Deriving Semantic Sessions From Semantic Clusters

Banafsheh Safarkhani
Department of Computer Engineering
Islamic Azad University North Tehran Branch
Tehran,Iran
B_safarkhani@yahoo.com

Mojde Talabeigi
Department of Computer Engineering
Islamic Azad University Qazvin Branch
Tehran, Iran
mojde.talabeigi@gmail.com

Mehran Mohsenzadeh
Department of Compyter Engineering
Islamic Azad University Science & Research Center
Tehran , Iran
mohsenzadeh@sr.Iau.ac.ir

Mohammad Reza Meybodi
Department of Compyter Engineering
Amir Kabir University of Technology
Tehran , Iran
mmeybodi@aut.ac.ir

Abstract- **A important phase in any web personalization system is transaction identification. Recently a number of researches have been done to incorporate semantics of a web site in representation of transactions. Building a hierarchy of concepts manually is time consuming and expensive. In this paper we intend to address these shortcomings. Our contribution is that we introduce a mechanism to automatically improve the representation of the user in the website using a comprehensive lexical semantic resource and semantic clusters. We utilize Wikipedia, the largest encyclopedia to date, as a rich lexical resource to enhance the automatic construction of vector model representation of user sessions. We cluster web pages based on their content with Hierarchical Unsupervised Fuzzy Clustering algorithms ,are effective methods, for exploring the structure of complex real data where grouping of overlapping and vague elements is necessary. Entries in web server logs are used to identify users and visit sessions, while web page or resources in the site are clustered based on their content and their semantic. Theses clusters of web documents are used to scrutinize the discovered web sessions in order to identify what we call *sub-sessions*. Each sub-session have consistent goal. This process engendered to improving deriving semantic sessions from web site user page views. Our experiments show that proposed system significantly improves the quality of web personalization process.**

Keywords- Semantic vectors; Semantic sub-session; Semantic cluster; Wikipedia

I. INTRODUCTION

With the emerging growth of World Wide Web the task of finding pertinent information for internet users has become tedious. In order to overcome this problem, personalized systems have been introduced which adapt the information or services to their users using knowledge derived from their interaction with web site implicitly.

Web usage mining is a sub domain of knowledge discovery which employs data mining techniques to extract patterns from web server logs. It has been used effectively as an approach to automatic personalization [16]. However, the pure usage based personalization has a number of deficiencies. For example, there are not enough usage data for some navigational actions in a web site or for new pages and features that have been added to it and are not yet visited. There have been a number of research studies that incorporated website content in order to improve personalization process. Most of them extract features from page contents and these features are usually keywords, but keyword-based approaches are incapable of capturing more complex relationships among objects at a deeper semantic level based on the inherent properties associated with the objects.

Several research studies express user's behavior in terms of an ontology and use them to enhance the underlying model of users. All of them either use manually created concept hierarchies or employ existing resources like WordNet [8] or CYC [9] for this purpose. Creating knowledge bases and updating them manually is very expensive and time consuming. In addition, most of the existing ontologies and taxonomies are too small and do not contain domain specific information. In this paper we try to overcome these problems. The first contribution is introducing a comprehensive lexical semantic resource namely *Wikipedia* [7] in the area of web personalization. The second one is using a *fully automatic* mechanism for deriving semantic user models.

We first extract keywords from web site pages and map them to the appropriate concepts (articles) in Wikipedia. Then map user's page views to vectors of concepts and then clustering these web page concept vectors by HUFC algorithm and obtain the semantic clusters. Then by semantic clusters and web server logs, we create a novel notion, sub-session, which useful for improving the website user model. All of the above tasks are done automatically.

The rest of the paper is organized as follows. In section 2 we review the research efforts that utilize web site semantics for automatic personalization. In section 3 we briefly introduce Wikipedia and its structure. Section 4 presents the overview of the system and its components. Section 5 presents the HUFC algorithm and clustering the sessions. Section 6 include introduce and procreate sub-session. In section 7 we explain 4 experimental results

and finally in section 8 we present conclusion and future work.

II. RELATED WORK IN SEMANTIC PERSONALIZATION

Several research studies proposed frameworks that express the users' navigational behavior in terms of concepts of ontology and use it for automatic personalization [15]. Proposed a web personalization framework that characterizes the usage profiles of a collaborative filtering system using ontologies. Each page is mapped to a set of relevant objects of an ontology. Then, usage profiles are transformed to domain level aggregate profiles. It is assumed that mapping of pages to ontology objects is done manually or using a supervised method[3]. Uses a manually constructed ontology to derive semantic log files for tracking users in a portal. It proposed a framework where data mining can be performed on these semantic log files to extract knowledge about users [4]. Utilizes a topic hierarchy for markov model-based recommender system. It maps each visited page to a topic and builds a tree and then estimates the parameters of the markov model defined on this tree. The semantic characterization of the context is done manually [5]. Constructed the user model from word senses in WordNet. Documents are processed and relevant senses are extracted and then build a semantic network. Based on this network, new documents for users are predicted [6]. Utilizes a taxonomy and disambiguation techniques to create semantically enriched logs, namely C-logs from web server log files and then uses it for recommendation [2]. Uses a list of manually selected concepts from Wikipedia categories and URL structure of visited pages to extract an initial list of concepts. These concepts are then disambiguated using WordNet with terms extracted from page contents and eventually build the user model. Consider a news portal in which many new pages are added each day. Assume a web user that is interested in his/her favorite celebrity's news. The personalized portal must be able to rank this celebrity's news in higher positions in the main page for this user. WordNet based methods do not address this case since they lack many name entities. On the other hand, updating manually constructed taxonomies is costly and it needs time. None of the above approaches is suitable for this case because of its incomplete concept source or the difficulty to update.

III. WIKIPEDIA

Wikipedia is a free online multilingual encyclopedia constructed by volunteer contributors from all over the world. Today it excels all other encyclopedias in size and coverage and is one of the most visited sites on the web. With more than 10 million articles in more than 250 different languages, one third in English, it is almost ten times as big as the *Encyclopedia Britannica.* In fact, it was found to be similar in coverage and accuracy to *Britannica* [17]. Wikipedia consists of articles, separate pages for images, discussions about article contents,

authors and so on. The basic element in Wikipedia is an *article* which describes an entity or event and has links to other articles or other web pages. Each article has a unique title. In fact, each article can be regarded as a singular concept. Wikipedia has a *category system* that organizes all articles in a hierarchy. The category hierarchy is not a tree, as some categories have multiple parent categories. Synonymy, spelling variations and abbreviations of a term can be found in article *redirects* which consist of a redirection hyperlink from an alternative name to the article actually containing the description of the entry. Polysemy in Wikipedia is addressed via *disambiguation pages* which list all articles that exist for a certain term and clarify their meanings by adding a disambiguation tag in parentheses.

The advantage of Wikipedia is in its size and its acceptable coverage of name entities and domain specific terms, which is the main liability of general purpose ontologies like WordNet. Experiments also proved that it is comparable even with manually constructed domain specific taxonomies [11].

IV. EXTRACTING CONCEPT VECTORS

The semi architecture of our system appears in Figure 1. The main difference of it with previous works is its fully automatic concept extraction mechanism using a much more comprehensive lexical semantic resource. The resulting semantic model can be used as the input to various data mining techniques to enhance their effectiveness with semantic knowledge.

A. Web page concept extraction

As noted before, in this paper we treat each Wikipedia article as a unique concept. The reason is that each article concentrates on a specific event or entity and gives well-formed information about it and its related issues. The concept extraction component is responsible for detecting concepts from web site pages. It consists of two sub components for keyword extraction and word sense disambiguation. Keyword extraction module identifies the most important words and phrases in the page. The task of Word sense disambiguation module is ascribing the relevant concept to each keyword. After the process of concept extraction, each page of the web site is transformed into a vector of most pertinent Wikipedia articles, each having a weight expressing its importance in the page.

1) Keyword extraction component

Keywords here are words and phrases that are considered important for the page. These usually include technical terms, named entities, new terminology as well as other concepts closely related to the content of the page. *Keyword extraction* is the task of automatic identification of keywords in a page. There are a number of supervised and unsupervised methods for keyword extraction. Supervised methods employ machine learning models such as Naïve Bayes, decision trees, etc using
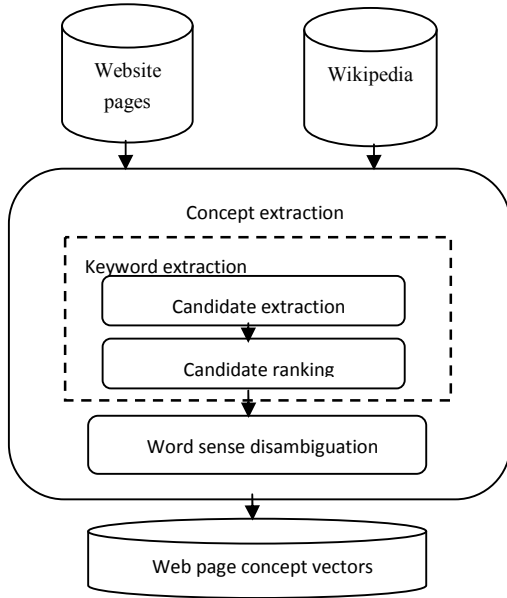
Figure. 1. Semi Architecture Of System

attributes such as syntactic features and others. Unsupervised methods have been also proved to be accurate for this task. There are a number of supervised and unsupervised methods for keyword extraction. Supervised methods employ machine learning models such as Naïve Bayes, decision trees, etc using attributes such as syntactic features and others. Unsupervised methods have been also proved to be accurate for this task. Our method for this task is mainly based on that of [1] which is an unsupervised technique that uses Wikipedia and can be considered as a state-of-the-art keyword extraction mechanism. Motivated by the Wikipedia guidelines for contributors, [1] constructs an initial vocabulary of all article titles in Wikipedia and then enriches it with *surface forms* which are various forms of article titles found in hyperlinks in Wikipedia articles. For example, consider this sentence:" *Shakespeare is buried in the [[chancel]] of [[Holy Trinity Church, Stratford-upon- Avon|Holy Trinity Church]] in [[Stratford-upon-Avon]].*".Here, *Holy Trinity Church* is a surface form for *Holy Trinity Church, Stratford-upon-Avon* article title. The resulting set is named *controlled vocabulary*.

The process of keyword extraction consists of two subtasks, namely *candidate generation* and *ranking*, respectively. The first task extracts from the given page, all n-grams that are also in the controlled vocabulary. The second task ranks them using several mechanisms. The ranking method which we selected is named *keyphrasesness* and measures the probability of a term to be selected as a keyword in Wikipedia and is obtained using equation (1)

$$keyphraseness(k) = \frac{num\ of\ articles\ containing\ k\ in\ link}{num\ of\ erticles\ containing\ k} \quad (1)$$

2) Word sense disambiguation

In order to map keywords to concepts (Wikipedia article titles) we must first disambiguate them in the context. Word sense disambiguation is the task of automatic assignment of the most appropriate sense to a polysemous word in the page.

There are mainly two approaches for this task. The first are knowledge based methods that rely on knowledge derived from dictionaries. The second are data driven methods that are based on probabilities collected from large amount of sense-annotated data.

We use Wikipedia disambiguation pages for this task. Similarity between a term and an article is obtained based on gloss overlap between the context of the ambiguous word and candidate articles using Lesk algorithm [10]. We use the current paragraph of the ambiguous word as a representation of its context and the first paragraph of candidate article page as its gloss For example in the sentence "*I sat on the bank of the lake*" when we face keyword *bank* we see a disambiguation page for *bank* in Wikipedia and candidate articles are *Bank(sea floor)*, *Bank Street(Ottawa)*,*Bank Station(OC Transpo)*, Looking at the first paragraph of each of these articles and using gloss overlap measure we finally find the most relevant meaning namely *Bank(sea floor)*.

B. Concept weightings

After the disambiguation phase each keyword in the page is mapped to its corresponding Wikipedia article. The importance of each concept in a page then is obtained using well known TF-IDF measure:

$$weight(page, conceot) = freq \times log_2 \frac{|W|}{df_{concept}} \quad (2)$$

freq is the frequency of *concept* in *page*. $|W|$ is total number of pages in the web site. $df_{concept}$ is the number of pages containing *concept*. As a result, each page P_i can be represented with a vector of concept-weight pairs:

$$P_i = \{ <c_{i1}, w_{i1}>, \ldots, <c_{in_i}, w_{in_i}> \} \quad (3)$$

V. UNSUPERVISED FUZZY CLUSTERING

To cluster web pages based on their content, we use Hierarchical Unsupervised Fuzzy Clustering (HUFC).Fuzzy clustering algorithms [20] are effective methods for exploring the structure of complex real data where grouping of overlapping and vague elements is necessary. Some experience has been accumulated in the medical field in diagnostics and decision-making support tools where a wide range of measurements were used as the input-data space and a decision result was produced by optimally grouping the symptoms together [20]. The algorithm does not require the number of clusters to discover as a constraint, but allows the definition of cluster sizes. This was the appealing property which made us select the algorithm. Indeed, we do not want either too

525

large or too small content cluster sizes. Very large clusters cannot help capture missions from sessions, while very small clusters may break potentially useful relations between pages in sessions [21].

The main part of the HUFC algorithm [20] is a recursive procedure HUFC(X,w) , where its inputs are a N×M data matrix X composed of M columns of data patterns $xj \in R, j = 1,...,M$, and a row vector $w \in R$ of M weights of each data pattern $wi = 1,...,M$ in the partitioning. The weight of each pattern is treated by the clustering algorithm as if wi patterns, which are equal to ith pattern xi, were included in the data matrix X. The final result of the HUFC algorithm is a global set Ug of Kg membership vectors uk, k = 1,..,Kg of all the data patterns in all the final Kg fuzzy clusters.

The HUFC algorithm is initiated by setting the global set Ug to be empty and the global number of clusters Kg to be zero and executed by calling HUFC(X0,w0) , where X0 is the matrix of the M0 original data patterns and w0 is a row vector of M0 ones. When the algorithm has terminated, Ug contains the final memberships of all the data patterns in all the Kg final clusters.

VI. DERIVING SEMANTIC SUB-SESSIONS

A. Log transformation

In order to derive semantic user model we must map web pages visited in a session to their corresponding concepts. Each session is a vector of page views and time durations:

$$S = \{<p_1, t_1>,...,<p_s, t_s>\} \qquad (4)$$

t_i is the amount of time which user spends on page

p_i . If we replace p_i with its concept vector, we obtain:

$$S = \{<\vec{c}_1, \vec{w}_1>,...,<\vec{c}_s, \vec{w}_s>\} \qquad (5)$$

Where

$$\vec{c}_i = (c_{i1}, c_{i2}, ..., c_{il_i})$$

$$\vec{w}_i = (w_{i1}, w_{i2}, ..., w_{il_i})$$

$$w_{ij} = \text{weight}(p_i, c_j) \times \frac{t_i}{s\_max} \qquad (6)$$

s_max is the maximum session duration [12]. In this paper we assume s_max=30 minutes. Equation (6) shows that the more time the user spends on a page, the more important are its concepts for him/her. The weight of each concept in the entire session is the sum of its weights in all of the concept vectors:

$$weight(s, c) = \sum_{\substack{i=1 \\ c_{ij}=c}}^{s} w_{ij} \qquad (7)$$

Example: suppose that our web site contains three pages namely p_1, p_2, p_3 . Assume a user session

$S = \{<p_1, 10>,<p_2, 8>,<p_3, 4>\}$,Table 1 shows the concept vectors for each page.

Taking page view durations into account, we obtain page-concept vectors for this session as shown in Table 2.

It should be noted that we do not have duration for the last page of each session and it should be estimated. The final session vector is:
S={<datastructures,0.55>,<tree,0.21>,<traversal,0.014>,<list,0.14>,<hash table,0.04>}

B. Visit sub-session Identification

Entries in web server log and semantic clusters are used to identify users and visit sessions, while web page or resources in the site are clustered based on their content and their semantic (see fig.2). Theses clusters of web documents are used to scrutinize the discovered web sessions in order to identify what we call sub-sessions. Each sub-session have consistent goal. Rather than dividing sessions into arbitrary transactions, we identify sub-sessions with coherent information needs. We assume that a visitor may have different information needs to fulfill during a visit, but we make no assumption on the sequence in which these needs are fulfilled. In the case of transactions in [19], it is assumed that one information need is fulfilled after the other. A *sub-session* would model a sub-session related to one of these information needs, and would allow overlap between sub-sessions, which would represent a concurrent search in the site [21].

Now how do we identify *sub-sessions*? The first approach we proposed to identify *sub-session* is based on web content [18]. While in the transaction-based model, pages are labeled as *content* pages and *auxiliary* pages, and a transaction is simply a sequence of auxiliary pages that ends with a content page, in the sub-session based model we propose, the identified sequence is based on the real content of pages. Indeed, a content page in the transaction-based model is identified simply based on the time spent on that page, or on backtracking in the visitor's navigation. We argue that sub-sessions could better model users' navigational behavior than transactions.

TABLE I. Page-concept vectors for assuming website

Page	Concept-weight vector
p_1	*<data structures,0.8>,<tree,0.62>,<traversal,0.43>*
p_2	*<data structures,0.73>,<list,0.54>*
p_3	*<data structures,0.75>,<hash table,0.36>*

TABLE II. page-concept vectors for the given session

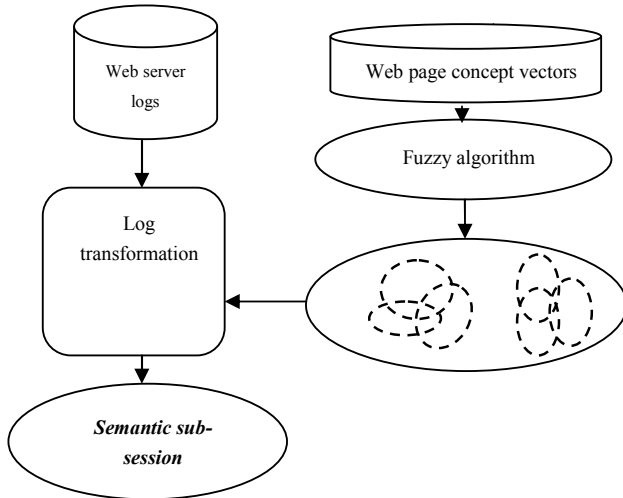| Page	Concept-weight vector
p_1	*<data structures,0.26>,<tree,0.21>,<traversal,0.14>*
p_2	*<data structures,0.19>,<list,0.14>*
p_3	*<data structures,0.1>,<hash table,0.04>*

Figure. 2. Overall System Architecture

In our model, users visit a web site with concurrent goals, i.e., different information needs. For example, a user could fulfill two goals in a visit session: *a, b, c, d*, in which pages *a* and *c* contribute to one goal, while pages *b* and *d* contribute to the other. Since pages related to a given goal in a visit session are generally supposed to be content coherent, whether they are neighboring each other or not, we use page content to identify sub-sessions within a visit session. All web site pages are clustered based on their content, and these clusters are used to identify content coherent clicks in a session. Let us give an example to illustrate this point. Suppose the text clustering algorithm groups web pages *a*, *b*, *c*, and *e*, web pages *a*, *b*, *c*, and *f*, and web pages *a*, *c* and *d* into three different content clusters (please note that our text clustering algorithm is a soft clustering one, which allows a web page to be clustered into several clusters). Then for a visit session: *a, b, c, d, e, f*, our system identifies three sub-sessions as follows: sub-session 1: (*a, b, c, e*); sub-session 2: (*a, b, c, f*); and sub-session 3: (*a, c, d*). As seen in this example, sub-session identification in our system is different from transaction identification in that we can group web pages into one sub-session even if they are not sequential in a visit session. We can see that our sub-session-based model subsumes the transaction-based model, since sub-sessions could become transactions if visitors fulfill their information needs sequentially.

VII. EXPERIMENTAL RESAULTS

In our experiments, we used JWPL, a java API developed by Torsten Zetch [13] for accessing Wikipedia. It can be downloaded from (http://www.ukp.tu-darmstadt.de/software/JWPL). It works with an optimized version of Wikipedia data downloaded from February 2007. To estimate the quality of web page semantic vectors and semantic sub-session, we define the following precision /recall measure: Precision indicates the accuracy of mapping, while Recall indicates the effectiveness of mapping.

We chose web logs of the Computer Engineering Department of Sharif University of Technology [14]. It includes nearly 6000 web pages about different courses. Web logs were collected for a one week period. The total web log size was 224883 hits. After sessionizing and keeping only sessions with length more than 2 and less than 10 pages similar to [2], 10408 sessions were remained. Out of these sessions, we randomly selected 100 sessions for our experiments. These web pages were mapped to concept vectors with the system. We also evaluated the system with WordNet. Two humans also mapped the web pages into corresponding concepts selected using searching *Computing* category in Wikipedia. The human-made vector was built by considering the concepts which were considered important by both humans. We determined different thresholds for the weights of concepts in the web pages and compared the resulting vectors with those which made by humans according to Precision, Recall and F-measure. Table 3 and Table 4 show the results for Wikipedia and WordNet respectively.

The precision and recall of the resulting semantic user model heavily depends on the precision and recall of the keyword extraction phase. As the concept weight threshold increases, the effectiveness of the resulting model also increases. The results are better than those obtained with WordNet. This is mainly due to using a more comprehensive concept source like Wikipedia compared to WordNet. Next, we used web page concept vectors and fuzzy clustering to generate semantic clusters and performed web server logs-to-concept mapping to derive the semantic sub-sessions.

For evaluate goodness of semantic sub-session, also we use these 2 parameters. At first we obtain precision and recall for propositional system. then we obtain these measures by tow humans that driving sub-sessions based on propositional system. Then we obtain sub-sessions

TABLE III . Precision, recall and F-measure according to different weight thresholds for Wikipedia

Concept weight threshold	Wikipedia		
	Precision	*Recall*	*F-measure*
0.60	*50.00*	*52.13*	*52.37*
0.50	*49.53*	*50.34*	*49.93*
0.44	*48.00*	*49.36*	*49.00*
0.35	*40.84*	*40.50*	*46.17*
0.28	*47.25*	*47.39*	*46.81*

TABLE IV. Precision, recall and F-measure according to different weight thresholds for WordNet

Concept weight threshold	WordNet		
	Precision	*Recall*	*F-measure*
0.6	*37.46*	*40.34*	*38.84*
0.5	*37.62*	*41.63*	*39.52*
0.4	*31.84*	*38.51*	*34.85*
0.3	*31.50*	*37.42*	*34.20*
0.2	*29.75*	*38.52*	*33.57*

527

based on keyword extraction (tf-idf) with no lexical resource and no attention to the semantic.

Finally, comparison between the advantages of experimental results of proposed system (semantic sub-session) and F-measure (human results) with third experimental results (keyword extraction) (see table 5 and 6) shows that driving sub-sessions based on propositional system (based on semantic relations) in web personalization process is increases precision . The semantic threshold indicates the cut-off level of the gloss-overlap measure calculated during the concept selection. For example, at 20% threshold, any session pairs with gloss-overlap score less than 20% are considered to have no semantic relatedness.

VIII. CONCULUSION AND FUTURE WORK

In this paper we explained that how a promising lexical resource namely Wikipedia can be utilized in order to automatically extract concepts from web pages and derive semantic user models. It can be used to address the main liability of general ontologies with its acceptable coverage of name entities and domain specific terms and alleviate the burden of manually constructing concept hierarchies for automatic web personalization.

One future direction can be defining numerous concept similarity measures using various lexical resources in Wikipedia (such as category structure, article links) and evaluate their effectiveness in clustering semantic user sessions. Defining an eligible similarity measure that requires low computational complexity in order to be incorporated into an online recommender engine can be a challenging task.

REFRENCES

[1] R.Mihalcea, and A.Csomai,"Wikify!: linking documents to encyclopedic knowledge", In Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management CIKM '07. ACM, New York, NY, 2007,pp.233-242.

[2] P.Achananuparp, H.Han, O.Nasraoui, and R.Johnson, "Semantically enhanced user modeling" In Proceedings of the 2007 ACM Symposium on Applied Computing SAC '07, ACM, New York, NY, 2007,pp. 1335-1339.

[3] D.Oberle, B.Berendt, A.Hotho, J.Gonzalez.,"Conceptual User Tracking", In Proceedings of the 1st Atlantic Web Intelligence Conf. (AWIC),Madrid,Spain, 2003.

[4] S. Acharyya and J. Ghosh, "Context-Sensitive Modeling of Web Surfing Behaviour Using Concept Trees", In Proc. Of the 5th WEBKDD Workshop, Washington, August 2003.

[5] B.Magnini and C.Strapparava, "User Modelling for News Web Sites with Word Sense Based Techniques", User Modeling and User-Adapted Interaction, vol. 14,2004, pp. 239-257.

[6] M.Eirinaki et al. "Introducing Semantics In Web Personalization: The Role of Ontologies", in proceedings of EWMF/KDO 2005, LNAI 4289, 2006,pp. 147-162.

[7] Wikipedia, http://en.wikipedia.org

[8] G.Miller et al, "WordNet: an On-line Lexical Database", International Journal of Lexicography, 1990, pp. 235-245.

[9] C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira, "An introduction to the syntax and content of Cyc", In AAAI Spring Symposium, 2006.

[10] M. E. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", in Proceedings of the SIGDOC Conference ,Toronto, June 1986.

[11] D.Milne, O.Medelyan and I.H.Witten, "Mining Domain-Specific Thesauri from Wikipedia: A case study", In 6 Proceedings of the IEEE/WIC/ACM International Conference, 2006.

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

[13] T.Zesch, C.Muller and I.Gurevych,"Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary",In Proceedings of the Conference on Language Resources and Evaluation (LREC) 2008.

[14] http://ce.sharif.edu

[15] H. Dai and B. Mobasher, "Using Ontologies to Discover Domain-Level Web Usage Profiles", in Proceedings of the 2nd Workshop on Semantic Web Mining, Helsinki, Finland, 2002.

[16] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. Spyropoulos,"Web usage mining as a tool for personalization: A survey", User Modeling and User- Adapted Interaction, vol.13, 2003,pp.311–372.

[17] J. Giles. "Internet encyclopedias go head to head", Nature, 438(7070), 2005,pp. 900–901

[18] J. Li and O. R. Za¨ıane. Combining usage, content, and structure data to improve web site recommendation. In 5th International Conference on Electronic Commerce and Web Technologies (EC-Web 2004), 2004.

[19] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1):5–32, 1999.

[20] A. B. Geva, "Hierarchical Unsupervised Fuzzy Clustering", In proceeding of the IEEE transaction on fuzzy systems.1999

[21] R.Zaiane, Jia Li, R.Hayward "Mission-based navigational behavior modeling for web recommender system"Springer webKDD 2004,pp.37-55,2006

TABLE V. Precision, recall and F-measure across different semantic thresholds for proposed system

Semantic Threshold	Proposed System		
	Precision	Recall	F-measure
20	41.7	51.5	46.84
25	40.9	48.6	44.52
30	42.8	48.1	45.45
35	47.8	52.7	50.20
40	47.0	51.6	49.57

TABLE VI. Precision, recall and F-measure across different semantic thresholds for system based on keyword extraction

Semantic Threshold	System based on tf-idf keyword extraction		
	Precision	Recall	F-measure
20	26.5	27.3	38.84
25	31.8	27.6	39.52
30	29.4	25.7	34.85
35	32.7	28.6	34.20
40	33.3	28.0	33.57