Semantically Enhanced User Modeling

Palakorn Achananuparp and Hyoil Han
College of Information Science and Technology
Drexel University
Philadelphia, PA
pkorn@drexel.edu, hyoil.han@ischool.drexel.edu

Olfa Nasraoui
Dept. of Computer Engineering & Computer Science
Speed School of Engineering
University of Louisville
olfa.nasraoui@louisville.edu

Roberta Johnson
Education and Outreach Department
University Corporation for Atmospheric Research
rmjohnsn@ucar.edu

September 18, 2006

ABSTRACT
Content-based implicit user modeling techniques usually employ traditional term vector as a representation of the user's interest. However, due to the problem of dimensionality in vector space model, a simple term vector is not a sufficient representation of the user model as it ignores the semantic relations between terms. In this paper, we present a novel method to enhance a traditional term-based user model with the WordNet-based semantic similarity techniques. To achieve this, we utilize word definitions and relationship hierarchies in WordNet to perform word sense disambiguation and employ domain-specific concepts as category labels for the derived user models. We tested our method on *Windows to the Universe*, a public educational website covering subjects in the Earth and Space Science and performed an evaluation of our semantically enhanced user models against human judgment. Our approach is distinguishable from existing work because we automatically narrow down the set of domain specific concepts from an initial domain concepts obtained from Wikipedia and because we automatically create semantically enhanced user model.

Keywords
Implicit User Modeling, Content-based User Modeling, Semantic Techniques

1. INTRODUCTION
A complex and rich information space such as the World Wide Web has provided challenges and opportunities for researchers to study on. With voluminous amount of web pages, it has become more and more difficult for information seekers to find relevant information without facing information overload problem. Many research areas have tried to aid the users by building personalized information systems that provide the right resources

tailored to a specific user. One of the critical components in such systems is a user model which helps the systems understand various aspects of their users, such as their background knowledge, topics of interest, etc. Without an accurate representation of a user, the systems cannot provide the correct resources to the information seeker no matter how good the algorithms are.

According to Rich's taxonomy of user models [35], user models can be classified into several dimensions, i.e., *short-term*/*long-term* dimension dealing with the user information overtime, *explicit*/*implicit* dimension dealing with the way the model is extracted, *individual*/*group* dimension dealing with whom the model belongs to, individual users or generalized groups. We are interested in the explicit/implicit dimension of user modeling as it can determine the viability of the systems. Specifically, we focus on the implicit user modeling approach for web users since it requires no additional efforts from the users to construct the models. We show that an implicit and content-based approach offers some potential for user modeling on web-based system.

A major shortcoming of content-based approaches exists in the representation of the user model. Content-based approaches often use term vector, extracted from the content of web documents, to represent each user's interest. In doing so, they ignore the problem of dimensionality caused by semantic relations between terms [9] of the vector space model in which indexed terms are not orthogonal and often have semantic relatedness between one another.

In this paper, we present a method to create a semantically enhanced user model based on an implicit and content-based user modeling approach. Our goal is to improve the representation of a user model in content-based approaches by incorporating semantic content into term vector. To this end, we utilize word definitions and relations provided by WordNet to perform word sense disambiguation and employ domain-specific concepts as category labels for the derived semantically enhanced user models. The implicit information pertaining to the user behavior was extracted from clickstream data or web usage sessions captured in web server log.

Our approach is distinguishable from existing work because we automatically narrow down the set of domain specific concepts from an initial domain concepts obtained from Wikipedia [44] and because we automatically create semantically enhanced user model.

The rest of the paper is outlined as follows. First, we briefly reviewed related literatures in web personalization systems and content-based user modeling as well as semantic techniques. Then we presented a detail description of our method. Finally, the experimental results and discussion are presented.

2. RELATED WORK

There has been ongoing research in several domains, such as user modeling and web usage mining, to leverage the users' web usage behavior in order to implicitly create a model of their interests and provide personalization [22][30][41]. These systems often employed techniques from the vector-space model, e.g., term frequency and inverse document frequency (tf-idf) [39], to extract terms from web documents and construct term vector that represents the user model. However, there are certain limitations with the term vector approach because it ignores semantic relationships between terms, which in turn tend to decrease the accuracy of the user model representation.

Recent efforts have tried to adapt semantic techniques from related fields, such as information retrieval and computational linguistics, to improve the existing term vector approach. These methods can be generally categorized into two approaches, the *statistical* approach and the *taxonomical* approach.

Statistical techniques have been adopted to search for hidden semantic relationships among co-occurring objects [5][11]. Latent Semantic Analysis (LSA) [5][8] is a good example of a technique which tries to solve the dimensionality problem of the vector space model. In LSA, term vectors are mapped into a lower dimensional space associated with higher-level concepts.

The taxonomical approach is commonly performed through the use of WordNet [28]. Over the past few years, researchers have been investigating an application of WordNet-based approaches to user modeling and web personalization [7][26]. Eirinaki et al. adapted Wu & Palmer's semantic similarity measure [46] to construct *C*-

logs (concept logs), semantically enhanced web usage logs, which is used as an input to augment the web usage mining process. Magnini and Strapparava [26][27] proposed sense-based user modeling in their SiteIF system. They constructed the user models from word senses in WordNet synsets. The majority of semantic similarity algorithms using disambiguation approaches have utilized concept hierarchies and semantic relationships in WordNet. These techniques can be further categorized into three groups. The first group comprises *path-based* approaches [10][20][46] which measure semantic similarity between two concepts by counting the number of edges in the concept hierarchy of WordNet between them. The next category is the *information content* approach [36] [16][24]. The information content [36] of a concept can be measured by calculating the probability of occurrence of a concept in a corpus. The last group contains *gloss-based* approaches which make use of glosses or term definitions [21][20][33]. Concept pairs with maximum gloss overlaps have the maximum semantic similarity.

Other efforts in the area of semantic Web mining have addressed the semantic augmentation of user logs to improve Web usage mining. These include Berendt et al. [49] and Mobasher et al. [48] However, these techniques start with the assumption that the content can be manually mapped to existing manually constructed ontologies, which have been made specifically for the website at hand. This approach which is in common with Eirinaki et al.'s approach [7] relies on heavy "manual" involvement as opposed to our automated approach.

3. METHODOLOGY

This section describes our methodology to semantically enhance a term-based user model. The overall process is shown in figure 1. First, we build a *term-based* user model. Next we build a semantically enhanced user model by applying semantic similarity measures and term-to-concept mappings to the term-based user model. For each web usage session, we started by retrieving web document content corresponding to the URL sequence in the session. Then, we extracted individual terms from the content and created a term vector that represents the session. The term vector serves as the initial term-based user model (IT-UM) that our approach is intended to improve.

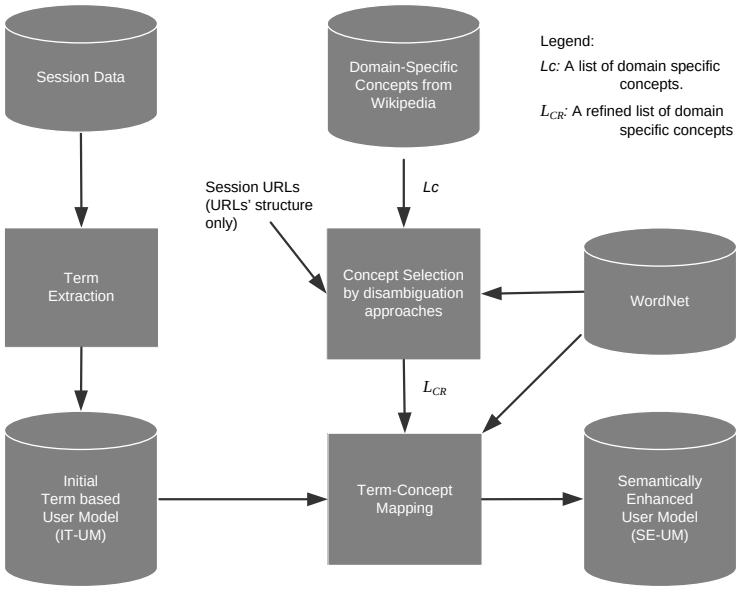

Figure 1: Overall Process

To build a semantically enhanced user model (SE-UM), we use refined domain-specific concepts. First we obtain a list (L_C) of domain-specific concepts from Wikipedia categories [44]. We disambiguate concepts in L_C by measuring their relatedness to keywords extracted from the structure of the URLs contained in the session. This

process gave us a more refined version (L_{CR}) of domain-specific concepts which take into account the contextual information in the session's URLs. Then we performed term-to-concept mapping between terms in the initial user model (IT-UM) and concepts in L_{CR} using WordNet taxonomy. The final product is a semantically enhanced user model (SE-UM) in which related terms are mapped to a high-level concept.

We begin by describing the two semantic similarity measures being used for concept selection process.

3.1 Semantic Similarity Measures

There are several approaches to determine semantic similarity between two terms. In this work, we implemented two semantic similarity measures, *gloss-overlap* and *path-based* measures that are used during the concept selection and term-to-concept mapping stages respectively.

3.1.1 Gloss-Overlap Measure

The *gloss-overlap* measure is a promising approach for performing word sense disambiguation [21][33]. Our algorithm is based on extending the gloss-overlap approach described in Pedersen et al. [33]. The extended gloss-overlap approach retains the advantage of the original gloss-overlap measure, i.e., the ability to determine semantic relatedness between two concepts independent of taxonomy structure in WordNet, while addressing the shortcoming of the original approach [21]. One drawback of the gloss-overlap approach lies in the length of glosses. Glosses of some concepts are too short to be used to effectively measure their relatedness [33]. To overcome this limitation, our extended approach measures the overlap between the initial two concepts as well as their related concepts from WordNet's concept hierarchies.

Our approach differs from Pedersen et al. [33], in which gloss vectors were used as representation of concepts. Instead, we use a more simplified approach by treating glosses as bags of words, as in the original algorithm, and use WordNet to retrieve glosses of their direct hypernyms and hyponyms. Before calculating the gloss overlap between the two concepts, we removed common stop words from the glosses and treated each gloss as a set of words. Given the two set of words, we define the similarity measure as follow:

$$Sim_{gloss}(a_i, b_j) = \frac{\left|Gloss(a_i) \cap Gloss(b_j)\right|}{\left|Gloss(a_i) \cup Gloss(b_j)\right|}$$

$$Sim_{gloss}(a, b) = \max\left[Sim_{gloss}(a_i, b_j)\right]$$

Where $Sim_{gloss}(a_i,b_j)$ is the gloss-overlap score between sense *i* of concept *a* and sense *j* of concept *b*. $Gloss(a_i)$ and $Gloss(b_j)$ are the set of words derived from the gloss of sense *i* of concept *a* and sense *j* of concept *b*, respectively. $Sim_{gloss}(a,b)$ is the overall gloss-overlap semantic similarity between concept *a* and *b*, which is derived from the sense pair of *a* and *b* which gives maximum gloss-overlap score. Notice that the measure is essentially a form of Jaccard coefficient [42], the most widely used similarity measure between term sets.

$$Gloss(a_1) = \{w_1, w_2, w_3\}$$
$$Gloss(a_2) = \{w_2, w_4\}$$
$$Gloss(b_1) = \{w_2, w_4\}$$
$$sim_{Gloss}(a_1, b_1) = \frac{\left|\{w_2\}\right|}{\left|\{w_1, w_2, w_3, w_4\}\right|} = 0.25$$
$$sim_{Gloss}(a_2, b_1) = \frac{\left|\{w_2, w_4\}\right|}{\left|\{w_2, w_4\}\right|} = 1$$
$$sim_{Gloss}(a, b) = 1$$

WordNet represents individual concepts as a synset (synonym set). Thus, to determine which senses of the two concepts are the most semantically related, we exhaustively calculated a gloss-overlap score between each sense of the two concepts. The score from the word sense pair that gives maximum gloss overlap will be chosen. Let's consider an example. Suppose we want to calculate a gloss-overlap score between the two concepts, *a* and *b*. We first look up both concepts in WordNet's taxonomy and obtain the synsets. Assuming that concept *a* has two word senses, a_1 and a_2; while concept *b* has one word sense, b_1. We start by measuring a gloss-overlap score between the first sense of *a* and *b*. Then we perform the same calculation with the second sense of *a* and the first sense of *b*. From the example above, we derive $Sim_{gloss}(a,b)$ from $Sim_{gloss}(a_2,b_1)$, which is equal to 1.

We tested our algorithm against a commonly used 28-noun benchmark set by human judgment [29][16][47]. The result (in Table 1) showed that our gloss-overlap algorithm (r = 0.849) correlates reasonably well with human judgment (r = 1). It also performs slightly better than Wu & Palmer's measure (r = 0.82).

Table 1: Comparison between Semantic Similarity Measures

Word Pair		**M&C means (human)**	**Gloss**	**W&P**
car	automobile	3.92	1	1
gem	jewel	3.84	1	1
journey	voyage	3.84	1	0.91
boy	lad	3.76	1	0.92
coast	shore	3.7	1	0.89
asylum	madhouse	3.61	1	0.94
magician	wizard	3.5	1	1
midday	noon	3.42	1	1
furnace	stove	3.11	1	0.46
food	fruit	3.08	0.14	0.73
bird	cock	3.05	1	0.93
bird	crane	2.97	0.22	0.82
tool	implement	2.95	1	0.93
brother	monk	2.82	1	0.93
crane	implement	1.68	0.09	0.67
lad	brother	1.66	0.33	0.67
journey	car	1.16	0.17	0
monk	oracle	1.1	0.06	0.53
food	rooster	0.89	0.14	0
coast	hill	0.87	0.75	0.6
forest	graveyard	0.84	0.17	0
monk	slave	0.55	0.13	0.67
coast	forest	0.42	0.25	0.25
lad	wizard	0.42	0.13	0.67
chord	smile	0.13	0	0.38
glass	magician	0.11	0	0.18
noon	string	0.08	0.07	0
rooster	voyage	0.08	0	0
Correlation with Miller & Charles (r)		1	0.849	0.82

3.1.2 Path-Based Measure

We used a simple edge-counting approach to help map terms to concepts. Our path-based measure is defined as follows:

$$Sim_{path}(a,b) = \min\left\lfloor Dist_{path}(a_i,b_j)\right\rfloor$$

Where $Dist_{path}(a_i,b_j)$ is a path-based similarity score between sense *i* of concept *a* and sense *j* of concept *b*. $Dist_{path}(a_i,b_j)$ is the distance between sense *i* of concept *a* and sense *j* of concept *b* in WordNet's taxonomy. To calculate such distance measure, we can simply count the number of edges on the path between two synsets that represent a_i and b_j. $Sim_{path}(a,b)$ is the overall path-based semantic similarity between concept *a* and *b*, which is taken from the sense pair with shortest distance. In addition, we set the maximum depth threshold to 12 edges as suggested by the literature [47]. As such, if $Sim_{path}(a,b)$ is greater than 12, it implies that they are not semantically related.

3.2 Term Extraction

We hypothesized that document content can be used as an implicit measure of the user's interests. In a typical web browsing session, a user will navigate through a number of URLs, either to look for specific information that satisfies their information need or casually skim through document content to see if there is anything that might be of interest to them.

We first obtain a term-based user model by constructing a term-document matrix from clickstream data. Clickstream data refers to a sequence of URLs that a user has visited within a particular period of time. For each URL, we extract individual terms from the whole document, perform stemming using WordNet, and remove common stopwords. After processing all the URLs in a session, we rank the extracted terms using term frequency and inverse document frequency (*tf-idf*) weighting approach. The *tf-idf* weight ($W_{tf\text{-}idf}$) is calculated as follow [39]:

$$tf.idf_{ij} = \frac{freq_{ij}}{\sum_l freq_{lj}} \times \log\frac{N}{n_i}$$

Where *tf* is the normalized frequency of a term k_i while the *Inverse Document Frequency* (*idf*) is a measure of importance of a term k_i with regard to other documents in the whole collection, $freq_{ij}$ is the frequency of a term k_i in the i^{th} document. *N* is the total number of documents (in our case, Web pages), and n_i is the number of these documents containing a term k_i. IDF is an indication of the importance of terms in the document corpus. The top-*k* terms in the ranking per session are used as an initial term based user model (IT-UM) for the session in figure 1. IT-UM is represented by a list of pairs (*term*, *weight*), where *weight* means $W_{tf\text{-}idf}$ for the term.

3.3 Concept Selection

Concept selection is a process to build and narrow down the set of domain-specific concepts for our experiment. Due to lack of public domain ontology for our experiment, we employed an alternative approach to finding domain-specific concepts through concept hierarchies [41]. We obtain a list (L_C) of domain-specific concepts from Wikipedia categories [44]. Then we narrow down the list of domain-specific concepts by strictly selecting concepts which are highly related to the contextual contents in a particular web usage session.

Based on the notion that hierarchical structure of web documents implicitly classifies the type of their content [3], we extracted keywords from a hierarchical structure of the URLs in a web usage session and utilized them to perform word sense disambiguation with domain-specific concepts. The Gloss-overlap measure (discussed in section 3.1.1) is the underlying disambiguation algorithm for this process. Any concepts with no semantic relatedness to URL keywords (Sim_{gloss} equal to 0) will be filtered out from the list. For example, suppose a session with *http://www.windows.ucar.edu/tour/link=* as a base URL, contains the following URL sequence:

/tour/link=/earth/Atmosphere/precipitation/rain.html
/tour/link=/earth/Atmosphere/rainbow.html
/tour/link=/earth/images/rainbow_image.html

From the structure of the URLs, we extract *earth*, *atmosphere*, and *precipitation* as URL keywords for this session (*tour*, *link*, and *images* are stop words in this case). Then, we calculate gloss-overlap measures between each URL keyword and every concept in the domain-specific concept list (L_C). This gives us a set of concepts which are semantically related to each URL keyword. The final list of refined domain-specific concepts for the session is a combined set of concepts that related to *earth*, *atmosphere*, or *precipitation*.

We repeat this process for all the sessions and finally build a refined list (L_{CR}) of domain-specific concepts for the sessions by merging refined concepts for each session. The list (L_{CR}) consists of 191 concepts related to topics in astronomy, geology, and biology.

3.4 Term-to-Concept Mapping

The final step is to build a semantically enhanced user model (SE-UM) by mapping terms to directly corresponding concepts (exact match) or concepts located higher (more general) in the concept hierarchies.

We mapped terms in the term vector to the list (L_{CR}) of refined domain-specific concepts which were obtained from the concept selection process. The goal is to find the most semantically related concept for each term, either the exactly matched concept or a more general concept. We used a path-based measure (discussed in section 3.1.2) between term-concept pairs. The advantage of using a path-based approach is in the explicit semantic relationship between concepts in WordNet hierarchies. Using hypernym and hyponym relations, a path-based measure is able to identify a superclass concept that a term should be mapped to.

We started from the first term (t_1) in the initial term-based user model (IT-UM) and exhaustively calculated a path-based measure with every concept in L_{CR}. A Concept in L_{CR} which gives the minimum path-based similarity score is selected as mapping assignment for that particular term, t_1. A new term weight is assigned to a selected concept by summing up the total weight of terms being mapped to it.

For example, suppose that the IT-UM in table 2 is given for a session and the refined domain-specific concept list (L_{CR}) contains two concepts, *primate* and *earth*. The weight in table 2 is given by the *tf-idf* approach (see Section 3.2).

Table 2: Example for the initial term based user model (IT-UM) from one session

Term	*monkey*	*Primate*	*forest*	*rain*	*gorilla*
Weight	7	6	5	4	3

We begin the mapping process with the first term in table 2, *monkey*. We calculate path-based measures, Sim_{path}(*monkey*,*primate*) and Sim_{path}(*monkey*,*earth*), respectively. If Sim_{path}(*monkey*,*primate*) is less than Sim_{path}(*monkey*,*earth*), then we will map *monkey* → *primate*. Next, repeat the same process with the second term in table 2, *primate*, and so on. The final mapping result is shown below:

Table 3: Example for the semantically enhanced user model (SE-UM) of a mapping result for table 2

Concept	**Terms**	**Weight**
primate	*monkey, primate, gorilla*	16
earth	*Forest*	5

The table 3 displays an SE-UM corresponding to an IT-UM in table 2 where terms are mapped to their most related domain-specific concepts. *Monkey*, *primate*, and *gorilla* are mapped to *primate* while *forest* is mapped to *earth*. Notice that *rain* does not have any mapping assignment since it has no semantic similarity with any of the concepts. The new term weights for *primate* and *earth* are 16 (= 7 + 6 + 3) and 5 respectively. This SE-UM in

table 3 indicates that the majority of the user interest in this web usage session is in topics related to *primate* and *earth*. A semantically enhanced user model (SE-UM) is represented by a list of pairs (*term*, *weight*), where *weight* is the summation of the weight $W_{tf\text{-}idf}$ of each term that is mapped to the *term*. (see Section 3.2 for $W_{tf\text{-}idf}$).

4. EXPERIMENTAL RESULTS

We used Windows to the Universe, a public educational website covering subjects in the Earth and Space Science as a source of our experimental data. We extracted clickstream sequences from about 3 hours and 40 minutes worth of access log data on http://www.windows.ucar.edu, comprising a total of 65,536 hits. These hits were later segmented into 2,400 individual user sessions accessing a total of 8,044 URLs. Sessions were segmented based on an IP address and a time out mechanism (a maximum of 45 minutes between consecutive accesses) as described in [32], with the only difference that the sequence information was preserved. This means that repeat visits to the same page are also recorded into the sequence. We then used a web crawler to gather the Web page content data for our experiments. The next step was to filter out 'unrealistic' sessions so that the number of outlier can be minimized. Hence, we define the characteristics of normal sessions as follows:

- The length of a session is from two to ten URLs. This is done to avoid sessions initiated by crawlers that tend to be too short or too long [12][30].
- The content type of the URLs is HTML document and their language is in English (this is because most web pages have a Spanish version).
- The URLs are syntactically and semantically well-formed (for the purpose of extracting meaningful words from the URL). This accounts for approximately 90% of all the URLs.

Any URLs that do not fit the above criteria were discarded. Ultimately, we selected 100 normal sessions for the experiment. Next, we used this session data to generate the initial term-based user models (IT-UM) and performed term-to-concept mapping to derive the semantically enhanced user models (SE-UM). A human evaluator, a graduate student with moderate knowledge in the application domains, annotated the same set of sessions with concepts from the domain-specific concept list (L_{CR}). The annotation is used to represent human-annotated user models (HA-UM). The evaluator had to familiarize herself with domain-specific concepts as well as subject contents in the website. The evaluator could use as many or as few concepts as she saw fit.

A comparison between the semantically enhanced user models (SE-UM) and human-annotated user models (HA-UM) was performed. To estimate the quality of user models produced by our method, we defined the following precision/recall measure: *Precision* is the number of correct concepts in the user model divided by the total number of concepts in the user model. *Precision* indicates the accuracy of mapping and is obtained by the formula C_R/C_{SE}, where C_R is the set of correct concepts in the user model, and C_{SE} is the set of total concepts in the user model. *Recall* is the number of correct concepts in the user model divided by the total number of concepts annotated by the human evaluator. *Recall* indicates the effectiveness of mapping and is obtained by the formula C_R/C_{HA}, where C_R is the set of correct concepts in the user model, and C_{HA} is the set of total concepts annotated by the human evaluator.

Table 4 shows our experimental results. The *semantic threshold* indicates the cut-off level of gloss-overlap measure calculated during the concept selection phase. For example, at 20% threshold, any word pairs with gloss-overlap score less than 20% are considered to have no semantic relatedness.

Table 4: Precision/recall across different semantic thresholds

Semantic Threshold (percentage)	*Precision*		*Recall*	
	Average (P_R)	Adjusted P_R	Average (R_R)	Adjusted R_R
20	24.5%	31.7%	28.3%	41.5%
25	29.8%	30.9%	25.6%	38.6%
30	30.4%	31.8%	24.7%	38.1%
35	30.7%	37.8%	26.6%	42.7%
40	31.3%	37.8%	26.0%	42.5%
45	31.3%	37.0%	26.0%	41.6%

(Note: Adjusted average precision and recall values were calculated after the outliers, any sessions with either 0% or 100% precision/recall, were excluded.)

HA-UM contains an average of 3.81 concepts per session with the maximum number of 15 concepts in a session. Based on a statistical F-test, the average number of concepts per session differ significantly across HA-UM and SE-UM derived from various semantic threshold levels ($F(6,579) = 3.527$, $p < 0.01$).

Furthermore, the average number of concepts per session and the maximum number of concepts per session derived from our method decrease as the semantic threshold increases. At the 35% semantic threshold or higher, the auto-generated user models contain comparable number of concepts with those of human-annotated ones. The similar trend can be observed in the average precision (P_R) measure. As the semantic threshold increases, the precision or accuracy of mapping increases. The results offer some insight into the characteristic of human-annotated user model as compared to auto-generated user models. Quantitatively, our method produced a reasonably equal number of concepts per session.

5. CONCLUSION

In this paper, we presented a method of constructing a semantically enhanced user model that represents the user's interests from clickstream data or web usage logs. The goal of incorporating the semantic content of the web pages to build the semantically enhanced user models requires to address the dimensionality problem and semantic inadequacy of the vector space model, on which the initial user model is based, and to map conceptually related terms. Our method makes use of a WordNet-based approach and a domain-specific concept list that is refined based on Wikipedia concepts and URL based information.

We evaluated the semantically enhanced user models against the user model derived from human annotation based on precision and recall. Our preliminary experimental results indicate that our method produced a fair result with respect to human annotation, especially at higher semantic thresholds.

Lastly, we recognized a few limitations in our method. First, the gross-overlap measures require a great amount of computation time and the efficiency does not scale well as the number of word pairs increases. Next, we assumed that semantic terms can be extracted from the URL structure. This is true in the case of our website, but it is probably inapplicable to some other URLs. For future work, we plan to improve the efficiency and scalability of our method, especially on a larger data set. We also plan to automatically construct a list of domain-specific concepts from several sources.

6. ACKNOWLEDGMENTS

Olfa Nasraoui is supported by National Science Foundation CAREER Award IIS-0133948.

7. REFERENCES

[1] Allen, R. User Models: Theory, Method and Practice. International Journal of Man-Machine Studies. International Journal of Man-Machine Studies 32 1990, 511-543

[2] Baglioni, M., Ferrara, U., Romei, A., Ruggieri, S. and Turini, F. (2003) Preprocessing and Mining Web Log Data for Web Personalization, in Proceedings of the 8th Italian Conf. on Artificial Intelligence, vol. 2829 of LNCS, September, 237-249.

[3] Chang, Domain Specific Word Extraction from Hierarchical Web Documents: A First Step Toward Building Lexicon Trees from Web Corpora

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

[5] Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S.,and Harshman, R. (1988) Using Latent Semantic Analysis to Improve Access to Textual Information. Proceedings of the Conference on Human Factors in Computing Systems, CHI. 281-286

[6] Dumais, S., Joachims, T., Bharat, K., and Weigend, A. (2003) SIGIR 2003 workshop report: implicit measures of user interests and preferences. SIGIR Forum 37, 2 (Sep. 2003), 50-54.

[7] Eirinaki, M., Vazirgiannis, M., and Varlamis, I. (2003). SEWeP: using site semantics and a taxonomy to enhance the Web personalization process. In Proceedings of the Ninth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Washington, D.C., August 24 - 27, 2003). KDD '03. ACM Press, New York, NY, 99-108.

[8] Foltz, P.W. (1990) Using latent semantic indexing for information filtering, in R.B. Allen (Ed.) Proceedings of the Conference on Office Information Systems, Cambridge, MA, 40-476

[9] Furnas, G.W., Landauer, T.K., Gomez, L. M., and Dumais, S. T. (1987) The vocabulary problem in human-system communication. Communications of the ACM, 30(11):964—971.

[10] Hirst, G. and D. St.Onge (1995). Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. WordNet. C. Fellbaum. Cambridge, MA, The MIT Press.

[11] Hofmann, T. (1999) Probabilistic latent semantic indexing. In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, Berkeley, CA, August.

[12] Hu, J. and Zhong, N. (2005) Clickstream Log Acquisition with Web Farming. In Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), 257-263.

[13] Hu, X., Yoo, I., Song, M., Zhang, Y., and Song, I. (2005) Mining undiscovered public knowledge from complementary and non-interactive biomedical literature through semantic pruning. In Proceedings of the 14th ACM international Conference on information and Knowledge Management, October 31 - November 05. ACM Press, New York, NY, 249-250.

[14] Humphreys, B. L. and Lindberg, D. A. (1993) The UMLS project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc. 1993 April; 81(2), 170–177.

[15] HTML Parser. http://htmlparser.sourceforge.net/.

[16] Jiang, J. and Conrath, D. (1997) Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings on International Conference on Research in Computational Linguistics, Taiwan, 19-33.

[17] Jin, X., Zhou, Y., and Mobasher, B. (2004) Web usage mining based on probabilistic latent semantic analysis. In Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining. ACM Press, New York, NY, 197-205.

[18] Kaur, I. and Hornof, A. J. (2005). A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM Press, New York, NY, 51-60.

[19] Kelly, D. and Teevan, J. (2003) Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37, 2, 18-28.

[20] Leacock, C. and Chodorow, M. (1998) Combining local context and WordNet similarity for word sense identification, in C. Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, 1998, 265-283.

[21] Lesk, M. (1986) Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, in Proceedings of the 5th annual International Conference on Systems Documentation, ACM Press, 24-26.

[22] Lieberman, H. (1995) Letizia: An Agent That Assists Web Browsing. Proceedings of the International Joint Conference on Artificial Intelligence.

[23] Lieberman, H., Van Dyke, N. W., and Vivacqua, A. S. (1999) Let's browse: a collaborative Web browsing agent. In Proceedings of the 4th international Conference on intelligent User interfaces. ACM Press, New York, NY, 65-68.

[24] Lin, D. (1997) Using syntactic dependency as a local context to resolve word sense ambiguity, in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, 64-71.

[25] Lipscomb, C. E. (2000) Medical Subject Headings (MeSH). Bull Med Libr Assoc. 2000 July; 88(3), 265–266.

[26] Magnini, B. and Strapparava, C. (2000) Experiments in Word Domain Disabmiguation for Paralle Texts, in: Proceedings of SIGLEX Workshop on Word Senses and Multi-linguality. Hong-kong, 27-33.

[27] Magnini, B. and Strapparava, C. (2004) User Modelling for News Web Sites with Word Sense Based Techniques. User Modeling and User-Adapted Interaction 14, 2-3 (Jun. 2004), 239-257.

[28] Miller, G. A. (1995) WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41.

[29] Miller, G. A. and W. G. Charles (1991). Contextual Correlates of Semantic Similarity. Language and Cognitive Processes 6(1): 1-28.

[30] Mobasher, B., Cooley, R., and Srivastava, J. (2000) Automatic personalization based on Web usage mining. Commun. ACM 43, 8 (Aug. 2000), 142-151.

[31] ODP – Open Directory Project. http://dmoz.org.

[32] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi. Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering, International Journal on Artificial Intelligence Tools, Vol. 9, No. 4, pp. 509-526, 2000.

[33] Pederson, T., Banerjee, S., and Patwardhan, S. (2005) Maximizing semantic relatedness to perform word sense disambiguation

[34] Rich, E. (1979) Building and Exploiting User Models. CMU Dissertation.

[35] Rich, E. (1983) Users are individuals: individualizing user models. International Journal of Man-Machine Studies, 18, 199-214.

[36] Resnik, P. (1995) Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 448-453.

[37] Richardson, R. and Smeaton, A.F. (1995) Using WordNet in a knowledge-based approach to information retrieval

[38] Salton, G. (1971) The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice Hall Inc., Englewood Cliffs, NJ.

[39] Salton, G. and Buckley, C. (1988) Term-weighting approaches in automatic retrieval. Information Processing and Management, 24(5):513-523.

[40] Salton, G. and Lesk M. E. (1968) Conputer evaluation of indexing and text processing. Journal of the ACM, 15(1):8-36, January.

[41] Sieg, A., Mobasher, B., and Burke, R. (2004) Inferring User's Information Context: Integrating User Profiles and Concept Hierarchies. Proceedings of the 2004 Meeting of the International Federation of Classification Societies, Chicago, IL.

[42] Tan, P., Steinbach, M., and Kumar, V. (2005) Introduction to Data Mining. Addison-Wesley.

[43] Wang, B. B., Mckay, R. I., Abbass, H. A., and Barlow, M. (2003). A comparative study for domain ontology guided feature extraction. In Proceedings of the Twenty-Sixth Australasian Computer Science Conference on Conference in Research and Practice in information Technology - Volume 16 (Adelaide, Australia), vol. 35, 69-78.

[44] Wikipedia Categories. http://en.wikipedia.org/wiki/Wikipedia:Browse.

[45] Wang, B. B., Mckay, R. I., Abbass, H. A., and Barlow, M. (2003) A comparative study for domain ontology guided feature extraction. In Proceedings of the Twenty-Sixth Australasian Conference on Computer Science: Research and Practice in information Technology. Australian Computer Society, Darlinghurst, Australia, 69-78.

[46] Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd Annual Meeting on Association For Computational Linguistics (Las Cruces, New Mexico, June 27 - 30, 1994), 133-138.

[47] Yang, D. and Powers, D. M. (2005). Measuring semantic similarity in the taxonomy of WordNet. In Proceedings of the Twenty-Eighth Australasian Conference on Computer Science - Volume 38 (Newcastle, Australia), vol. 102, 315-322.

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

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