Smart and Effective Web Page Recommendation System using Ontology

Prof. Mrunal Pathak[1], Pagore Mayur[2], Dayanand Shelke[3], Akshar Urade[4], Powar Ajit[5]
[1, 2, 3, 4, 5] Department of Information Technology
[1, 2, 3, 4, 5] 12345AISSMS Institute Of Information Technology

Abstract- *In the expanding era of the Internet web-page recommendation plays a vital role in intelligent Web systems. The knowledge used in web-page recommendations is very essential and pivotal for the design of new successful web pages/websites. The use of semantic-enhancement in the recent years by integrating the recommendations assists a lot in the process of Web page recommendation. Two new models are designed to represent the domain knowledge. The first model uses an ontology to recommend web pages. The second model uses a accordingly generated semantic network to represent domain terms, Web pages and relations between them. In addition to that, an advanced model the conceptual prediction model, is able to automatically generate a semantic Web usage knowledge, which is the combination of domain knowledge and Web usage knowledge. A set of recommendation strategies have been developed. The recommendation strategies have been developed to overcome the shortcomings of existing system of Web Usage Mining(WUM)method. The experimental results demonstrate that the proposed approach produces notably higher performance than the WUM approach.*

Keywords- Web Usage Mining, Web-page recommendation, ontology, semantic network.

I. INTRODUCTION

WEB-PAGE recommendation has become increasingly popular, and is shown as links to related stories, related books, or most viewed pages at websites. When a user browses a website, a sequence of visited Web-pages during a session (the period from starting, to existing the browser by the user) can be generated. This sequence is Organized into a Web session S = d1,d2...dk, where di(i = [1...k]) is the page ID of the i th visited Web-page bythe user. The objective of a Web-page recommender system is to effectively predict the Web-page or pages that will be visited from a given Web-page of a website. There are a number of issues in developing an effective Web-page recommender system, such as how to effectively learn from available historical data and discover useful knowledge of the domain and Web-page navigation pat- terns, how to model and use the discovered knowledge, and how to make effective Web-page recommendations based on the discovered knowledge. A great deal of research has been devoted to resolve these issues over the past decade. It has been reported that the approaches based on tree structures and probabilistic models can efficiently represent Web access sequences reorganization.(WAS) in the Web usage data [1]. These approaches learn from the training datasets to build the transition links between Web-pages. By using these approaches, given the current visited Web-page (referred to as a state) and k previously visited pages (the previous k states), the Web-page(s) that will be visited in the next navigation step can be predicted. The performance of these approaches depends on the sizes of training datasets. The bigger the training dataset size is, the higher the prediction accuracy is. However, these approaches make Web-page recommendations solely based on the Web access sequences learnt from the Web usage data. Therefore, the predicted pages are limited within the discovered Web access sequences, i.e., if a user is visiting a Web-page that is not in the discovered Web access sequence, then these approaches cannot offer any recommendations to this user. We refer to this problem as "new-page problem" in this study. Some studies have shown that semantic-enhanced approaches are effective to overcome the new-page problem [2], [3] and have therefore become far more popular. The use of domain knowledge can provide tremendous advantages in Web-page recommender systems [4]. A domain ontology is commonly used to represent the semantics of Web-pages of a website. It has been shown that integrating domain knowledge with Web usage knowledge enhances the performance of recommender systems using ontology based Web mining techniques [4]–[6]. Integrating semantic information with Web usage mining achieved higher performance than classic Web usage mining algorithm However, one of the big challenges that these approaches are facing is the semantic domain knowledge acquisition and representation. How to effectively construct the domain ontology is an ongoing research topic. The domain ontology can be constructed manually by experts, or by automatically learning models, such as the Bayesian network or a collocation map, for many different applications. Manually building an ontology of a website is challenging given the very large size of Web data in today's websites, as well as the well-known drawbacks of being time consuming and less reusable. According to Stumme, Hotho and Brandt, it is impossible to manually discover the meaning of all Web-

pages and their usage for a large scale website [10]. Automatic construction of ontologies. on the other hand, can save time and discover all possible concepts within a website and links between them, and the resultant ontologies are reusable. However, the drawback of this automatic approach is the need to design and implement the learning models which can only be done by professionals at the beginning. Therefore, the trade-off between the two approaches to ontology construction needs to be considered and evaluated for a given website. This paper presents a novel method to provide better Web-page recommendation based on Web usage and domain knowledge, which is supported by three new knowledge representation models and a set of Web-page recommendation strategies. The first model is an ontology based model that represents the domain knowledge of a website. The construction of this model is semi-automated so that the development efforts from developers can be reduced. The second model is a semantic network that represents domain knowledge, whose construction can be fully automated. This model can be easily incorporated into a Web-page recommendation process because of this fully automated feature. The third model is a conceptual prediction model, which is a navigation network of domain terms based on the frequently viewed Web-pages and represents the integrated Web usage and domain knowledge for supporting Web-page prediction. The construction of this model can be fully automated. The recommendation strategies make use of the domain knowledge and the prediction model through two of the three models to predict the next pages with probabilities for a given Web user based on his or her current Web-page navigation state. To a great extent, this new method has automated the knowledge base construction and alleviated the new-page problem as mentioned above. This method yields better performance compared with the existing Web usage based Web-page recommendation systems

II. THE EXISTING MODEL

We roughly classify the research work related to Web-page recommendation into the following two categories. 2.1 Traditional Approaches that use Sequence Learning Models In applying sequence learning models to Web-page recommendation, association rules and probabilistic models have been commonly used. Some models, such as sequential modelling, have shown their significant effectiveness in recommendation generation [2]. In order to model the transitions between different Web-pages in Web sessions, Markov models and tree-based structures are strong candidates [2], . Some surveys have shown that tree-based algorithms, particularly PreOrder Linked WAP-Tree Mining (PLWAP-Mine for short), are outstanding in supporting Web-page recommendation, compared with other sequence mining algorithms. Furthermore, the integration of PLWAP-Mine and the higher-order Markov model can significantly enhance mining performance. 2.2 Semantic-Enhanced Approaches The semantic-enhanced approaches integrate semantic information into Web-page recommendation models. By making use of the ontology of websites, Web-page recommendation can be enriched and improved significantly in the systems . In the systems, a domain ontology is often useful for clustering documents, classifying pages or searching subjects. A domain ontology can be obtained by manual or automatic construction approaches, for example, ontologies have been developed for distance learning courses , course content, personalized e-learning, contracts, and software. Depending on the domain of interest in the system, we can reuse some existing ontologies or build a new ontology, and then integrate it with Web mining. For example, ontology concepts are used to semantically enhance Web logs in a Web personalization system . In this system, an ontology is built with the concepts extracted from the documents, so that the documents can be clustered based on the similarity measure of the ontology concepts. Then, usage data is integrated with the ontology in order to produce semantically enhanced navigational patterns. Subsequently, the system can make recommendations, depending on the input patterns semantically matched with the produced navigational patterns. Liang Wei and Song Lei employ ontology to represent a website's domain knowledge using the concepts and significant terms extracted from documents. They generate online recommendations by semantically matching and searching for frequent pages discovered from the Web usage mining process. This approach achieves higher precision rates, coverage rates and matching rates. On the other hand, by mapping Web-pages to domain concepts in a particular semantic model, the recommender system can reason what Web-pages are about, and then make more accurate Web-page recommendations .Alternatively, since Web access sequences can be converted into sequences of ontology instances, Web-page recommendation can be made by ontology reasoning . In these studies, the Web usage mining algorithms find the frequent navigation paths in terms of ontology instances rather than normal Web-page sequences. Generally, ontology has helped to organize knowledge bases systematically and allows systems to operate effectively.

III. DOMAIN ONTOLOGY OF A WEBSITE FORWEB-PAGE RECOMMENDATION

The ontology based system works in the following manner as depicted in the figure. The Fwap a fast wireless access protocol sends the request to one of the FVTP clients which in turn alerts the CPM to activate the request to access the web page. The process repeats itself over and over and hence a pattern is formed which resembles an ontological pattern

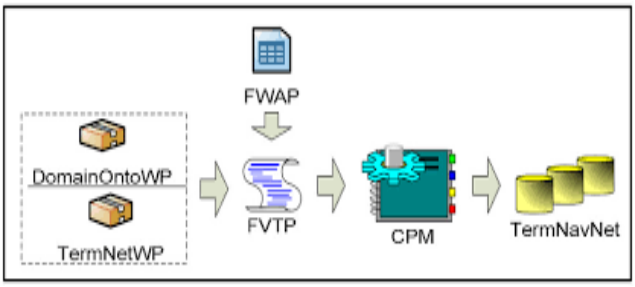

Fig 1: System Architecture

A domain ontology is defined as a conceptual model that specifies the terms and relationships between them explicitly and formally, which in turn represent the domain knowledge for a specific domain . The three main components are listed as follows :

1) Domain terms (concepts),
2) Relationships between the terms (concepts), and
3) Features of the terms and relationships.

Ontologies are often implemented in a logic-based language, such as OWL/RDF, to become understandable to software agents or software systems. Therefore, ontology-based knowledge representation allows sharing and interchanging semantic information among Web systems over the Internet. It also enables the reuse of the domain knowledge, and reasoning the semantics of Web-pages from the existing facts . Furthermore, ontological representation of discovered knowledge from different sources can be easily integrated to support Web-page recommendation effectively.

Depending on the purposes of Ontologies, they can be designed as domain conceptualizations of various degree of formality and can be in the form of concept schemes, taxonomies, conceptual data models, or general logical theories . In this section, we will construct a conceptual data model as a domain ontology for a given website. Since this ontology is used to support Web-page recommendation, we take a Web-page as a unit and assume each page title is well defined to represent key information about the content of the page. The rationale behind this assumption can be seen from two aspects. One aspect is that a Web-page contains a collection of objects (represented by HTML tags) documented in metadata, which is data about data. Metadata embraces the core elements of title, meaning, descriptive context, structure, and overall context of a Web-page. By analyzing the metadata, such as Web-page title, the meaning of a Web-page can be understood and captured. The second aspect is from the professional practice in Web development. In well-designed Web-pages, the TITLE tag should contain the meaningful keywords which are relatively short and attractive to support Web search or crawling. In practice, the terms in page titles are usually given higher weights by search engines, such as Google. Consequently, professional website developers need to define the Web-page titles very seriously because they want their Web-pages to be correctly identified during Web search or crawling and use the Web-page titles to convey accurate information about the Web-page. Because of these facts, we use the Web-page titles as clues to represent the domain knowledge of a website. It implies that although there are numerous models for extracting topics of Web-pages, making use of Web-page titles is simple and easy to implement.

This section now presents a procedure for constructing the domain ontology using the Microsoft (MS) web-site (www.microsoft.com) as an example. The dataset was downloaded from http://kdd.ics.uci.edu/databases/msweb/msweb.html. The ontology will be constructed based on the titles of visited Web-pages so that it is the domain knowledge perceived by users. Queries are then provided based on this domain ontology.

IV. DOMAIN ONTOLOGY CONSTRUCTION

There are three steps in the procedure for constructing the domain ontology.

Step 1: Collect the terms

In order to collect the terms, we will: (i) collect the Web log file from the Web server of the website for a period of time (at least seven days), (ii) run a pre-processing unit to analyse the Web log file and produce a list of URLs of Web-pages that were accessed by users, (iii) run a software agent to crawl all the Web-pages in the URL list to extract the titles, and (iv) apply an algorithm to extract terms from the retrieved titles, i.e., single tokens are extracted first by removing stop words from the titles, some single tokens are then combined into composite terms if these single terms often occur at the same time and there is never any token appears between these tokens, and the remaining single tokens will become single word terms. Using the MS Web dataset, we obtain the Web-page titles and paths. Based on the extracted terms, we can generalize them to domain concepts .

Step 2 Define the Concepts

It is possible for some extracted terms to share the same features, so it is better for them to be instances of a concept, rather than standalone concepts. In this step, the domain concepts will be defined for the given website based on the extracted terms. In this paper, we present the MS website as an example. This website focuses on the application software, such as MS Office, Windows Operating System ,and Database. Therefore, the identified domain

concepts of this website are Manufacturer, Application, Product, Category, Solution, Support, News, Misc, and SemPage, where the concept SemPage refers to the class of Web-pages, and the other concepts refer to the general terms in the MS website.

Step 3: Define taxonomic and non-taxonomic relationships

According to Uschold and Gruminger [31], there are three possible approaches to develop the taxonomic relationships, such as, (4) a top-down development process starts from the most general concepts in the domain and then identifies the subsequent specialization of the general concepts, (5) a bottom-up development process starts from the most specific concepts as the leave nodes in the concept hierarchical structure/tree structure, then groups these most specific concepts into more general concepts, (hybrid development process is the combination of the top-down and bottom-up approaches. We identify the core concepts in the domain first and then generalise and specialise them appropriately. With the MS website example, we applied a hybrid approach to define taxonomic relationships. We started with the concept Application and Product. Considering the consistsOf relation, which indicates that a concept comprises a number of parts that are also concepts, we particularly have an application that may consist of some sub-applications. For instance, an application software of Office has some sub-applications, including Word, Access, PowerPoint, etc. Considering the includes relation, we have a product that may include some sub-products, e.g. Office software product includes sub-products such as MS Word, MS Access, and MS Power Point. Considering the belongsTo relation, we have a product that may belong to a certain category software, hardware, entertainment service. The non-taxomonic relationships can be the relationship types used in a relational database except for the relationships between a super-set and a sub-set, such as self-referencing, 1-M and M-N relationships. In the MS website example, the main types of non-taxonomic relationships are listed as below.

1. The ‘provides’ relation describes the M:N relation- ship between concept Manufacturer and concepts Product, Solution, Support & News. The ‘isProvided’ relation is the inverse of the ‘provides’ relation.
2. The ‘has’ relation describes the M:N relationship between concept Application and concepts Product, Solution, Support News. The ‘isAppliedFor’ relation is the inverse of the ‘has’ relation.
3. The ‘hasPage’ relation describes the M:N relation- ship between a concept, such as Application and This domain ontology is constructed at three levels:

1. General level, which holds the concepts that present the general domain terms of Web-pages and relationship definition sets;
2. Specific level, which holds the specific domain terms corresponding to the domain concepts, e.g. terms “Database” and “Office” are the instances of concept Application, and the relationships between terms;
3. Web-page level, which holds all the Web-pages within the given website, and the association relationships between Web-pages and terms. The general level is presented as an ontology schema, and the specific and Web-page levels are presented as ontology instances. Such an ontology model supports modular development, scalability and reusability at different levels. The general terms within a domain are usually stable and have little changes over the time while the specific terms can increase frequently with the evolution of the Web-page. For instances, when a new Web-page is generated in the site, its title can very likely contain specific terms that are part of existing general terms, but not in the domain ontology. With this ontology structure, these specific terms can be easily added into the ontology at the specific level and the Web-page can be included easily at the Web-page level. The association relations between the concept SemPage and other domain concepts allow the machine to interpret Web-pages or identify what Web-pages are about. However, one problem is how to assign numerous Web-pages to domain terms appropriately. The clue is keywords existing in Web-page titles. Hence, each term instance needs to be specified by relevant keywords. By matching keywords in terms and Web-page titles, the system can automatically map the Web-pages with respect to the domain terms. This domain ontology, namely DomainOnto WP is implemented using OWL in Protégé. With the help of OWL, we can perform the following queries for the use in the later recommendation process.

V. CONCEPTUAL PREDICTION MODEL(CPM)

In order to obtain the semantic Web usage knowledge that is efficient for semantic-enhanced Web-page recommendation, a conceptual prediction model (CPM) is proposed to automatically generate a weighted semantic network of frequently viewed terms with the weight being the probability of the transition between two adjacent terms based on FVTP. We refer to this semantic network as TermNavNet hereafter. .

According to the Markov model [32], a kind of model efficient to represent a collection of navigation records, CPM is developed as a self-contained and compact model. It

has two main kinds of elements: (4) state nodes, and (5) the relations between the nodes. One node presents the current state, e.g. current viewed term, and may have some previous state nodes and some next state nodes. By scanning each term pattern F ∈ F, each term becomes a state in the model. There are also two additional states: a start state, S, representing the first state of every term pattern; and a final state, E, representing the last state of every term pattern. There is a transition corresponding to each pair of terms in a pattern, a transition from the start state S to the first term of a term pattern, and a transition from the last term of a term pattern to the final state E. The model is incrementally built by processing the complete collection of FVTP.

VI. CONCLUSION

In conclusion, this paper has presented a new method offer better Web-page recommendations through semantic enhancement by three new knowledge representation models. Two new models have been proposed for representation of domain knowledge of a website. One is an ontology based model which can be semi-automatically constructed, namely Domain Onto, and the other is a semantic network of Web-pages, which can be automatically constructed, namely Term Newt. A conceptual pre- diction model is also proposed to integrate the Web usage and domain knowledge to form a weighted semantic net- work of frequently viewed terms, namely TermNavNet. A number of Web-page recommendation strategies have been proposed to predict next Web-page requests of users through querying the knowledge bases. The experimental results are promising and are indicative of the usefulness of the proposed models. Compared with one of the most advanced Web usage mining method, i.e. PLWAP-Mine, the proposed method can substantially enhance the performance of Web-page recommendation in terms of precision and satisfaction. More importantly, this method is able to alleviate the "new-page" problem mentioned in the introduction because it based on not only the Web usage knowledge, but also the semantics of Web-pages. For the future work, a key information extraction algorithm will be developed to compare with the term extraction method in this work, and we will perform intense comparisons with the existing semantic Web-page recommendation systems.

REFERENCES

[1] B. Mobasher, "Data mining for web personalization," in The Adaptive Web, vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 90–135.

[2] G. Stumme, A. Hotho, and B. Berendt, "Usage mining for and on the Semantic Web," in Data Mining: Next Generation Challenges and Future Directions. Menlo Park, CA, USA: AAAI/MIT Press,2004, pp. 461–480

[3] N. R. Mabroukeh and C. I. Ezeife, "Semantic-rich Markov modelsfor Web prefetching," in Proc. ICDMW, Miami, FL, USA, 2009, pp.465–470

[4] M. O'Mahony, N. Hurley, N. Kushmerick, and G. Silvestre,"Collaborative recommendation: A robustness analysis," ACM Trans. Internet Technol., vol. 4, no. 4, pp. 344–377, Nov. 2004.

[5] G. Stumme, A. Hotho, and B.Berendt, "Semantic Web mining: State of the art and future directions," J. Web Semant., vol. 4, no.2, pp. 124–143, Jun. 2006.