ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Semantic Web Usage Mining Techniques for Predicting Users' Navigation Requests

S.Kaviarasan[1], K.Hemapriya[2], K.Gopinath[3]

Assistant Professor, Department of CSE, Panimalar Institute of Technology, Chennai, India

ABSTRACT: The explosive growth of the World Wide Web (WWW) has resulted in intricate Web sites, demanding for tools and methods to complement user skills in the task of searching for the desired information. In this context Web usage mining techniques have been developed for the discovery and analysis of frequent navigation patterns from Web server logs, which can be used as input for recommendation engines. Web usage mining techniques have been associated with Web content mining approaches in order to increase the accuracy of recommendation mechanisms. Existing approaches represent Web pages' content essentially by means of keywords, N-grams or ontologies of concepts, being, therefore, incapable of capturing the semantic information and the relationships among pages at the semantic level. Herein, we propose a method that combines usage patterns extracted from server logs with detailed semantic data that characterizes the content of the corresponding pages. Thus, a method to extract and analyze frequent semantic navigation patterns which are fed into a recommendation engine is proposed. We argue that by integrating usage and Web pages' detailed semantic information in the personalization process we will be able to increase the recommendation accuracy. The proposed method is an example of semantic Web mining that combines two fast developing research areas; Semantic Web and Web Usage Mining. We conducted an extensive experimental evaluation that provides strong evidence that the recommendation accuracy increases with the integration of semantic and usage data. The results show that the proposed method is able to achieve 15-17% better accuracy than a usage based model, 5-7% better than a N-gram based model and 4-6% better than a ontology based model. Also the proposed method is able to address the new item problem of solely usage based techniques by augmenting navigation patterns with newly added pages in a Web site.

KEYWORDS: Semantic Web Usage Mining, Web Usage Mining, Recommendation, Prediction

I. INTRODUCTION

Web mining is the application of data mining techniques to discover useful patterns from the Web, and it is usually divided into three general categories: Web content mining, Web structure mining and Web usage mining [1]. This research field became very important due to the rapid growth rate of the Web, development of e-commerce, Web services, and Web-based information systems. The field of Web usage mining focuses on developing techniques to model and study users' Web navigation data [2]. The task of modeling and predicting users' navigational behavior on a Web site is useful to many Web applications such as Web caching, Web page recommendation, Web search engines and personalization. According to [3], most Web usage mining techniques are based on association rules, sequential patterns and clustering. Up to now, many research efforts have tried to incorporate Web page content into Web usage mining and personalization techniques, but very few have performed this using detailed semantic data inferred by means of Semantic Web Technologies [4].

Inspired by the work of Haibin Liu *et al.* [5], we propose a method to combine Web usage patterns and semantic data characterizing the Web pages' content in order to model and analyze semantic navigation patterns. The semantic navigation patterns are then used as input for a recommendation engine. More precisely, we propose to extend the technique described in [5], which was devised to take into account the Web page content by means of a N-gram representation, in a way that takes into account the semantic of the Web pages. We argue that models that represent the content of web pages by means of N-grams or ontology are incapable of capturing the semantics of Web pages being, therefore, incapable of capturing relationships among theWeb pages at the semantic level.[7] reveal that the inclusion of detailed semantic data instead of ontology enhances the recommendation accuracy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

II. RELATED WORK

Several models have been proposed for modeling user browsing behavior on a Web site and generating recommendations for its users. Most approaches used consist in models based on solely usage data or models that combine usage data with Web site pages' content or its structure data.

Many approaches also differ in the method used to represent Web pages' content, being common to represent it by N-grams, keywords or by ontologies of concepts.

2.1 Approaches Based on Traditional Content Mining

Jin *et al*. [8] proposed a unified framework which provides dynamic and personalized recommendations taking into account both usage data and Web site content. Guo*et al*. [9] investigated an approach for combined mining of both Web logs and Web page content. J. Li and O.R. Zaïane [10] proposed a Web recommendation system that generates navigationalmodels by combining usage, content, and structure data in a Web site. Haibin Liu and VladoKešelj [5] proposed a novel approach for classifying navigation patterns and predicting users' future requests. The approach is based on a combined mining of Web server logs and content of the Web pages. The major limitation of this model is the N-gram representation of Web pages content does not take into account its semantics. Miao Wan *et al.* [11] proposed a Random Indexing approach that is based on a vector space model, to discover intrinsic characteristics of Web users' activities.

2.2 Approaches Based on Ontology

Honghua Dai and BamshadMobasher [12] explore various approaches for integrating semantic knowledge into the personalization process that are based on integration of domain ontologies and usage patterns. Eirinaki*et al*. [7] presented a semantic Web personalization framework that combines usage data and Web content (annotated in terms of ontology) in order to generate useful recommendations. Stuart Middleton *et al.* [13] presented a recommender system for online academic publications where user profiling is done based on a research paper topic ontology. Mehdi Adda*et al*. [14] use sequence association rules as a pattern structure and incorporate semantic information in terms of ontology into the pattern generation process. OlfaNasraoui*et al*. [15] presented a complete framework and findings inmining Web usage patterns that has all the challenging aspects of real-life Web usage mining, including evolving user profiles and external data describing an ontology of the Web content. P. K. Bhowmick*et al*. [16] proposed an ontology based user profiling strategy to capture the shift in the information needs of the users. HakanYilmazandPinarSenkul [17] proposed a navigation behavior extraction approach that makes use of page semantics and navigation sequence information. Pinar Senkul and SuleymanSalin [18] proposed a technique for integrating semantic information into Web navigation pattern generation process. The frequent navigational patterns are composed of ontology instances instead of Web page addresses and these are used for generating recommendations. ThiThanh Sang Nguyen *etal.* [19] proposed a novel ontology-style model of Web usagemining that enables integration of Web usage data and domain knowledge. The recommendations are generated by using Web user access sequences that are represented in

Web Ontology Language (OWL). Juan D. Velásquez*et al*. [20] proposed a methodology for identifying Website Key Objects. Website Key Objects are used to find the desired information.MehdiAdda*et al.* [21] studied a ontology based pattern space and proposed the mining method xPminer that performs a complete and non-redundant traversal of the pattern space and discovers all frequent patterns. The mined frequent patterns are used to generate recommendations.

2.3 Other Approaches

Patricia Kearney *et al*. [22] investigated how Web visitor usage data may be combined with semantic domain knowledge to provide a deeper understanding of user behavior. BlažFortuna*et al.* [23] proposed an approach to build flexible, comprehensive and scalable user model that takes into account usage data, page content and users' registration information. Fabian Abel *et al* [24] presented a framework for mining of users profiles from usage data and semantic enrichment of user profiles on the Social Web. The user data is aggregated and linked with data from social Web systems and RDF data from Linked Data services. A. C. M. Fong *etal.* [25] proposed a semantic Web usage mining

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

approach fordiscovering periodic Web access patterns from annotated Web usage logs. Hemant Kumar Singh and Brijendra Singh [26] proposed an efficient clustering algorithm based on disjoint sets to cluster user navigation patterns. The method is based on solely usage data and can be extended to take into page content and site structure.

2.4 Summary and Discussion

In summary, the works referred attempt to improve recommendation accuracy by integrating usage data with Web site structure or with Web page content represented by means of keywords, N-grams or ontologies of concepts. Such methods of characterization do not take into account the semantics in Web pages' content being unable to include in the recommendation set pages having semantically relevant content. We believe that more effective recommendations can be generated by incorporating detailed semantic data in the personalization process. Thus, we propose a method that combines usage and semantic data in order to induce semantic usage patterns that are fed to a personalization mechanism.

To the best of our knowledge, this is the first work proposing the use of detailed semantic metadata inferred from Web pages and represented using semantic Web technology in the process of recommendation generation.

III. PROPOSED METHOD

In this work, we propose to integrate semantic and usage data in order to more accurately predict users' future requests and generate better recommendations. We make use of semantic metadata generated by semantic annotation of Web pages' content and expressed using Resource DescriptionFramework (RDF). Figure 1 denotes the overall architecture of the proposed method which is detailed in the following subsections.

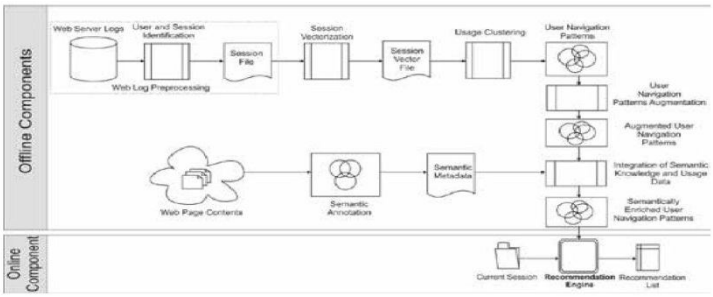


Fig. 1 : The structure of the proposed methodology

3.1 Web Log Pre-processing

As shown in Figure 1, the pre-processing task is the first step in Web usage mining, being responsible for reading the Web logs and inducing the corresponding user Web navigation sessions. In the process, Web log data is cleaned in order to remove entries that are not useful to represent user Web navigation behavior and for repairing erroneous data.

Also, users are identified based on information available in the log file, such as IP address, type of operating system and browsing software. The proposed system makes use of Web log pre-processing techniques described in [27].

3.2 Session Vectorization

The second phase of the method makes use of the session vectorization technique described in [5]. User navigation sessions are transformed into a multidimensional space of vectors of Web pages in order to facilitate partitioning these sessions into groups of similar sessions. Let P be a set of pages accessed by all Web users, $P=\{p1,p2,p3,...,pm\}$ with

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

each page uniquely represented by its associated URL and *S* be a set of all user access sessions, *S*={*s*1,*s*2,...,*sn*}. Each session *s* is represented by a *m* dimensional vector, *s*={*w*(*p*1,*s*), *w*(*p*2,*s*),..., *w*(*pm*,*s*)}, where w(*pi*,*s*) is a weight assigned to the*i*thWebpage ($1 \leq i \leq m$) visited in a session *s*. The weights can be determined in a number of ways, for example, binary weights can be used to represent existence or non-existence of a Web page access in a given session. Alternatively, weights can be seen as a measure of user degree of interest on a Web page.

3.3 Session Clustering and Navigation Patterns Generation

For generating the navigation patterns we make use of a clustering algorithm that allocates sessions into groups of similar sessions according to a similarity measure. In the proposed method the K-means algorithm available in the WEKA machine learning toolkit[3] is used. As a result we obtain a set of clusters, *C*={*c*1,*c*2,...,*ck*), in which each cluster *ci* (1≤*i*≤*k*) is a subset of user sessions *S*, and *k* is the number of clusters. The number of cluster is a parameter of the method.

Each cluster $c \in C$is represented by its mean vector *mc* that is computed by the ratio between the sum of the weights of the pages in the sessions of the cluster and the total number of sessions in that cluster. A weight threshold *Wmin* is used to filter out Web pages having a mean value below the threshold in the cluster. Each mean vector represents the users' navigation pattern of a cluster [5]. The set of users' navigation patterns represents common browsing characteristics among a group of users and are represented as, *NP* = {*np*1, *np*2,...,*npk*}, in which each *npi* is a subset of *P*, a set of Web pages accessed in the Web log.

3.4 Semantic Annotation

Semantic Web provides a common framework that allows data to be shared and reused across applications, and enterprises, in a manner understandable by machines. Semantic annotation is a key component for the realization of the Semantic Web that formally identifies concepts and relations between concepts in the documents. Our method makes use of the OpenCalais and the AlchemyAPI Web services for generating the semantic annotation, which includes topics, social tags, concept tags, keywords, search terms and other metadata.

The system crawls a Web site to collect the Web pages.OpenCalais processes these pages and returns annotated semantic metadata as RDF payloads serialized as XML data containing the topics that the content discusses, the identified entities, facts, events, and social tags. The semantic metadata induced by OpenCalais also includes a feature called social tags which is an emulation of how a human being would tag specific pieces of content. The metadata also contains relations that involve at least one recognized entity from the content. Relations are generally all subject predicate-object relationships without predefining their types. Web pages are also processed by Alchemy API to generate complementary semantic metadata. AlchemyAPI utilizes statistical algorithms, natural language processing technology and machine learning algorithms to analyze Web page content and extract keywords, search terms, concept tags, and information about people, places, companies, topics, languages and more. The AlchemyAPI has a concept tagging feature that automatically tags documents and text in a manner similar to human-based tagging. The results are also returned as RDF payloads.

The resulting XML data is parsed to extract semantic metadata and store it in the RDF data store. We make use of AllegroGraph RDF Data Store6, which is a modern, high-performance, persistent RDF graph database.
The semantic metadata information is used to calculate the semantic similarity between pages using the method described in [28], which returns a value between 0 and 1. A similarity score of 1 means that the instances have exactly the same properties and of 0 that there are no shared properties. In our method, if the similarity score is above a certain threshold it means that the Web pages are considered to be semantically similar. This similarity information is used to augment the navigation patterns with pages that are highly similar to the pages in the navigation patterns, but are not present in the log files (newly added pages in the Web site) as discussed in Section 3.5.

3.5 Navigation Patterns Augmentation

As discussed in Section 3.3, navigation patterns generated by clustering user sessions consist of Web pages identified

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

in the log files. Thus, recommendations based solely on usage data will not suggest pages for which requests are not available in the log files. This is commonly referred as the new item problem.
In the semantic annotation step all pages in the Web site are processed. Thus, in order to address the new item problem, each navigation pattern is augmented with new pages that are highly semantically similar to the pages composing it, that is, included in the cluster. The augmented navigation patterns are represented as $ANP = \{anp1, anp2, ..., anpk\}$, where k is the number of navigations patterns. These new pages areassigned a weight which equals the average weight of the pages in the corresponding pattern. As a result the method is able to address the new item problem.

3.7 Recommendation Engine Using Semantic Navigation Patterns

As stated in [29], "Web recommendation is a promising technology that attempts to predict interests of Web users, by providing users information and/or services that they need without users explicitly asking for them".

The recommendation engine is the online component of a recommender system. In our method, the active user session is classified into a semantic navigation pattern, *vi, vi ∈ PS, where i=1,2,...,k, that better describes the user information goals. Recommendations are then generated from a set of pages in the corresponding augmented navigation pattern*

*anpi.*The current active user session is represented by a vector *tfp*using the technique described in Section 3.6. In order toclassify the active user session *p* into the adequate semantic navigation pattern *vi*, a dissimilarity metric $D(p,vi)$, $i=1,2,...,k$ is calculated that assesses the similarity between the active session and each of the patterns. The active user session is associated to the pattern corresponding to the smallest value

*ofD(*The*p,vi)*accuracy with which a session is classified into a pattern depends on a clever choice of dissimilarity measure.

In the proposed method we use three dissimilarity measures $d1$, $d2$, and $d3$presented in [30]. Let*tfp*be vector representationof current active user session and *tfvi* be vector representation of navigation pattern *vi*. The dissimilarity measures are given by,

$$\mathbf{tf, tf} = \frac{2 \times \mathbf{tf}\ x - \mathbf{tf}\ x}{\mathbf{tf}\ x + \mathbf{tf}\ x} \quad (7)$$

$$\mathbf{tf, tf} = \frac{2\ x\ \mathbf{tf}\ x - \mathbf{tf}\ x}{\mathbf{tf}\ x + \mathbf{tf}\ x} \quad (8)$$

$$\mathbf{tf, tf} = \frac{\mathbf{tf}\ x - \mathbf{tf}\ x}{\mathbf{tf}\ x \times \mathbf{tf}\ x + 1} \quad (9)$$

In fact, these measures only differ in their normalization schema. The dissimilarity measures $d1$ and $d2$ use arithmetic mean while d_3 uses geometric mean for normalization.
It is likely that pages unrequested by a user in the navigation pattern may be accessed by a Web user during his further interaction with a Web site. In order to determine which Web pages are to be recommended, the average weight of a page in the augmented navigation pattern *anpi* is used as a threshold. The recommendation set is generated from a set of Web pages in the augmented navigation pattern *anpi* whose average weight is above a certain threshold value, excluding the already visited pages. The method to calculate the average weight of a Web page in navigation pattern is discussed in Section 3.3.

IV. EXPERIMENTAL EVALUATION

This Section provides a detailed experimental evaluation of the proposed method. Section 4.1 presents a description of the data sets used, Section 4.2 a description of the evaluation metrics used and Section 4.3 the experimental results and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

its discussion.

4.1 Data Sets Description

For the experimental evaluation of the proposed method it is necessary that data sets provide both server log data and the content of the Web pages. The experiments have been conducted on the publicly available Music Machine data set (DS-1), on data from the Semantic Web dog food Web site (DS-2), and on a synthetic usage data generated for a university Web site (DS-3). The DS-1 data set is provided cleaned and sessionized, and we have used access entries in a four month period, from January to April 1999. For DS-2, which corresponds to the Semantic Web Dog Food Web site, we have used access entries from June 2010 to December

2010. This is a very active Web site of publications, people and organizations in the Web and semantic Web fields, covering several of major conferences and workshops. Finally, DS-3 corresponds to a Web site of a technical university including

Web pages of students and teachers, news group and courses, for which usage data was generated using a technique similar to the described in [31].

Table 1 depicts summary statistics of the data sets.

For each data set, we indicate the total number of: access entries; of clean access entries (after removing entries that are not useful to represent navigation behavior); of pages occurring in the log; of pages identified by the crawler; and of users identified. We also give the total number of sessions derived from each data set and the number of sessions of length greater than two; session length is measured by the corresponding number of requests. We assume that sessions having more than two requests are more suitable for the experiments since single page sessions do not carry sufficient information about the users' intention on the Web site.

Table 1 : Statistics of Experimental Dataset

Attributes	DS-1	DS-2	DS-3
Total access entries	936677	452192	1325198
Clean access entries	936677	430252	1325198
Pages accessed in log	850	1919	835
Pages identified by Crawler	1037	2105	1050
Different access users	116183	23245	50000
Total identified sessions	143633	26667	50000
Total identified sessions (≥2 requests)	91926	21067	49040

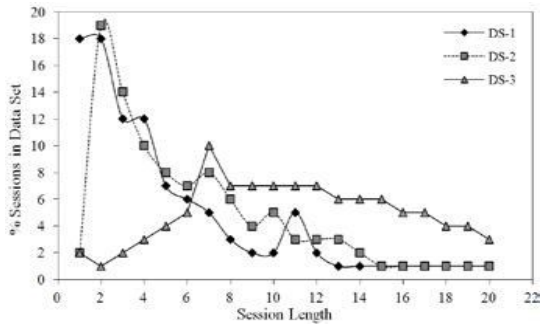
Figure 2 shows the distribution of session length forthe three data sets. For example, session length of two indicates the percentage of sessions with two page requests that occur in the collection of sessions. As shown in Figure 2, the percentage of sessions decreases when session length increases.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015



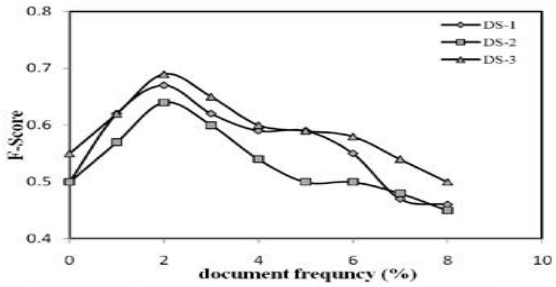


Fig. 2 : Sessions distribution of the data sets *Fig. 3 : Accuracy of Recommendation vs. document frequency for dissimilarity measure d3*

In order to evaluate the effectiveness of the recommendations generated the performance was measured by means of three metrics, namely precision, coverage, and the F-Score [32]. Among these, precision and coverage have been widely used in recommender system research. As precision and coverage are inversely related, a combination measure called F-Score, which gives equal weight to both precision and coverage, is used. While, precision measures the degree to which the recommendation engine produces accurate recommendations, coverage measures the ability of the recommendation engine to provide all the pages that are likely to be visited by a user.

We conducted a series of experiments focused on evaluating the accuracy of the proposed method (*M*4) and compared it to a method based solely on usage data (*M*1), to a method based on both usage and pages' content data (*M*2), and a method that integrates usage data with content semantics expressed in ontology terms (*M*3). The method (*M*1) is the Profile Aggregations based on Clustering Transactions (PACT) method [6] and is based solely on usage data (*M*1). In the PACT method, sessions are clustered and aggregated profiles based on each cluster are generated. These aggregated profiles are used to generate recommendations to a user. The method (*M*2) combines usage data and pages content, represented using N-grams, and makes use of the approach presented in [5]. The method *M*3 combines usage data and page content modeled using an ontology of concepts [7]. We stress that our method has the advantage of taking into account detailed semantics of pages' content, which is integrated with usage data.Cross validation with k=5 subsets was used, being the sessions split into k subsets, the model built from k-1 subsets, leaving the k^{th} subset as a test set. In order to simulate active user sessions, each test session is split into two parts. The first part of the session simulates an active session and thesecond part the pages that the user will request during his further navigation. That is, the first part of the session is used to predict its second part. Each active session is then fed into the recommendation engine in order to produce a recommendation set. The recommendation set obtained is then compared to the second part of the test session in order to compute the precision, coverage, and F-Score metrics.

Two sets of experiments were conducted. First, the parameters (document frequency and number of clusters) of the proposed method were varied in order to tune their values for best performance. In the second set of experiments our method is compared to the other approaches. We will now discuss results of the first experiment. Figure 3 illustrates the variation of the F-Score with document frequency of semantic metadata items for the three data sets. The concept of document frequency of semantic metadata items is discussed in Section 3.6 and is used to filter out semantic metadataitems that are perceived less important. The value of document frequency was made to vary between 0% and 8%. A value of 0% indicates that all the semantic metadata items are taken into account irrespective of their document frequency and a value 5% denotes that items present in less than 5% of the pages are filtered out. A very small value of document frequency will lead to semantic navigation patterns containing more semantic metadata items probably leading to too general users' profiles, which may not be able to accurately capture users' information goals. On the other hand semantic navigation patterns containing few semantic metadata items may not be adequate to predict users' further navigational behavior. The results in Figure 3 show that the recommendation accuracy changes with the value of document frequency parameter. It is clear that the best accuracy is achieved for document frequency of 2% for the three data sets. Values of document frequency below 2% lead to general user profiles, which are not suitable for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

generating accurate recommendations. Also decrease in recommendation accuracy for values of document frequency higher than 2% is due to inadequate number of semanticmetadata items in the user profile.

Figure 4 shows the effect of the number of clusters on the accuracy of the proposed model. The number of clusters was made to vary from 10 to 52. The results show that maximum recommendation accuracy is attained for the range between25 and 35 clusters. The optimal number of clusters depends on the data set being used. For higher values of number of clusters it was observed that many clusters are very small and contain less than 1% of the total training sessions and that patterns generated from such clusters do not accurately represent navigational behavior of a Web user leading to low recommendation accuracy.

We will now discuss results of the second set of experiments focused on comparing the accuracy of the proposed method with the three competing methods proposed in [6,5,7].

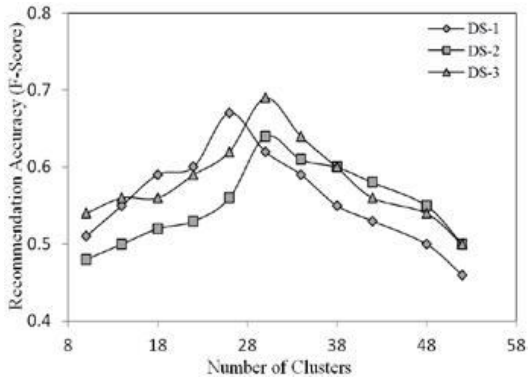


Fig. 4 : Accuracy of Recommendation vs. Number of Clusters for dissimilarity measure d_3

Table 2 shows the values of the three metrics for each of the four methods *M*1, *M*2, *M*3, and *M*4. For the comparative analysis the parameter settings that provide best performance for each of the methods was chosen. In the case of *M*1, the recommendation threshold value was set to 0.6. For *M*2, we have set the value of document frequency to 10% and the N-gram size to 6. For our method, *M*4, the document frequency was set to 2%. In addition, the number of clusters was set independently for each data set and the size of recommendation set was set to ten.

The proposed method (*M*4) and the approach in [5] (*M*2) have been evaluated for the three distance metrics *d*1, *d*2, and *d*3. The concept of distance metric is not applicable for *M*1and *M*3. It is observed from the figures in Table 2 that the dissimilarity metric *d*3 that is based on the geometric mean outperforms the other dissimilarity metrics. The geometric mean is more sensitive to differences on smaller values. In our example we have used normalized frequency to represent semantic navigation patterns and current active user session. These frequency values are very small and due to sensitivity of geometric mean for small values we are getting betterperformance for dissimilarity measure based on geometric mean. The results in Table 2 reveal that the proposed method *M*4is able to achieve higher accuracy than the competingmethods *M*1, *M*2 and *M*3. In fact, the inclusion of semantic data results in an increase of 15-17% in accuracy relatively to using solely usage data. Also, an increase of 5-7% in accuracy is obtained over the approach based on using page content represented by N-grams and a 4-6% improvement compared to the ontology based method. The proposed method outperforms the approaches presented in [5,7], being a clear indication that inclusion of detailed semantic data, instead of N-grams or ontology, improves accuracy of the recommendations generated. The better performance of the proposed method over the ontology based method is explained by the semantic metadata that is extracted from page content enabling to induce a more realistic model of user behavior than ontological model. It is observed from Table 2 that results are consistent for the three data sets in the sense that the proposed method outperforms the other methods for the three data sets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Data Set	Dissimilarity Measure	Precision				Coverage				F-Score			
		M_1	M_2	M_3	M_4	M_1	M_2	M_3	M_4	M_1	M_2	M_3	M_4
DS-1	d_1	0.46	0.55	0.62	0.59	0.50	0.58	0.64	0.63	0.48	0.56	0.63	0.61
	d_2		0.56		0.61		0.59		0.64		0.57		0.62
	d_3		0.61		0.65		0.63		0.69		0.61		0.67
DS-2	d_1	0.45	0.55	0.58	0.59	0.51	0.59	0.63	0.63	0.47	0.56	0.60	0.61
	d_2		0.55		0.59		0.58		0.63		0.56		0.60
	d_3		0.58		0.62		0.62		0.66		0.59		0.64
DS-3	d_1	0.50	0.57	0.61	0.62	0.53	0.59	0.66	0.64	0.51	0.57	0.63	0.63
	d_2		0.58		0.63		0.61		0.67		0.59		0.65
	d_3		0.62		0.68		0.65		0.71		0.63		0.69

Table 2 : Results of Recommendation Engine for three Data Sets DS-1, DS-2, and DS-3

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we show the benefits of using semantic metadata items together with usage data for building recommender systems. Previous methods were based on usage data alone or on usage data combined with a representation of pages' content by means of keywords, n-grams, or ontologies of concepts. We argue that by using semantic metadata it is possible to take into account meaning and semantic of pages' content.

A novel method is presented to predict users' future requests by combining usage data and detailed semantic information extracted from page content. The semantic metadata items are extracted by means of natural language processing and text mining techniques and are expressed in a Resource Description Framework (RDF). The semantic metadata is then combined with usage data to generate semantic navigation patterns, which areused in the personalization process. Three dissimilarity measures are used to classify a user session into one of the semantically enriched navigation patterns and then generate recommendations.

Results of an extensive experimental evaluation conducted on three data sets are reported. The experimental results show that by integrating semantic Web and usage mining, we are able to generate a more accurate classification of navigation patterns, and leads to more accurate recommendations. The experiments show that the proposed method is able to outperform a solely usage based method, a method that combines usage data and Web page contents represented in terms of N-grams, and a method based on an ontology of concepts to represent the contents.

As future work we refer that there are some aspects in which the proposed method can be improved. The method can be extended in order to take into account Web site structure.The proposed method can also be extended for a database backed Web site that generates Web pages dynamically based on structured queries performed against backend databases.The content of Web page depends on query parameters, hence these parameters must be taken into account in the personalization process.

REFERENCES

[1] O. R. Zaiane, "Resource and Knowledge Discovery from the Internet and Multimedia Repositories," Ph.D. Thesis 1999.
[2] Robert Cooley, BamshadMobasher, and JaideepSrivastava,
"Web mining: Information and pattern discovery on the World
Wide Web," in *IEEE international conference on tools withartificial intelligence, Newport Beach, CA, USA.*, 1997.
[3] Sungjune Park, Nallan Suresh, and Bong-KeunJeong, "Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm," *Data & Knowledge Engineering*, vol. 65, pp. 512-543, 2008.
[4] Bing Liu, *Web Data Mining*, Second Edition ed.: Springer, 2011.
[5] Haibin Liu and VladoKešelj, "Combined mining of Web server logs and Web contents for classifying user navigation patterns and predicting users' future requests," *Data & KnowledgeEngineering*, vol. 61, no. 2, pp. 304-330, 2007.
[6] BamshadMobasher, Honghua Dai, Tao Luo, and Miki Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," *Data Mining and KnowledgeDiscovery*, vol. 6, no. 1, pp. 61-82, Jan 2002.
[7] MagdaliniEirinaki, DimitriosMavroeidis, George Tsatsaronis, and MichalisVazirgiannis, "Introducing Semantics in Web Personalization: The Role of Ontologies," in *Proc. EWMF/KDO'2005*, 2005, pp. 147-162.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

[8] Xin Jin, Yanzan Zhou, and BamshadMobasher, "A Unified Approach to Personalization Based on Probabilistic Latent Semantic Models of Web Usage and Content," in *AAAI Workshopon Semantic Web Personalization (SWP'04)*, July 2004.
[9] J. GuoGuo, Kešelj V., and Q. Gao, "Integrating web content clustering into web log association rule mining," in *Proceedingsof Canadian AI'2005*, Victoria, BC, Canada, 2005.
[10] J. Li and O. R. Za"ane, "Combining usage, content, and structure data to improve web site recommendation," in *EC-Web*, 2004, pp. 305-315.
[11] Miao Wan, Arne Jönsson, Cong Wang, and Lixiang Li, "Web user clustering and Web prefetching using Random Indexing with weight functions," *Knowl Information Systems*, October 2011.
[12] Honghua Dai and BamshadMobasher, "A Road map to More Effective Web Personalization: Integrating Domain Knowledge with Web Usage Mining," in *Proceedings of the InternationalConference on Internet Computing*, Las Vegas, Nevada, 2003.
[13] Stuart Middleton, Nigel Shadbolt, and David Roure, "Ontological
User Profiling in Recommender Systems," ACM *Transactionson Information Systems*, vol. 22, no. 1, pp. 54-88, 2004.
[14] Mehdi Adda, PetkoValtchev, RokiaMissaoui, and ChabaneDjeraba, "Toward recommendation based on ontology-powered web-usage mining," *IEEE Internet Computing*, vol. 11, no. 4, pp. 45-52, 2007.