EFFECTIVE KNOWLEDGE REPRESENTATION INTEGRATED WITH WEB USAGE MINING FOR WEB PAGE RECOMMENDATION

Namita Ganjewar

Namita Ganjewar, Dept. of Computer Engineering, Pune Institute of Computer Technology, India.

ABSTRACT— *Web page recommendation plays an important role in intelligent web systems. There are many techniques that serve this purpose. The review of existing systems considering their performance and limitations is represented here. And to improve the efficiency, this paper proposes a new framework for a semantic-enhanced Web-page recommender system, and a suite of enabling techniques which include semantic network models of domain knowledge and Web usage knowledge, querying techniques, and Webpage recommendation strategies. The framework enables the system to automatically discover and construct the domain and Web usage knowledge bases, and to generate effective Web page recommendations.*

Keywords —Web page recommendation (WPR), Web usage mining, Domain knowledge modeling, Knowledge representation, Semantic network.

1. INTRODUCTION

The explosive growth of information on the World Wide Web with the development of advanced electronic devices has made Web information increasingly important in almost everybody's life.

The rapid introduction of current websites has overwhelmed Web users by offering many choices [1]. Consequently, Web users tend to make poor decisions when surfing the Web due to an inability to cope with enormous amounts of information. Recommender systems have proved in recent years to be a valuable means of helping Web users by providing useful and effective recommendations or suggestions [1] [3]. The core techniques in recommender systems are the learning and prediction models which learn users‟ behavior and evaluate what users would like to view in the future. In particular, a recommender system can suggest interesting items from a large set of items based on the knowledge gained about an active user. Web-page recommender systems are one kind of recommender systems, which can automatically recommend Web-pages that are most interesting to a particular user based on the user's current Web navigation behavior. Since a website is usually designed to show the index pages on the home page, the index pages take the role of guiding users to the content pages on the website through Web-page links, whereas with the index pages, a user usually has to navigate a number of Web-pages to reach the content page they are interested in. If the index pages of a website are not well designed, which is often the case, Web users will struggle to find useful pages and are very likely to leave the site. For a commercial website, this means losing potential customers [2]. For an e-government Website, this will mean that the citizen's needs are not satisfied. Therefore, Web-page recommender systems have become increasingly valuable for helping Web users to find the most interesting and useful Web-pages on specific websites. Good Web-page recommendations can improve website usage and Web user satisfaction [6].

2. LITERATURE SURVEY

The theme of this literature review is semantic-enhanced recommender systems. Web mining and ontology are the two pillar techniques to semantic-enhanced recommender systems, and evaluation metrics are the core part to make sure that the techniques are useful for semantic recommender systems [5]. Web mining techniques are used to discover useful patterns from Web data. They can be classified into three streams as Web content mining, Web structure mining and Web usage mining. Web usage mining stream is the focus of Web mining in this study. Ontology is an advanced knowledge organization technique as the backbone of the Semantic Web technology.

2.1 WEB MINING

Web mining (WM) is the process of discovering useful knowledge from Web data. Depending on different types of Web data, appropriate mining techniques are selected. There are three main broad categories of Web mining.

- Web content mining (WCM) is used to mine Web content, such as HTML or XML documents.

- Web structure mining (WSM) focuses on Web structure, such as hyperlinks on Webpages.

- Web usage mining (WUM) is applied to Web usage data, such as Web logs or clickstreams, from a website.

2.2 WEB USAGE MINING

Web usage mining aims to discover some useful patterns from the Web usage data, such as, clickstreams, user transactions and users‟ Web access activities, which are often stored in Web server logs. A Web server log records user sessions of visiting Web-pages of a website day by day. It can be used to discover potentially useful Web usage knowledge,

E.g. the navigational behavior of Web users [10]. Generally speaking, a Web usage mining process includes three phases: pre-processing, mining, and applying mining results. After pre-processing Web log files, Web access sequences (WAS), for example, are generated and filed in a dataset. An element of this dataset is a sequence of representing a user browsing session. In the mining phase, some sequential pattern mining techniques, such as clustering, classification, association rules, and sequential pattern discovery can be applied to the WAS to extract the frequent Web access patterns (FWAP), which is useful Web usage knowledge. In the third phase, the discovered knowledge will be used in a specific application, e.g., a Web-page recommender system, in which FWAP are used for generating the recommendation rules to support on-line Web-page recommendation. The mining phase using sequential pattern mining techniques is the core phase in a WUM process and plays a crucial role in a Web-page recommender system to support users to make better decision based on their current Web navigation history.

2.3 ONTOLOGY

The Semantic Web has been proposed by Tim Berners-Lee in 2000s, as an extension of the current Web, namely a machine-readable Web which facilitates human-computer cooperation. The vision of the Semantic Web is to enable machines to interpret, understand, and process information in the World Wide Web in order to respond users‟ requests. Ontology has been considered as the backbone of the Semantic Web technology for representing and sharing knowledge between Web applications since 1980s [13].

According to Gruber, ontology is an explicit and formal specification of a conceptualization. It defines a set of representational primitives that are relevant for modeling a domain of knowledge or discourse. The representational primitives typically consist of a set of concepts (or entities within a domain), relationships (that may exist among these concepts), and properties (or attributes that distinguish each concept) [6].

There should be three main components in an ontology, as follows:

- Domain terms (concepts),

- Relationships between the terms (concepts), and -

Features of the terms and relationships.

2.4 RECOMMENDATION SYSTEMS

Recommender systems were developed to learn Web user experience in order to model the interaction between users and items described on Web-pages and to recommend the interesting items to the users. The popularity of recommender systems is increasing with the rapid growth of the Internet since the mid-1990s [1]. In the systems, recommended items may be Web-pages (links), articles, books or products. An intelligent recommender system will support Web users to make better decisions to rapidly reach their own target pages during a browsing session. Therefore, recommender systems become more and more important in Web-based applications, such as e-commerce, e-government, and e-services.

At the beginning, traditional recommender systems have been developed merely based on Web mining. Web recommendations mostly rely on the informally represented data patterns which are discovered from Web data, e.g., Web server log files, user profile, and Web content. In the 2000s, the advent of the Semantic Web has changed the World Wide Web [5].

The Semantic Web offers a good basis to enrich Web mining by discovering the semantics in the data and make the discovered knowledge explicit. Semantic Web mining has emerged as an advanced technique which can improve the effectiveness of Web mining based on the ontology technology. Ontology has significantly contributed to semantic knowledge representation in order to semantically enhance information process in recommender systems, as known as semantic (enhanced) recommender systems [1].

2.4.1 TYPES OF RECOMMENDER SYSTEMS

There are six types of recommender systems that vary in terms of the used knowledge, the addressed domain, and the recommendation algorithm [8].

- **Content-based**: The system recommends items that are similar to the ones that the user liked in the past. The similarity of items is calculated based on their features.

- Collaborative filtering: In the system, an active user who is surfing the Web is suggested items that other users with similar taste liked in the past. The similarity in taste of users is calculated based on the similarity in their activity history. This technique is considered to be the most popular one in recommender systems.

- **Demographic:** The system recommends items based on the demographic profile of the user.

- **Knowledge-based:** The system recommends items based on explicit domain knowledge about how certain item features meet the user needs and references, and how the items are useful for the user.

- **Community-based:** The system recommends items based on the references of the user‟s friends.

- **Hybrid recommender systems:** These recommender systems combine some of the above mentioned techniques by taking the advantages of the used techniques to optimize Web recommendation.

2.5 SEMANTIC RECOMMENDER SYSTEM

The integration of semantic knowledge with Web mining plays an important role in the development of robust recommender systems. In particular, domain ontology is useful for clustering documents, classifying pages or searching subjects. Ontology concepts could help to enhance a Web personalization process with content semantics. In this process, ontology is built with the concepts extracted from the documents, so that the documents can be clustered based on the similarity measure of ontology terms [5]. 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. Using ontology, the relationships between document categories accessed by users can be formally represented in order to design effective Web mining models. Ontology can be extracted from accessed documents to represent the concept models of Web user information needs.

2.5.1 SEQUENTIAL PATTERN MINING ALGORITHMS

Sequence mining algorithms focus on discovering useful patterns in sequential datasets. Real world sequential datasets include e-commerce transactions, banking transactions and customer queries. The goal of sequence mining algorithms is to extract frequent Web access patterns from WAS. A frequent Web access pattern is a sequence of the events that frequently occurred in a specific order. A sequential pattern is a subsequence of a Web access sequence. A support threshold is used to filter FWAP from WAS.

2.5.1 WAP TREE BASED APPROACH

The WAP-tree based approach aims to build a tree-structure of the WAS database to overcome the limitations of Apriori-like algorithms and to extract frequent patterns efficiently. At the beginning, Han et al. proposed a frequent pattern tree (FP-tree) to store the crucial information of frequent patterns. Based on the FP-tree, a large database can be compressed into a condensed and smaller data structure so that mining the complete set of frequent patterns becomes faster and more effective. More importantly, using a FP-tree avoids the problems of multiple database scan and the generation of an explosive amount of candidates.

In addition, using a FP-tree converts the search for the FPs in the large search space to traverse a FP-tree so that the search becomes more efficient. To fulfill the potential of a FP-tree, Pei et al. proposed a highly compressed data structure, namely Web access pattern tree (WAP-tree), to store the database of WAS [7]. Constructing a WAP-tree entails scanning the database twice: the first scan is to find all frequent individual events, and the second is to construct a WAP-tree over the set of frequent individual events of each session. The WAP-tree facilitates the development of mining algorithms which can handle a large database of Web access patterns, such as WAP-Mine, Conditional Sequence Mining (CS-Mine) and pre-order linked WAP-tree mining (PLWAP-Mine) [9]. By mining a WAP-tree for frequent patterns, these algorithms avoid the problem of generating the explosive number of candidates as encountered in the Apriori-based algorithms. Some experimental results have shown that the WAP-based algorithms perform faster than traditional sequence mining techniques (e.g., the Apriori-based algorithms). The main reason is that the WAP-based algorithms use the compact structure of WAP-tree, which is a fundamental advancement compared with the Apriori-based algorithms. Moreover, the WAP-tree allows some novel sequence search strategies effectively used in the mining process. For example, the WAP-Mine algorithm recursively constructs the intermediate conditional WAP-tree by using the conditional suffix patterns. Two outstanding algorithms

From this approach are PLWAP-Mine and CS-Mine.

2.5.2 PRE-ORDER LINKED WEB ACCESS PATTERN TREE MINING

The PLWAP-Mine algorithm scans the database of WAS twice to find all frequent individual events and construct a PLWAP-tree over the set of individual frequent events. While constructing the PLWAP-tree [9], the binary position codes are assigned to each node of the tree. The binary code assignment technique is performed by using a rule similar to Huffman code generation. Based on the position codes, the algorithm can determine the suffix trees of any prefix event of frequent patterns. Therefore, the algorithm can recursively mine the PLWAP-tree using common prefix pattern search to find out all FWAP. Frequent m-sequences are computed and discovered using frequent (m-1)-

Sequences and the appropriate suffix sub-trees. As a result, a complete set of frequent patterns are efficiently discovered from the search space, i.e., the PLWAP tree.

2.5.3 CONDITIONAL SEQUENCE MINING

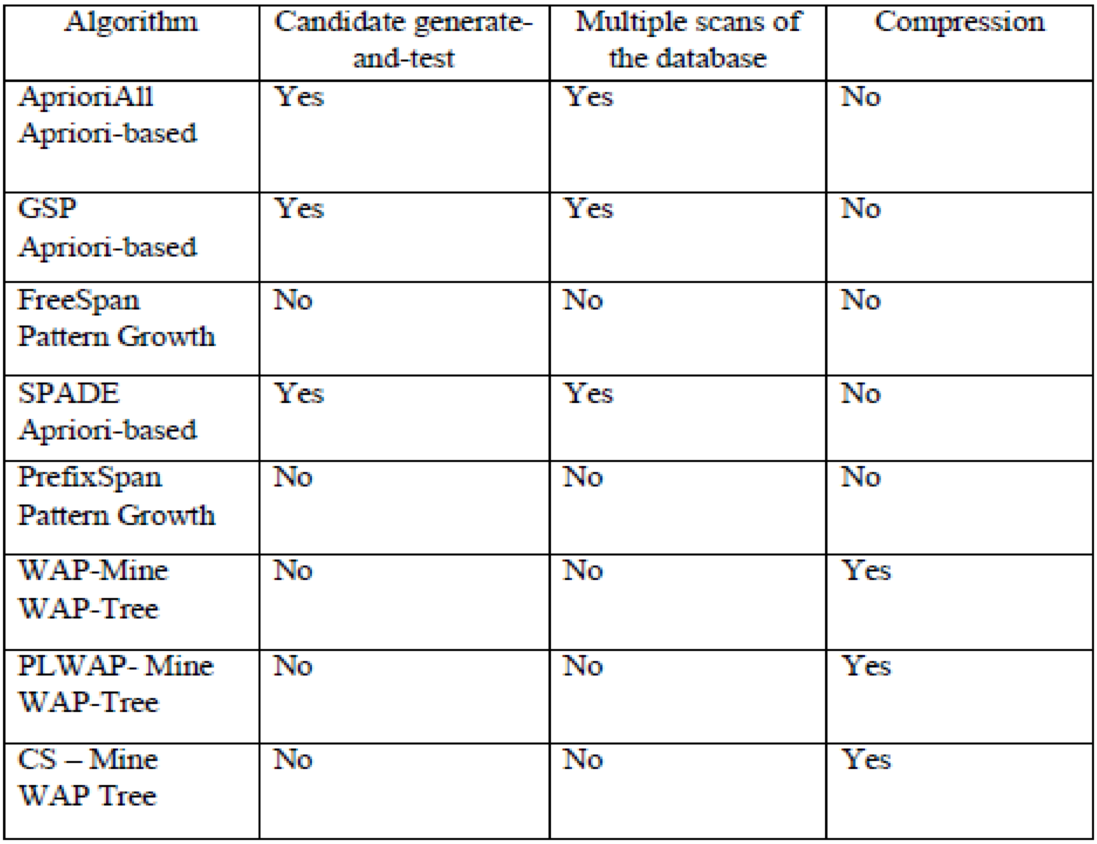
The CS-Mine algorithm scans the database of WAS once to build a WAP-tree. Given the WAP-tree, the conditional sequence base of each frequent event is initialized. That means the database is sub-divided to search frequent patterns based on these conditional sequence bases.If the combination of the conditional sequences of each frequent event is a single sequence, then a frequent pattern will be generated [12]. If this combination is failed, then the sub-conditional sequence base of each frequent event will be re-constructed and the algorithm recursively test if a frequent pattern exists. Two major processes in this algorithm are database division and the combination of conditional sequences. The database division makes search space smaller than other sequence mining algorithms, while the combination of conditional sequences can obtain a set of frequent patterns in a limited search space. As a result, the algorithm might miss out some frequent patterns because it does not consider the combination of all sequences in the entire search space. Although experimental results show that CS-Mine algorithm [12] is significantly fast in cases of smaller support threshold and larger database size, compared with other sequence mining algorithms, the effectiveness of the frequent WAP obtained from this method is questionable for making user behavior predictions.

2.5.4 COMPARISON OF ALL TECHNIQUES

- The main drawback of most of the Apriori-based algorithms is that the bottlenecks of candidate generating-and-testing may occur when applied to mine long sequential patterns due to the huge set of candidates that need to be generated and the need of multiple scans of database. In other words, most of these algorithms have the limitations of costly candidate generation and multiple database scan, so their performance is often not satisfactory in mining long patterns.
- the algorithms from the pattern-growth approach perform faster than the Apriori based algorithms except for SPADE which is faster than Free Span. Generally speaking, they are more efficient in dealing with large sequence databases because they avoid the expensive candidate generate-and-test and reduce the times of database scan.

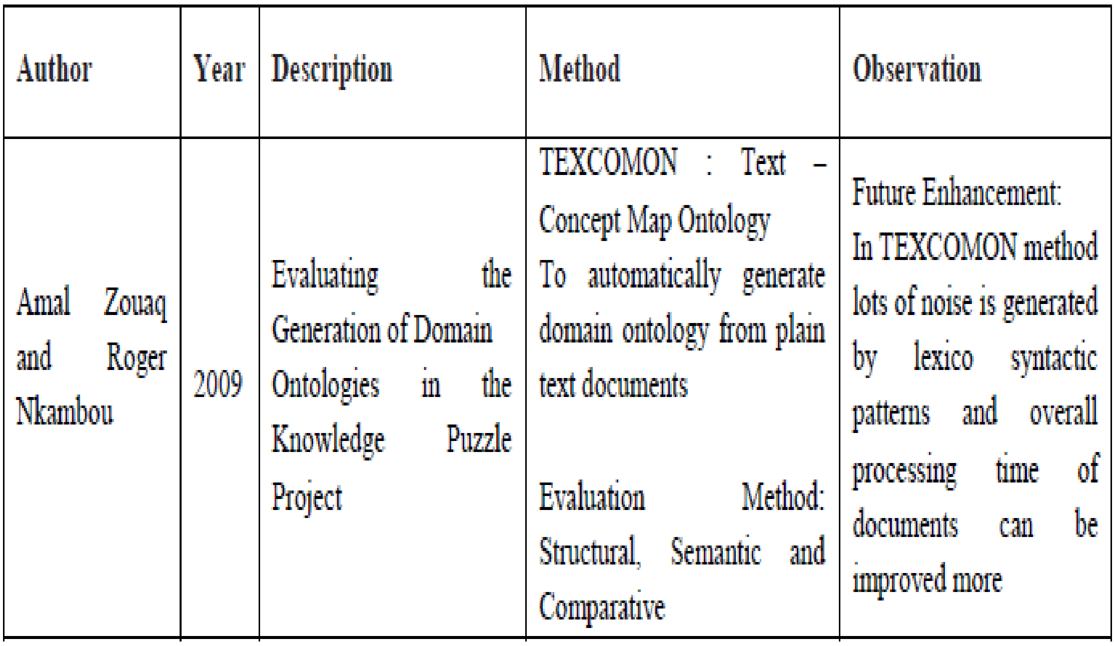
- Overcoming the problems of candidate generation-and-test and multiple database scan, the WAP-tree based approach aims to build a WAP-tree, an effective solution of data storage and retrieval. Based on the WAP-tree, sequence mining strategies can be applied more efficiently. As explained earlier, WAP-Mine encounters the costly reconstruction of intermediate WAP-trees during mining, while CS-Mine and PLWAP-Mine avoid this problem.

- The performance comparisons of the sequential pattern mining algorithms from the Above mentioned three approaches are summarized in Table 3-1, in which the algorithms are sorted in descending order of the execution time. It can be seen that Prefix Span is more efficient than the Apriori-based algorithms because it avoids the generation-and-test and multiple scans of the database. However, Prefix Span cannot compete with WAP-Mine, CS-Mine and PLWAP-Mine because they use an uncompressed data structure. In other words, the tree-based algorithms outperform the Other sequence mining techniques. Among the WAP-tree based algorithms, PLWAPMine and CS-Mine outperform WAP-Mine.

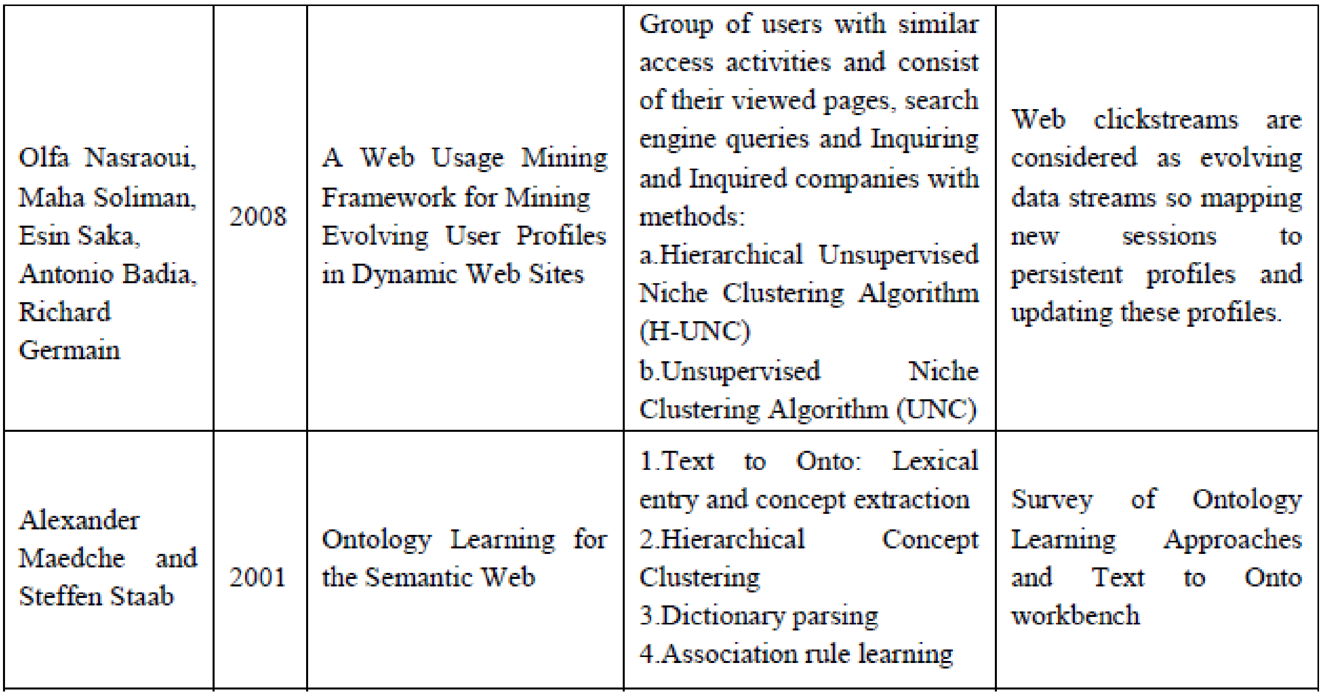


Algorithm	Candidate generate-and-test	Multiple scans of the database	Compression
AprioriAll Apriori-based	Yes	Yes	No
GSP Apriori-based	Yes	Yes	No
FreeSpan Pattern Growth	No	No	No
SPADE Apriori-based	Yes	Yes	No
PrefixSpan Pattern Growth	No	No	No
WAP-Mine WAP-Tree	No	No	Yes
PLWAP- Mine WAP-Tree	No	No	Yes
CS – Mine WAP Tree	No	No	Yes

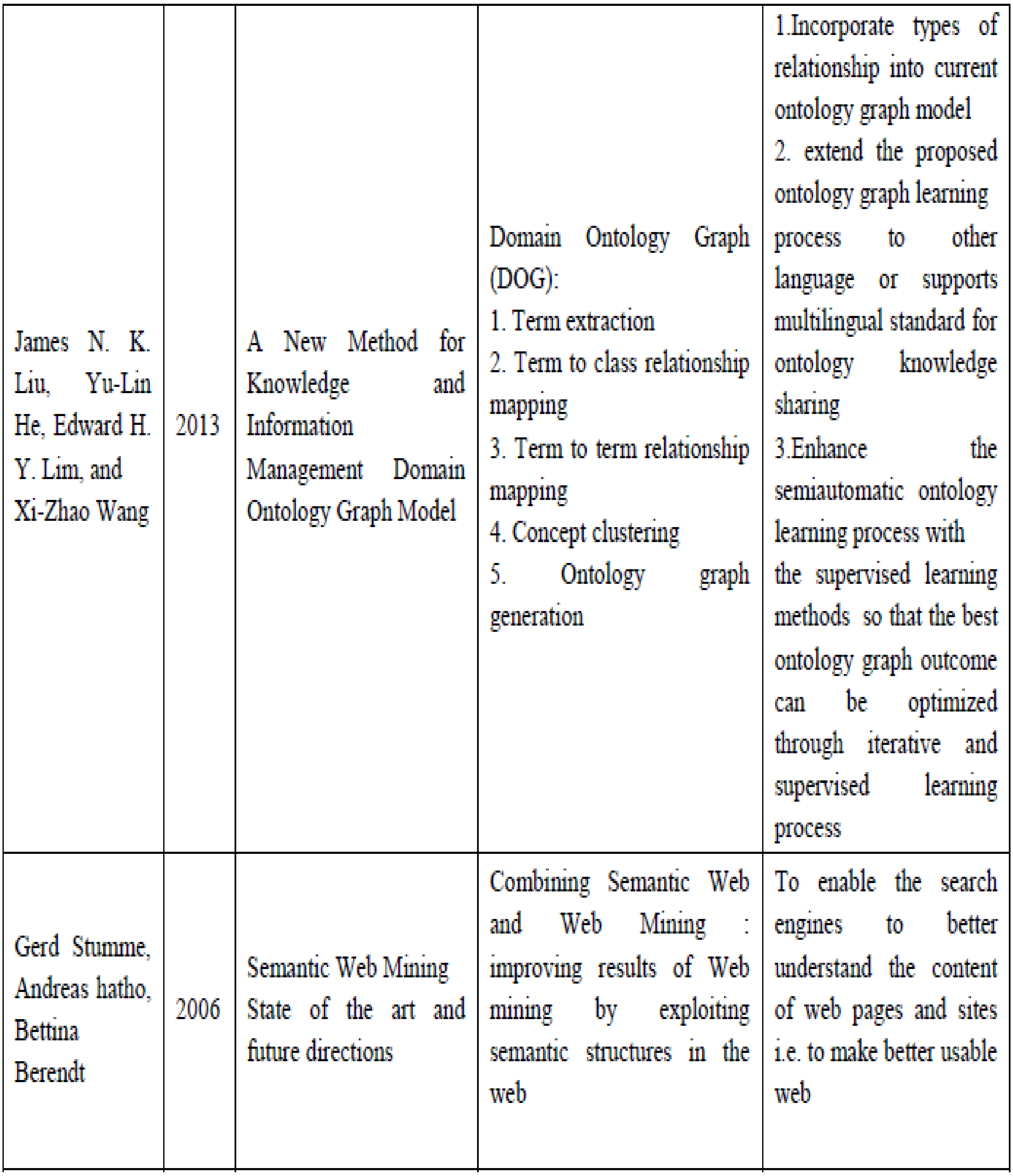
TABLE1. Comparison of Sequential Pattern Mining Algorithms



Author	Year	Description	Method	Observation
Amal Zouaq and Roger Nkambou	2009	Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project	TEXCOMON : Text – Concept Map Ontology To automatically generate domain ontology from plain text documents Evaluation Method: Structural, Semantic and Comparative	Future Enhancement: In TEXCOMON method lots of noise is generated by lexico syntactic patterns and overall processing time of documents can be improved more



Olfa Nasraoui, Maha Soliman, Esin Saka, Antonio Badia, Richard Germain	2008	A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites	Group of users with similar access activities and consist of their viewed pages, search engine queries and Inquiring and Inquired companies with methods: a.Hierarchical Unsupervised Niche Clustering Algorithm (H-UNC) b.Unsupervised Niche Clustering Algorithm (UNC)	Web clickstreams are considered as evolving data streams so mapping new sessions to persistent profiles and updating these profiles.
Alexander Maedche and Steffen Staab	2001	Ontology Learning for the Semantic Web	1.Text to Onto: Lexical entry and concept extraction 2.Hierarchical Concept Clustering 3.Dictionary parsing 4.Association rule learning	Survey of Ontology Learning Approaches and Text to Onto workbench



James N. K. Liu, Yu-Lin He, Edward H. Y. Lim, and Xi-Zhao Wang	2013	A New Method for Knowledge and Information Management Domain Ontology Graph Model	Domain Ontology Graph (DOG): 1. Term extraction 2. Term to class relationship mapping 3. Term to term relationship mapping 4. Concept clustering 5. Ontology graph generation	1.Incorporate types of relationship into current ontology graph model 2. extend the proposed ontology graph learning process to other language or supports multilingual standard for ontology knowledge sharing 3.Enhance the semiautomatic ontology learning process with the supervised learning methods so that the best ontology graph outcome can be optimized through iterative and supervised learning process
Gerd Stumme, Andreas hatho, Bettina Berendt	2006	Semantic Web Mining State of the art and future directions	Combining Semantic Web and Web Mining : improving results of Web mining by exploiting semantic structures in the web	To enable the search engines to better understand the content of web pages and sites i.e. to make better usable web

Sinead Boyce, Claus Pahl	2011	Developing Domain Ontologies for Course Content	1.a. Domain and source 1.b. Purpose and Scope 2. Existing domain ontologies 3. Important domain terms 4. Class/Concept hierarchy 5. Concept properties (Internal Structure) 6. Concept Facets (Constraints and Types) 7. Class/Concept Instances	This ontological model can provide an interface to the content. A combination of concept ontology and associated content can be used to generate a separate content representation.
C.I. Ezeife	2012	Mining Web Log Sequential Patterns with Position Coded Pre-Order Linked WAP-Tree*	Frequent m-sequences are computed and discovered using frequent (m-1) sequences and appropriate suffix sub-trees.	Complete set of frequent patterns are efficiently discovered from the search space that is PLWAP Tree.

TABLE2. Survey of Reference papers

3. PROPOSED SYSTEM

There are three models in the system to effectively recommend the web pages to the user as per his search behavior and interest. The models are as follows:

Model 1: ontology based domain knowledge model which represents domain knowledge of a website which is done as follows:

1. Collect the terms

A) Web log file

B) Preprocessing □ List of URLs

C) Crawl the webpages and extract the titles

D) Extract terms from titles

2. Define the Concepts

3. Define Taxonomic and Non-taxonomic relationships

A) Top down Development Process

B) Bottom up Development Process

C) Hybrid

Domain Ontology is constructed at 3 Levels:

1. General Level: General domain terms of Webpages and Relationship definition sets.

2. Specific Level: Specific domain terms and their relationships.

3. Webpage Level: All pages within website and association relationships between webpages and terms

Model 2: Semantic Network of Domain Terms which is knowledge map that represents domain terms, their associations and web-pages. This can be described as follows:

1. Collect the titles of visited web-pages

2. Extract term sequences from the webpage titles

3. Build the semantic network - TermNetWP

Each node represents a term in the extracted term sequences and the order of the terms in sequences determines the „from-Instance" and „to-Instance" relations of a term between other terms.

Model 3: Conceptual Prediction Model CPM which is Navigation network of domain terms based on frequently viewed webpages and represent the integrated web usage and domain knowledge for supporting webpage prediction.

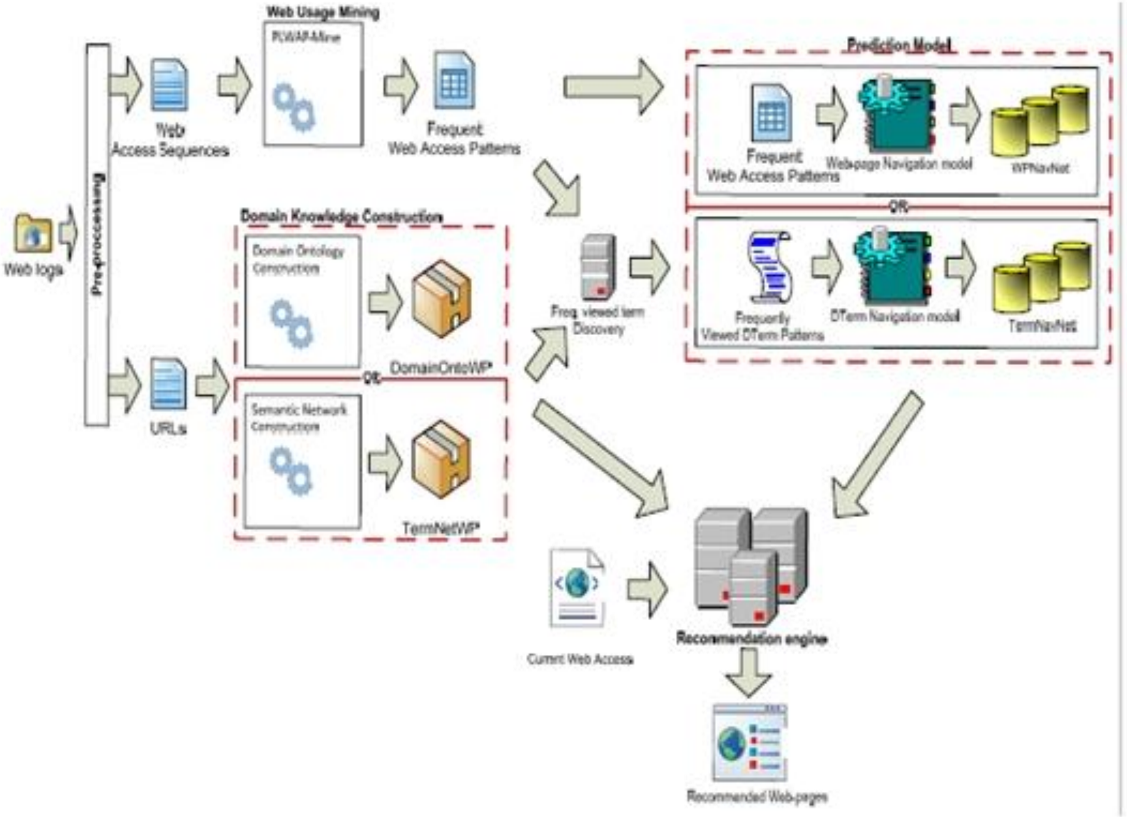

Web Usage Mining
PLWAP-Mine
Web Access Sequences
Frequent Web Access Patterns
Pre-processing
Web logs
URLs
Domain Knowledge Construction
Domain Ontology Construction
DomainOntoWP
OR
Semantic Network Construction
TermNetWP
Prediction Model
Frequent Web Access Patterns
Web-page Navigation model
WPNavNet
OR
Frequently Viewed DTerm Patterns
DTerm Navigation model
TermNavNet
Freq. viewed term Discovery
Current Web Access
Recommendation engine
Recommended Web-pages

FIGURE 1. Framework of Web page Recommendation System

4. SUMMARY AND CONCLUSION

Semantic Network of Domain Terms which is knowledge map that represents domain terms, their associations and web-pages. This can be described as follows this paper has presented a new method to offer better web page recommendations through semantic enhancement by three knowledge representation models. The proposed method can substantially enhance the performance of web page recommendation in terms of precision and satisfaction.

REFERENCES

[1] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 10, October 2014.

[2] James N.K. Liu, Yu-Lin He, Edward H.Y. Lim, Xi-Zhao Wang, "A New Method for Knowledge and Information Management Domain Ontology Graph Model" IEEE Transactions on Systems, MAN, and Cybernetics Systems, Vol. 43, No. 1, January 2013.

[3] Olfa Nasraoui, Maha Soliman, Esin Saka, Antonio Badia, Richard Germain, "A Web Usage Mining Framework for Mining Evolving user Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 2, February 2012.

[4] Amal Zouaq and Roger Nkambou, "Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 11, November 2009.

[5] Gerd Stumme, Andreas hatho, Bettina Berendt, "Semantic Web Mining State of the art and future directions", Web Semantics: Science, Services and Agents on the World Wide Web 4 (2006) 124-143.

[6] Jun S. Boyce and C. Pahl, "Developing domain ontologies for course content", Educ. Technol. Soc., vol. 10, no. 3, pp. 275-288, 2007.

[7] C. Ezeife, Y. Liu, "Fast incremental mining of Web sequential patterns with PLWAP tree", IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 4, pp. 721-732, Apr. 2012.

[8] N. R. Mabroukeh, C. I. Ezeife, "Semantic-rich Markov models for Web prefetching", ACM Comput. Surv., vol. 34, no. 1, pp. 1-27, Mar. 2008.

[9] C. I. Ezeife and Y. Lu, "Mining Web log sequential patterns with position coded preorder linked WAP-tree", IEEE Intell. Syst., vol. 16, no. 2, pp. 72-81, Oct. 2011.

[10] B. Liu, B. Mobasher, and O. Nasraoui, "Web usage mining," in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Germany: Springer-Verlag, 2011, pp. 527-603.

[11] T. T. S. Nguyen, H. Lu, T. P. Tran, and J. Lu, “Investigation of sequential pattern mining techniques for Web recommendation,” Int. J. Inform. Decis. Sci., vol. 4, no. 4, pp. 293-312, 2012.

[12] B. Zhou, S. C. Hui, and A. C. M. Fong, “CS-Mine: An efficient WAP-tree mining for Web access patterns,” in Proc. Advanced Web Technologies and Applications. vol. 3007. Berlin, Germany, 2004, pp. 523-532.

[13] D. Oberle, S. Grimm, and S. Staab, “An ontology for software,” in Handbook on Ontologies, vol. 2, S. Staab and R. Studer, Eds. Berlin, Germany: Springer, 2009, pp. 383-402.