thus permitting the software agents to search for and find desired content, share information and knowledge [5]. Therefore, there is an increasing effort in annotating Web pages and objects in terms of semantic information by using ontologies [23,24].

Web mining is the branch of data mining that extracts useful information from Web resources, including content, structure, and navigation logs. In Web usage mining, the goal is to extract navigation behavior of users. The extracted patterns are useful in several different areas including recommendation, Web site restructuring, prefetching, etc. Various

P. Senkul (✉) · S. Salin
Computer Engineering Department, Middle East Technical University,
06531 Ankara, Turkey
e-mail: senkul@ceng.metu.edu.tr

techniques can be used for Web usage pattern extraction such as clustering [22,37], neural networks [8], genetic programming [1], sequence mining, and sequence alignment [13]. Some of the studies emphasize more specific issues such as pattern extraction under constraint [27], mining evolving patterns [21], and time performance [14]. In conventional Web usage mining techniques, resulting patterns are presented in terms of Web pages (i.e., Web page addresses), hence, semantics of the extracted navigation profiles is not explicit. Site owners generally want to learn which category or group is preferred, rather than which page is more frequently visited. In order to extract the semantics of the navigation, the patterns generated by conventional Web usage mining techniques must be analyzed by humans, and this process is erroneous and time-consuming.

Recommendation is a heavily studied research subject, where Web usage patterns can improve the accuracy of the task. In recommender systems, conventionally, collaborative filtering techniques [2,4,36] are frequently used. Association-based and graph-based methods [12,15] and symbolic data analysis tools [6] are also elaborated in recommender system studies. Personalized search engines and crawlers [26,35] are also closely related with recommendation in the sense that best recommendations are generated as a result of a query. Employing Web usage mining for recommendation makes use of similar users' behavior patterns as in collaborative filtering, however, it avoid the drawbacks of collaborative filtering such as subjective nature of ratings [2,18].

In this work, we present an approach to generate frequent navigation patterns enriched with semantic information of the Web pages. As the navigation pattern structure, sequential association rules are used. A conventional sequence association rule generation technique is extended in such a way that the navigation patterns are constructed in terms of ontological instances.

To date, previous work on Web usage mining with semantic information is very limited. There are studies that aim to generate patterns in terms of semantic information as in [5,11,33]. However, generally, these works map out the results of classical Web usage mining with ontological terms and concepts. Clustering appears as the major pattern generation technique, and the resulting patterns are used for generating Web page recommendations [19,39]. Unlike most of the previous approaches, we preferred to use sequential association rule mining, since the sequence information in the navigation is retained in the generated patterns. In addition, in the proposed technique, semantic information is used within pattern generation, rather than in postprocessing. We have conducted several experiments in order to evaluate the effect of semantic information integration into the pattern generation process through the proposed approach in several directions. The quality of the generated patterns are evaluated within a recommendation environment. By making use of the semantically enhanced patterns, recommendations are generated for a set of test cases. The results are evaluated by comparing the generated recommendation with the visitor's actual next page in terms of precision and coverage. A shorter description of the proposed work with limited experimental results is presented in [25].

This paper is organized as follows: In Sect. 2, related work is presented. In Sect. 3, proposed approach is described. Section 4 gives the experimental results. Finally, Sect. 5 includes the final remarks.

2 Related work

The number of studies that combine Web usage mining and semantic Web is quite limited. Most of them emphasize the use of Web usage mining for extracting semantics of Web site

with minimal manual intervention [34,41]. The proposed work belongs to the group of studies that aims to use semantics for improving the patterns generated by Web usage mining.

One of the first studies on improving Web usage mining by incorporating semantics is presented in [5,34]. They suggest that Web site log files should register the user behavior in terms of ontology concepts and ontology terms, so that this log can then be mined by clustering or association rule mining algorithms.

Another work given in [33] presents basic ideas on how the Semantic Web can improve Web usage mining, and how usage mining can help to build the Semantic Web. The work assumes that the Web site logs contain Web requests grouped in sessions, and the ontology and knowledge base of the Web site are available. Each Web request is mapped to concepts at a chosen level of abstraction. The level of abstraction can be found dynamically by using an algorithm [19,31], or it can be done manually. Following this, any data mining method such as sequence miner [7,3,28] or classification [30,29] can be used to discover the navigation patterns. The work given in [33] and the proposed work share common ideas, however [33], does not present the details of pattern generation with semantic information and it does not include recommendation generation on the basis of constructed navigation patterns.

In [11,19], a general framework has been presented for integrating domain ontologies with Web usage mining and personalization. In [11], the emphasis is on the personalization process and the frequent pattern generation is done by clustering.

The work presented in [17] is similar to this work in such a way that both studies use sequence association rules as the pattern structure and aim to incorporate semantic information into the pattern generation process. However, the techniques to incorporate semantics into pattern generation differ considerably. In [17], the basic idea is to discover navigations by following the relations in the ontology, and this increases the time complexity of the algorithm considerably. The work does not present any analysis on the quality of the resulting patterns. Therefore, the effectiveness of the proposed algorithm remains unclear.

Another similar work that aims to enhance association rules with domain ontologies is given in [16]. However, the work presented in [16] emphasizes the improvement of time efficiency for online next page prediction, rather than pattern quality.

There are efforts to increase the precision and recall of search engines by analyzing previous queries of a user by using formal concept analysis [9,10]. Such studies share similar points with the proposed work on the use of concepts. However, there are fundamental differences such that the main objective in semantically enhanced Web usage mining is to generate useful navigation patterns on a Web site rather than Web search queries. As a result of this, the nature of the analyzed data and the employed techniques differ considerably.

3 Using semantic information in web usage mining

The proposed framework consists of three basic phases: preprocessing, rule extraction, and evaluation. The details of these steps are described in the rest of this section.

3.1 Preprocessing

Web server log files generally contain a large amount of noisy and irrelevant data. In addition to this, it is necessary to integrate semantic information of the Web pages with their log data. In this step, Web server log files are pruned, transactions are extracted, and ontology class individuals are mapped to the Web page addresses.

In the pruning step, non-responded requests and requests made by software agents such as Web crawlers are eliminated by using the error codes and access information in the logs. In our experiments, we have observed that, on the average, 5% of the Web requests are non-responded and 30% are made by Web crawlers.

The next step is the extraction of the navigation history of each session from the log file. The navigation history is the set of Web objects requested by the user in his/her active session time. In this work, we pruned the Web object requests that are not relevant to the analysis (such as a request for a picture on a Web page) and set the session timeout duration (i.e., idle amount of time that terminates the session) to be 20 min. Generally, the Web servers store the information on active sessions in the memory of the Web server, and this information is accessible by an identifier called session-id. The session-id is stored in the end user's cache or disk and at each request and it is sent to the Web server using various kinds of information exchange techniques, such as browser cookies, query strings, or a hidden field in the Web page. In our study, we used session-id's determined by the Web server for determining the navigation history of sessions.

The last step of preprocessing is the mapping between ontology instances and the requested Web address in the Web server log. In most of the similar studies, existence of semantically annotated Web pages is assumed. The navigation pattern generation technique of this work relies on such an assumption, as well. However, semantically annotated Web logs are not available as benchmark data and research groups generally generate their own data sets. Data sets used in this work are log data obtained from live Web sites. However, originally these Web pages are not semantically annotated. Since the automatic extraction of semantic information is a hard task and it is a research problem on its own, we accomplished this task manually. Firstly, we constructed ontologies and ontology instances in OWL [38] for the Web sites. For this task, we used the domain information obtained from Web site administrators and the database structure used in the Web site. Construction of the ontology did not take a long time (it took about 3–4 h) as the underlying database structures and Web page structures almost already reflected the ontology structure. For other data sets, this process may take slightly longer time, however, Web page structures are always helpful starting points. If the ontology is already available and Web objects are already annotated with the ontological objects, they can be directly included in the algorithm.

Figure 1 presents a snapshot from the Book ontology defined for an e-bookstore Web site, whose logs are used for experiments of this work. This ontology contains the concepts of *Book, Author, Category, Price, and Publisher*. The Category concept includes a hierarchy relation in which higher lever categories includes more detailed lower level categories (such as *Computer Books* category having a child named *Computer Graphics Books* category). Once the semantic information is defined in the form of ontologies, the second task is the semantic annotation of Web pages. Since we assumed a navigation history to be consisting of Web page URL's, we defined correspondence between the Web pages and the ontology instances. On the basis of the ontology structure, one URL may be mapped to several ontology instances. For instance, for the e-bookstore data set, a Web page about a certain book may be mapped to a book instance, author instance(s), and a book category instance. This task is offline and manual. The decision as to which concept instances are suitable for mapping is made on the basis of the content of the Web page. In the data sets used in the experiments, the Web page structures are homogeneous and always mapped to the same type of instances (instances of e-bookstore data set, book, author, book category, and publisher concepts). During pattern generation, one particular selected concept is considered and only the mappings to the instances of the selected concept is considered. The concept selection is performed by the user through the graphical user interface of our system. The user interface

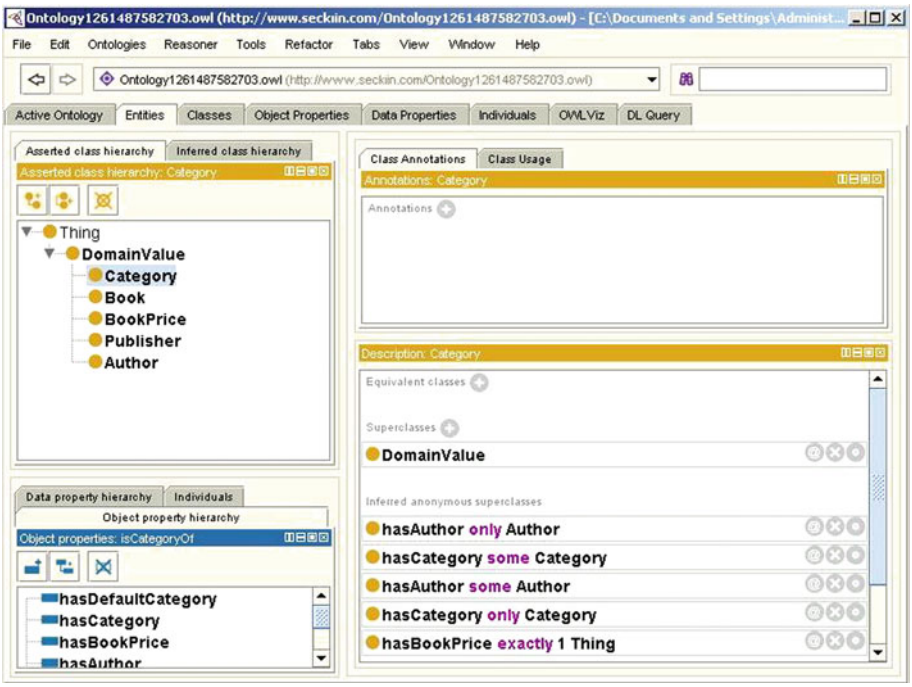

Fig. 1 A snapshot of the ontology defined for E-bookstore Web site

shows the ontology associated with the Web site whose logs are to be analyzed. On the presented ontology, the user can specify the concept she/he is interested in. Then, the navigation patterns are generated in terms of the selected concept. In other words, the mapping between the selected concept and the Web page addresses is used in pattern generation.

3.2 Frequent navigation pattern generation

The pattern language employed in our study is the sequential association rule structure. The reason to this choice is that, frequent item sequences keep the sequence relation among the items and sequential association rules can describe the subsequences that frequently follow another subsequence. Due to its time efficiency, SPADE [40] is used as the sequential association rule mining algorithm. However, several modifications are applied on the algorithm. The first modification is for generating patterns consisting of ontology instances. In addition to this, taxonomies are introduced into pattern generation, as well. Although an ontology may include more complex relationships, we limited our scope of the relationships to taxonomy structures. There are two basic reasons for this. Firstly, taxonomy is one of the basic relationship types and it exists in most of the ontology definitions in the form of *is-a* relation among the concepts. Secondly, ontology structures annotating the Web pages used in the experiments of this work are not very complicated and the relationships are in the form on hierarchies.

The SPADE algorithm uses a vertical data set in order to facilitate the pattern generation. Since ontology instances become the main element, the vertical data sets are populated with ontology instances. For this step, the mapping, which is described previously in Sect. 3.1,

Table 1 Sample rules in terms of category class for e-bookstore Web site

Rule	Confidence
Computer ⇒ Computer	0.40
Economics ⇒ Economics	0.40
Law, Periodicals ⇒ Law	0.76
Law→Law ⇒ Law	0.61
System_Administration_Network⇒ Internet	0.55

Table 2 Sample rules in terms of group class for METU CENG Web site

Rule	Confidence
metu.ceng.ses ⇒ metu.ceng.news	0.80
metu.ceng.test ⇒ metu.ceng.test → metu.ceng.undergrad	0.89
metu.ceng.test → metu.ceng.test ⇒ metu.ceng.announcement	0.90

is used. An important point here is that the user chooses the concept (i.e., ontology instance mapping) over which the patterns will be generated. As the result of iteration through candidate generation and frequent itemset extraction steps, the generated patterns are composed of ontology instances.

If a taxonomy exists in the ontology, an additional step is performed. Taxonomy information is read from the ontology file and it is written into a tree data structure that keeps the *is-a* relations between concepts. The vertical data set is extended in such a way that each item is also associated with the item's taxonomical parent. For example, Computer Engineering is a sub-category of Engineering so that each item from Computer Engineering also belongs to the Engineering category. In order to prevent generating too many overlapping association rules, uninteresting rules are pruned. As defined in [31,32], an uninteresting rule is a rule whose support count is less than an expected value. The expected value of any rule is the expected support count of the rule, relative to its taxonomical parent rule.

Some examples for the generated frequent sequence rules are presented in Table 1 and Table 2. Table 1 shows sample rules in terms of Category class for an e-bookstore Web site. Table 2 includes sample rules in terms of Groups for Middle East Technical University Computer Engineering (METU CENG) newsgroup Web site. The rules are shown in the form $A \Rightarrow B$, where A and B denote sequences. A sequence A is represented as $A_1 \rightarrow \cdots \rightarrow A_n$, where each A_i is either a single item or an itemset (i.e., set of items separated with commas). For example, the rule $Law \rightarrow Law \Rightarrow Law$ with confidence [1] 0.61 tells that the two consecutive visits to Law category books will be followed by another visit to a Law category book with 61% confidence.

[1] Following the confidence value calculation in the literature, confidence of an association rule is calculated as the division of number of sessions including the sequence that is concatenation of body of the rule to the head of the rule to the number of sessions including the sequence in the rule head. For a sequential association rule, $a \rightarrow b \Rightarrow c$, $a \rightarrow b$ is the head of the rule and c is the body of the rule. When body is concatenated to the head, we end up with the sequence $a \rightarrow b \rightarrow c$. Therefore, confidence for this rule is ((number of sessions including $a \rightarrow b \rightarrow c$) / (number of sessions including $a \rightarrow b$)). This confidence value tells us how strongly $a \rightarrow b$ is followed by c.

3.3 Evaluation of generated semantically enhanced web traversal patterns

The generated patterns can be used for various purposes such as restructuring Web sites, reorganizing links and recommending Web pages. Since it is not quite straightforward to evaluate the effect of semantic information on the pattern quality, we developed an evaluation framework in which generated patterns are used for Web page recommendation. We preferred to use Web page recommendation task since the effect of patterns is more obvious and there are established methods for evaluating recommendations. It is important to note that recommendation is basically used for observing and evaluating the generated navigation patterns. The use of additional methods for improving the accuracy of recommendation is outside the scope of this study.

Recommendation generation is performed by comparing sequential association rules and active user's navigation history. Generally, a very early page that the user visited is less likely to affect the next page since users generally make the decision about what to click by the most recent pages. Therefore, the concept of window count is introduced. *Window Count* parameter defines the maximum number of previous page visits to be used while recommending a new page.

Since the association rules are composed of ontology individuals, the user navigation history is converted into the sequence of ontology instances. Afterward, the association rules and user navigation history are joined in order to produce recommendations.

In the recommendation phase, firstly the most recently navigated item is taken as the search pattern. All the association rules are scanned, and the association rules whose antecedent part is equal to the search pattern are added to the *recommendation set*. This step iterates window count times and at each iteration, the search pattern is extended by one item. For each association rule in the recommendation set, consequent part is extracted. The ontology instances in the rule consequents are mapped back to Web page addresses. There may be many-to-one mappings, as in mapping from a book category to individual book Web pages. In such cases, all of the mappings are considered in recommendation. Under the existence of taxonomy, this includes mappings for the subconcepts as well. This may lead to overly high number of recommendations. We conducted experiments to evaluate the accuracy as well as coverage of the generated recommendations.

As an example consider, the active user navigation(WP stands for Web_Page)

$WPBook : Introduction\_to\_Law \rightarrow WPBook : Art\_of\_Internet$, where these two pages are concrete Web pages with URLs. Assume that the selected concept in Book ontology is Category. As the first step, user navigation history is mapped to ontology instances. Assume that $WPBook : Introduction\_to\_Law$ is mapped to *Law* category instance and $WPBook : Art\_of\_Internet$ is mapped to *Internet* category instance. As the result of this step, user navigation is changed to $Law \rightarrow Internet$. Assume that the only rule we have is $Computer \Rightarrow Programmming$. It can be used for recommendation since $Computer$ is the parent of $Internet$ category. If $WPBook : C\_Programming$, $WPBook : Java\_Programming$ and $WPBook : Python\_Programming$ are the Web addresses that are mapped to the instance of *Programming* concept and its subcategories, then $WPBook : C\_Programming$, $WPBook : Java\_Programming$ and $WPBook : Python\_Programming$ are recommended to the user.

The evaluation method of this work is based on the techniques introduced in [20]. The effectiveness of the recommendation is measured in terms of coverage and precision. Tenfold cross-validation is performed for each of the data sets. Each transaction t in the test set is divided into two parts. The first part is the first n items in t for recommendation generation. The other part, which is denoted as $eval_t$, is the remaining portion of t to evaluate the

recommendation. Once the recommendation phase produces a set of page views, which is denoted as Rec_t, the set is compared with $eval_t$ page views.

Precision is defined as the proportion of the number of relevant recommendations to the number of all recommendations. In other words, precision measures the accuracy of the recommendations.

$$precision_t = \frac{|Rec_t \cap eval_t|}{|Rec_t|} \tag{1}$$

Coverage measures the ability of the recommendation system to produce all the page views that are likely to be visited by the user. In other words, it shows how well the recommendation covers all the pages that the user is likely to visit.

$$coverage_t = \frac{|Rec_t \cap eval_t|}{|eval_t|} \tag{2}$$

In this study, two new metrics, namely *precision-with-threshold* and *coverage-with-threshold*, are introduced. For certain domains or users, a threshold value on precision and coverage determines the *goodness* of the recommendation. In such cases, a recommendation is considered successful if the precision/coverage value is greater than a certain value; on the other hand, it is a total failure when the precision/coverage value is lower than this value.

The motivation for devising these new metrics is to measure the success of proposed technique for satisfying a given level of precision and coverage rather than to obtain the average precision and coverage values. For instance, when threshold for precision-with-threshold is set to 100%, it means that 100% precision should be obtained for the recommendation for a user navigation to be counted as successful. Therefore, precision-with-threshold value tells the portion of recommendations that satisfies the given level of precision.

The precision-with-threshold is an extension to the precision measurement. The value of new precision can be either 0 or 1. If the precision is greater than the given precision threshold τ, then the value of precision-with-threshold is 1; otherwise, the value is 0.

$$precision - with - threshold_t = 0, \quad if precision_t < \tau$$
$$precision - with - threshold_t = 1, \quad if precision_t \geq \tau$$

The definition for coverage-with-threshold follows the same structure.

$$coverage - with - threshold_t = 0, \quad \text{if coverage}_t < \tau$$
$$coverage - with - threshold_t = 1, \quad \text{if coverage}_t \geq \tau$$

As an example consider the active user navigation:

$WPBook : Introduction\_to\_Law \rightarrow WPBook : Art\_of\_Internet \rightarrow WPBook : Python\_Programming.$

Assume that $eval_t$ is $WPBook : Python\_Programming$. As described in the example of the previous section, Rec_t generated for the user navigation

$WPBook : Introduction\_to\_Law \rightarrow WPBook : Art\_of\_Internet$ is as follows:

$$\{WPBook : C\_Programming, WPBook : Java\_Programming,$$
$$WPBook : Python\_Programming\}.$$

The coverage for this recommendation is 100% whereas precision is only 33.3%. Coverage-with-threshold is 100% under any given threshold. However, precision-with-threshold is 0% for any threshold above 33.3% and is 100% for any threshold equal to or below 33.3%.

4 Experimental evaluations

For the experiments, the server logs from two different actual Web sites are used. The first one belongs to an electronic book store (e-store) with about 10,000 books in 300 categories. It contains 6,800 Web page views and 1,700 unique sessions, daily, after preprocessing. The average number of Web pages in a session is 4. The ontology model of the domain includes the concepts Book, Author, PublishYear, Category, and Publisher concepts. The Category concept has a hierarchical structure, i.e., it includes taxonomy. In this ontology, there are 4,545 Authors individuals, 249 Publisher individuals, 34 PublishYear individuals, and 266 Category individuals.

The second Web site log belongs to Computer Engineering (CENG) Department of Middle East Technical University (METU). It consists of many sub-Web sites including Web pages of individuals (i.e., students, teachers), newsgroups, and courses. In the experiments, only the Web logs of the newsgroups are used. The logs of CENG newsgroup include about 8,717 postings in 103 groups. After preprocessing, there remains 17,000 page views and 1,100 distinct sessions daily. The average number of Web pages navigated is 15. The newsgroup ontology includes concepts of Thread, Author, and Group.

In order to evaluate the generated recommendations, four groups of tests are performed. Under both semantic and URL-based patterns, the recommendation environment recommends Web pages (i.e., URLs). However, the issue is whether the prediction of the URLs is facilitated by the semantic information. For URL-based patterns, the recommender can make use of only this very solid and rigid information. However, semantic-based patterns are more abstract and hold the transitions among concepts rather than transitions among the concrete Web pages. For this reason, semantic-based patterns carry the potential to predict a previously unseen transition among URLs by using the transitions among concepts. The first set of experiments evaluate this potential of semantic patterns. The following three groups of experiments are for the further analysis on semantic-based patterns. The second group of experiments shows the effect of precision and coverage threshold on the performance of recommendation. The third group of experiments investigates the effect of increasing the amount of semantic information for recommendation. In the last group of tests, the effect of number of association rules used in the recommendation to the precision and coverage values is revealed.

There are several reasons for not providing comparison with other work from the literature. The first one is that the structures of the patterns generated are very different in each work and they are not always compatible with the recommendation framework we use. In addition, the implementations are not publicly available or described clearly enough. If we implement them ourselves (which we did for some of the studies in the literature), they will rely on our own assumptions or implementations for the details missing in the literature. As the last point, there is no publicly available benchmark data set for semantic Web usage mining yet, so that published results can be used for comparison.

4.1 Experiments on the effect of incorporating semantic information

The first set of experiments compare the recommendation results by using URL-based patterns and semantically enhanced patterns. Table 3 (a) shows the result when patterns are extracted without semantic information. In this experiment, generated traversal patterns consist of URL addresses of the pages. As seen in the table, precision and coverage values are quiet low in this experiment.

Table 3 Experiments on support threshold change on the accuracy

Rec by URL for e-store Web site (a)	Supp. thr.	0.625%	0.875%	1.125%	1.375%	1.625%	1.875%
	Prec.	0.048	0.025	0.03	0.03	0.03	0
	Cov.	0.032	0.02	0.01	0.005	0.005	0
	Prec.-w-thr.	0.055	0.025	0.03	0.03	0.03	0
	Cov.-w-thr.	0.04	0.025	0.01	0.005	0.005	0
Rec by category for e-store Web site (b)	Supp. thr.	2.5%	4%	5.5%	7%	8.5%	10%
	Prec.	0.32	0.48	0.52	0.45	0.25	0.3
	Cov.	0.6	0.4	0.32	0.26	0.15	0.13
	Prec.-w-thr.	0.38	0.55	0.58	0.52	0.31	0.3
	Cov.-w-thr.	0.6	0.44	0.54	0.28	0.15	0.14
Rec by posting for CENG Web site (c)	Supp. thr.	0.4%	0.8%	1.2%	2%	4%	6%
	Prec.	0.075	0.05	0.05	0.05	0.05	0
	Cov.	0.0025	0.002	0.002	0.002	0.002	0
	Prec.-w-thr.	0.074	0.05	0.05	0.05	0.05	0
	Cov.-w-thr.	0.0025	0.002	0.002	0.002	0.002	0
Rec by group for CENG Web site (d)	Supp. thr.	2%	9%	16%	23%	30%	38%
	Prec.	0.3	0.34	0.1	0.09	0.1	0
	Cov.	0.5	0.2	0.08	0.05	0.02	0
	Prec.-w-thr.	0.4	0.35	0.1	0.09	0.1	0
	Cov.-w-thr.	0.58	0.2	0.08	0.05	0.02	0

This experiment is conducted under window count 3, confidence threshold 10%, precision threshold 30%, coverage threshold 30%

Table 3 part (b) displays the evaluation results in terms of Category concept. Since the number of individuals of the Category concept is less than those of URL addresses, the generated association rules have higher confidence and support counts. Therefore, higher support threshold values are used in this experiment. In Table 3 (b), coverage decreases with the increase in support threshold. This is an expected result due to the fact that when the support threshold increases, the number of generated patterns and, thus, the number of recommendations drops. The behavior in precision is more interesting. Up to a certain level (support threshold of 6%), the precision increases. This is due to the facts that, when support threshold increases, weaker association rules are pruned and more precise recommendations can be generated. However, after this breaking point, precision starts to decrease. This may be due to over-fitting that results in the generation of specific rules for the training data that fails to perform well enough for the test data. The resulting precision and coverage values are higher than the results shown in Table 3 part (a). As expectedly, using the category concept provides abstraction and leads to generation of more abstract rules. Such rules express the semantics of the navigation better, resulting with higher precision and coverage values.

Similar experiments are also conducted on CENG Web logs. Part (c) of Table 3 shows the evaluation results in terms of individual postings. Since semantic information does not contribute to the pattern structure in this experiment, the generated patterns and recommendations have low precision and coverage values. Part (d) of Table 3 shows the evaluation results in terms of Group concept. The coverage values show a smooth decrease similar to that of the e-store Category concept-based recommendation. Similar to part (b), at first, the precision value increases until a breakpoint value. Following this, there is a sharp decrease around support threshold of 14%. However, afterward, the precision value demonstrates a smooth behavior. As in the previous case, comparison between part (c) and part (d) Table 3

Table 4 Experiments on precision and coverage threshold change on the accuracy

Prec. thr.	0%	20%	40%	60%	80%	100%
Prec.-w-thr.	1	0.65	0.6	0.4	0.35	0.3
Cov. thr.	0%	20%	40%	60%	80%	100%
Cov.-w-thr.	1	0.5	0.4	0.3	0.28	0.28

This experiment is conducted by using category concept on e-store web site under window count 1, support threshold 0.45%, confidence threshold 10%

reveals that using a concept provides abstraction and improves pattern and recommendation quality.

4.2 Experiments on the effect of precision and coverage thresholds

The second set of experiments is for analyzing the new metrics, precision-with-threshold and coverage-with-threshold. The experiments in this group are performed under support threshold 0.45%, confidence threshold 10% and window count 1.

The first two rows of Table 4 shows the effect of change in precision threshold on precision-with-threshold value for patterns in terms of Category individuals under coverage threshold 30%. Since other metrics (precision, coverage, and coverage-with-threshold) are not effected from precision threshold, results on this metrics are not presented. Expectedly, when the threshold increases, the number of recommendations satisfying the given precision level decreases. However, this experiment shows that about 30% of the recommendations are 100% precise.

Similarly, the last two rows of Table 4 shows the effect of change in coverage threshold on coverage-with-threshold value for patterns in terms of Category individuals under precision threshold 30%. Expectedly, when the threshold increases, the number of recommendations satisfying the given coverage level decreases. As in precision-with-threshold results, in this experiment, about 30% of the user navigations are 100% covered by the generated recommendations.

4.3 Experiments on the effect of using all of the concepts

In the next group of experiments, the recommendation is generated by using the rules constructed in terms of all of the concepts. The combination of each concept under the same support threshold is not feasible. Therefore, support thresholds are proportionalized.

Part (a) of Table 5 displays the evaluation results under all concepts in the e-store Web site. In this experiment, precision and coverage behaviors are similar to the previous cases. However, two important improvements are observed in this experiment. The first one is that precision is more robust to the changes in the support threshold. The second improvement is on the noticeable increase in precision and coverage values.

Part (b) of Table 5 displays the evaluation results in terms of the combination of all concepts in the CENG Web site. As in part (a), this experiment shows an overall increase in coverage and precision when the amount of semantic information contribution increases. However, this increase is less than that given in part (a), since the patterns generated with Group concept already have the highest precision values among all concepts in this data set. Therefore, it appears to be the one that contributes most to the recommendation. These experiments show that the pattern and recommendation quality are improved when the combination of different aspects of the semantic information contributes to the pattern generation.

Table 5 Experiments on the effect of using patterns in terms of all concepts

Rec by all concepts for e-store Web site (a)	Supp. thr.	1.25%	2%	2.75%	3.5%	4.25%	5%
	Prec.	0.55	0.68	0.76	0.7	0.5	0.5
	Cov.	0.9	0.68	0.6	0.48	0.3	0.3
	Prec.-w-thr.	0.7	0.84	0.8	0.77	0.55	0.55
	Cov.-w-thr.	0.88	0.68	0.58	0.44	0.3	0.3
Rec by all concepts for CENG Web site (b)	Supp. thr.	2%	3.6%	5.2%	6.8%	7.6%	8%
	Prec.	0.35	0.35	0.35	0.35	0.35	0.35
	Cov.	0.52	0.4	0.35	0.28	0.22	0.2
	Prec.-w-thr.	0.44	0.4	0.4	0.39	0.37	0.37
	Cov.-w-thr.	0.58	0.48	0.4	0.3	0.25	0.2

This experiment is conducted under window count 3, confidence threshold 10% (for parts (a),(b)), support threshold 0.45% (for part (c)), precision threshold 30% and coverage threshold 30%

Table 6 Experiment with patterns in terms of category concept for e-store web site

Rule percent	20%	50%	70%	90%	100%
Precision	0	0.18	0.58	0.58	0.52
Coverage	0	0.1	0.24	0.32	0.36
Precision-w-threshold	0	0.18	0.58	0.64	0.54
Coverage-w-threshold	0	0.1	0.22	0.3	0.34

This experiment is conducted under window count 1, support threshold 0.45%, confidence threshold 10%, precision threshold 30% and coverage threshold 30%

4.4 Experiments on the effect of number of rules

Recommendation generation requires a scan of all generated rules in order to find the ones matching the user navigation. In some circumstances, there can be too many association rules (especially under low support and confidence thresholds) and their scan can require considerable time. In the last experiment, the effect of eliminating some of the association rules on the accuracy of the recommendation is tested.

Table 6 displays the effect of change in the percentage of rules considered in recommendation on precision, coverage, precision-with-threshold, and coverage-with-threshold, in terms of Category concept. As seen in the table, precision value increases when the association rule percentage increases. This is the expected result since the increase in the number of associations may increase the accurate recommendations as well. However, after a certain percentage, the precision starts to drop. The reason for this is that a further increase in the number of association rules leads to an increase in the number of inaccurate recommendations. Hence, the precision value starts to decrease.

It can be concluded from the above experiment that, if the time is limited for recommendation, then up to 30% of the association rules can be eliminated to reduce the time of the recommendation, and the recommendation accuracy will still be within the reasonable ranges.

5 Conclusion

The extraction of usage patterns is highly important for the Web site developers. The results can be used in many areas, such as generating Web page recommendation, product recommendation, content improvement, or page displacement. In this study, a technique for incor-

porating semantic information into frequent navigation pattern extraction is proposed and the effect of semantic information on pattern quality is elaborated. By introducing semantic information, Web usage mining patterns are generated in terms of ontology instances instead of Web page addresses. Such patterns can reflect the semantics of navigation behavior more explicitly and accurately.

The effect of semantic information on pattern quality is evaluated through a recommendation framework. Recommendations are generated by either considering the frequent navigation patterns in terms of a single concept or considering combination of frequent navigation patterns in terms of several concepts. Experimental results show that integrating semantic information provides abstraction that results in considerable improvement of pattern quality. In addition, this approach handles new item problem in recommendation [2]. Both single concept and the combined association rules have higher precision and coverage values than the classical Web usage mining (without the use of semantic information). The improvement is higher for combination of association rules, hence, we can deduce that, when the amount of contributing semantic information increases, the pattern quality increases as well. The analysis on the single concept patterns may be used for understanding the user's intent. The one that has the highest precision and coverage may reflect the user's intent for navigation. Another observation is that the increase in window count has a negligible effect on the precision and the coverage, hence most recent visit appears to be the most effective one on the recommendation. An interesting result concluded from the experiments is that, in order to speed up the recommendation generation, up to 30% of the rules can be eliminated with little decrease in the quality.

This work can be extended in several research directions. One possible direction is elaborating the use of semantically enhanced patterns for automatic or semi-automatic Web site adaptation and link restructuring. For the use of patterns in recommendation task, there are issues that can be further studied, as well. One of them is the analysis of the time efficiency. Pattern generation is an offline process, however, especially in a recommender application, recommendation generation is an online job whose time efficiency is an important factor for the success of the recommender system. We have observed that, on the average, recommendation generation lasts about a few seconds for a given navigation, however, this duration gets longer as the number of rules increases. A future work for this problem is finding techniques for setting a threshold and narrowing down the recommendation set.

Acknowledgments This work is supported by METU BAP-1 Grant number BAP-03-12-2009-01 and by TUBITAK Grant number TUBITAK-109E239.

References

1. Abraham A, Ramos V (2003) Web usage mining using artificial ant colony clustering and linear genetic programming. In: Proceedings of congress on evolutionary computation (CEC), pp 1384–1391
2. Adomavicius G, Tuzhilin E (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749
3. Baumgarten M, Buchner AG, Anand SS, Mulvenna MD, Hughes JG (2000) User-driven navigation pattern discovery from internet data. In: Proceedings of workshop on web usage analysis and user profiling, pp 74–91
4. Becchetti L, Colesanti U, Marchetti-Spaccamela A, Vitaletti A (2010) Recommending items in pervasive scenarios: models and experimental analysis. Knowl Inf Syst, September [Online]
5. Berendt B, Hotho A, Stumme G (2002) Towards semantic web mining. In: Proceedings of international semantic web conference (ISWC), pp 264–278. Springer, Berlin
6. Bezerra BLD, Carvalho FAT (2010) Symbolic data analysis tools for recommendation systems. Knowl Inf Syst, February [Online]

7. Borges JL, Levene M (2000) Data mining of user navigation patterns. In: Proceedings of workshop on web usage analysis and user profiling, pp 31–36
8. Britos P, Martinelli D, Merlino H, Garcia-Martinez R (2007) Web usage mining using self organized maps. Int J Comput Sci Netw Secur 7(6):45–50
9. Cho WC, Richards D (2004) Improvement of precision and recall for information retrieval in a narrow domain: Reuse of concepts by formal concept analysis. In: Proceedings of the 2004 IEEE/WIC/ACM international conference on web intelligence, pp 370–376, Washington, 2004. IEEE Computer Society
10. Cho WC, Richards D (2007) Ontology construction and concept reuse with formal concept analysis for improved web document retrieval. Web Intell Agent Syst 5(1):109–126
11. Dai H, Mobasher B (2005) Integrating semantic knowledge with web usage mining for personalization. In: Scime A, (eds) Web mining: applications and techniques. IRM Press, Idea Group Publishing
12. Daoud M, Lechani L, Boughanem M (2009) Towards a graph-based user profile modeling for a session-based personalized search. Knowl Inf Syst 21(3):365–398
13. Hay B, Wets G, Vanhoof K (2004) Mining navigation patterns using a sequence alignment method. Knowl Inf Syst 6(2):150–163
14. Masseglia l, Teisseire M, Poncelet P (2003) HDM: a client/server/engine architecture for real-time web usage mining. Knowl Inf Syst 5(4):439–465
15. Leung CW, Chan SC, Chung F (2006) A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. Knowl Inf Syst 10(3):357–381
16. Mabroukeh NR, Ezeife CI (2009) Using domain ontology for semantic web usage mining and next page prediction. In: Proceedings of conference on information and knowledge management (CIKM), pp 1677–1680
17. Missaoui R, Valtchev P, Djeraba C, Adda M (2007) Toward recommendation based on ontology-powered web-usage mining. IEEE Internet Comput 11(4):45–52
18. Mobasher B, Cooley R, Srivastava J (2000) Automatic personalization based on web usage mining. Commun ACM 43(8):142–151
19. Mobasher B, Dai H, Luo T, Sun Y, Zhu J (2000) Integrating web usage and content mining for more effective personalization. In: Proceedings of international conference on E-Commerce and web technologies (ECWeb2000), pp 165–176, Greenwich, UK
20. Nakagawa M, Mobasher B (2003) Impact of site characteristics on recommendation models based on association rules and sequential patterns. In: Proceedings of IJCAI'03 workshop on intelligent techniques for web personalization, Acapulco, Mexico, August
21. Nasraoui O, Soliman M, Saka E, Badia A, Germain R (2008) A web usage mining framework for mining evolving user profiles in dynamic web sites. IEEE Trans Knowl Data Eng 20(2):202–215
22. Park S, Suresh NC, Jeong B (2008) Sequence-based clustering for web usage mining: a new experimental framework and ann-enhanced k- means algorithm. Data Knowl Eng 65(3):512–543
23. Richards D (2004) Addressing the ontology acquisition bottleneck through reverse ontological engineering. Knowl Inf Syst 6(4):402–427
24. Rohn E (2010) Generational analysis of variety in data structures: impact on automatic data integration and on the semantic web. Knowl Inf Syst 24(2):283–304
25. Salin S, Senkul P (2009) Using semantic information for web usage mining based recommendation. In: Proceedings of international symposium on computer and information sciences (ISCIS 09), pp 236–241, September
26. Shchekotykhin K, Jannach D, Friedrich G (2009) xCrawl: A high-recall crawling method for web mining. Knowl Inf Syst, November [Online]
27. Shyu M, Haruechaiyasak C, Chen S (2006) Mining user access patterns with traversal constraint for predicting web page requests. Knowl Inf Syst 10(4):515–528
28. Spiliopoulou M (1999) The laborious way from data mining to web mining. Int J Comput Syst Sci Eng 14:113–126
29. Spiliopoulou M, Pohle C (2001) Data mining for measuring and improving the success of web sites. Data Mining Knowl Discov 5(1–2):14–85
30. Spiliopoulou M, Pohle C, Teltzrow M (2002) Modelling and mining web site usage strategies. In: Proceedings of multi-konferenz wirtschaftsinformatik, September
31. Srikant R, Agrawal R (1995) Mining generalized association rules. In: Proceedings of international conference on very large databases (VLDB), pp 407–419, Zurich, Switzerland, September
32. Srikant R, Agrawal R (1996) Mining sequential patterns: Generalizations and performance improvements. In: Proceedings of international conference on extending database technology (EDBT) pp 3–17, France, March
33. Stumme G, Berendt B, Hotho A (2002) Usage mining for and on the semantic web: next generation data mining. In: Proceedings of NSF workshop, pp 77–86, Baltimore, November

34. Stumme G, Hotho A, Berendt B (2006) Semantic web mining: state of the art and future directions. J Web Semant Sci Serv Agents World Wide Web 4(2):124–143
35. Tan A, Ong H, Pan H, Ng J, Qiu-Xiang Li (2006) Towards personalised web intelligence. Knowl Inf Syst 6(5):595–616
36. Vucetic S, Obradovaic Z (2005) Collaborative filtering using a regression-based approach. Knowl Inf Syst 7(1):1–22
37. Wang W, Zaine OR (2002) Clustering web sessions by sequence alignment. In: Proc. of the 13th International Workshop on Database and Expert Systems Applications (DEXA 2002), pp 394–398, Aix-en-Provence
38. World Wide Web Consortium (W3C). Web ontology language(OWL)
39. Yilmaz H, Senkul P (2010) Using ontology and sequence information for extracting behavior patterns from web navigation logs. In: Proceedings of IEEE ICDM workshop on semantic aspects in data mining (SADM'10), pp 549–556, December
40. Zaki MJ (2000) Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12(3):372–390
41. Zhou B, Hui SC, Fong ACM (2005) Web usage mining for semantic web personalization. In: Proceedings of workshop on personalization on the semantic web (PerSWeb), pp 66–72, July

Author Biographies

Pinar Senkul is currently Assistant Professor in Computer Engineering Department, Middle East Technical University (METU). She received her Ph.D. from the same department in 2003. She worked as a visiting researcher in State University of New York (SUNY) at Stony Brook. Her research interests include data mining, Web usage mining, semantic Web services, Web service discovery and composition, workflow modeling, and analysis. She has been serving as editorial board member of several journals, and she has been involving in the organization of various international conferences.

Suleyman Salin received his B.S. degree in Bilkent University, Ankara, Turkey in 2005. He received M.S. degree from Computer Engineering Department at Middle East Technical University in 2009. He is currently a Ph.D. student at Computer Engineering Department, METU. His research interests include Web usage mining, next page prediction, recommendation, and social networks.