Vol 05, Article 04437; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

A STUDY OF WEB RECOMMENDATION SYSTEMS AND WEB DOCUMENT CLUSTERING BASED ON DATA EXTRACTION

S. SUNEETHA[1], M. USHA RANI[2]
Research Scholar[1], Professor & BOS Chair Person[2],
Department of Computer Science, SPMVV, Tirupati.
suneethanaresh@yahoo.com[1], musha_rohan@yahoo.com[2]

ABSTRACT
World Wide Web is the largest and the most popular resource of varied and rich information. Despite enormous growth of the continuously changing Web, requisite information is not effectively mined. Majority of users are browsing a lot to get the desired information. Moreover, due to the heterogeneity and lack of structure of the Web data, Web data extraction is treated as a difficult task. This offers challenges for mining the Web. The main purpose of Web Recommender Systems is to provide the right information to the right user in right time. The obtained Web documents can be clustered to increase the effectiveness of information retrieval and to ease decision making after distinguishing and extracting relevant information from the noisy content. The central theme of this paper is to study the current approaches and concepts related to Web Recommendations, Web Data Extraction and Clustering with a view towards improvising them in the near future.

Keywords— Web Mining, Web Personalization, Web Recommender Systems, Web Data Extraction, Web Document Clustering.

I. INTRODUCTION

The Web is drowning with data World-Wide, but the thrust for requisite information is still in its way. Searching the Web for desired information is similar to dragging the Web across the surface of an ocean [4]. Users seeking information of their specific interest traverse to and fro across the Web via links and this offers ample opportunities and challenges for Data Mining.

'Web Mining' is the process of analysing Web data to discover knowledge from it. 'Web Personalization' is an action that upgrades the data or services offered by a Web site according to the user(s) requirements using the knowledge from the user's navigational behavior and individual interests. Web Recommendation Systems, a major application of Web Personalization Systems provide relevant Web pages to the requested user by personalizing various Websites.

Current recommender systems may not offer complete professional services because of several limitations of Web mining and this call for improvising them. A high potential Web mining technique is desired to obtain the required information in a timely manner. Classification and Clustering the Web pages increases the efficiency and effectiveness of information retrieval by arranging Web pages to support their contents.

Different Websites may not provide completely relevant information. Though the obtained Web pages appear informative to the users, they may not be structured and may vary in formats, patterns or types. As a result of this non-uniformity and lack of structure, data integration and finding desired information will become complex. Detection of data region in the Web page and presentation of information located on the Web to the user are problems in Web data extraction.

Semantically similar objects are usually clustered together and resemble each other. Visual clues about where the content to be extracted reside will ease information extraction. In this paper, a study of Web recommendation systems, Web Document Clustering and Web Data Extraction is done with the intent of improvising them.

II. STUDY OF WEB RECOMMENDER SYSTEMS

Web Recommender Systems are active filtering systems that provide desired information – Web pages, music, movies, news, books etc., to the user. The goal of intelligent recommendation system is to determine the Web pages that are likely to be accessed next by the current user in the near future, based on the past behaviour patterns of the user in the Web.

A. Web Personalization.

'Web Personalization' is the process of customizing the Web sites according to the specific needs of the user, by taking advantage of the knowledge from the analysis of the user's navigational behaviour in correlation with the information collected in the Web context (content, structure and user profile data) [10]. The purpose of a Web personalization system is to provide users with the required information with minimal effort.

Vol 05, Article 04437; December 2015 International Journal of Engineering Sciences Research-IJESR
http://ijesr.in/ ISSN: 2230-8504; e-ISSN-2230-8512

The high-level steps of the Web Recommendation process:
Data Collection->Data Pre-Processing->Web Data Mining/Analysis-> Web Recommendations

B. Applications & Challenges of Recommendation Systems.

Web recommendation is a very active research area and based on Web users demands it is improving on a regular basis. Web Recommender systems have their relevance to information retrieval in wide-ranging areas such as, Web browsing, information filtering and especially in e-commerce for converting browsers to buyers, increasing cross-sell and building loyalty. The most common recommender systems applications include:

- Entertainment - recommendations for movies, music, and TV programs.
- Content - personalized newspapers, recommendation for documents, recommendations of Web pages, e-learning applications, and e-mail filters.
- E-commerce - recommendations for consumers of products to buy such as books, cameras, PCs etc.
- Services - recommendations of travel services, recommendation of experts for consultation, recommendation of houses and sites to purchase/rent/sale or matchmaking services.

Still several challenges subsist in recommender systems. They include [5]:

- Scalability: With enormous growth of data over internet, recommender systems experience difficulty. Methods proposed for handling this scalability problem and speeding up recommendation are based on approximation mechanisms. Though they improve performance, they mostly result in accuracy reduction.
- Heterogeneity: It is difficult to develop recommendation algorithms for data acquired from different sources at different time intervals.
- Overspecialization: Users are restricted from getting recommendations that resemble theirs. Over Specialization prevents user from discovering new items and other available options. This problem can be dealt by using genetic algorithms.
- Data Sparsity: The main reason behind data sparsity is that most users do not rate most of the items and the available ratings are usually sparse. Researches attempted to alleviate this problem; but this area demands more research still.
- Cold Start Problem: It refers to the situation when a new user or item enters the system. Three kinds of cold start problems are: new user problem, new item problem and new system problem. In such cases, it is really very difficult to provide recommendations. However, content based methods can provide recommendation in this case.
- Metrics: Development and usage of metrics is complex. Selection of an appropriate metric to evaluate the accuracy of recommender systems is a challenging task.
- Privacy and Trust challenges have also to be dealt with.

C. Recommendation Methods.

Recommendation methods are classified as, pure probability-based methods and weighted probability-based methods.

Path Recommendation, History-based Recommendation, Future Prediction based and Shortcut Recommendation as their names imply are pure probability based methods that are based on the transition probabilities from the Web log data.

Weighted Probability-based methods are based on the probability weighted by factors such as, time of stay, Web policy and Web content structures. Such methods include: Weighted by time of stay, Weighted by the most recent access and Weighted by number of references.

D. Comparison of Web Recommendation Techniques

Recommendation system techniques are generally categorized into 4 major types [1] [10] as in Table 1 below:

TABLE I
COMPARISON OF WEB RECOMMENDATION TECHNIQUES

Content-based Filtering	**Collaborative Filtering**	**Rule-based Filtering**	**Hybrid Approaches**
Based on item description and a profile of user's interests.	Based on the assumption that users with similar behavior have analogous interests.	Based on "if this, then that" rules processing.	Combines content-based and collaborative methods
Explicit focus is on similarity of text.	Types are: Model-based and Memory-based.	Relies on the lists or words or regular expressions.	Makes use of Artificial Intelligence techniques.
Adequate to gather	Hinders deficient	Based on the replies to a set	Mixed, weighted,

Vol 05, Article 04437; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

feedback from customers about their precedence and does not require domain knowledge. User independent & transparent.	suggestions / recommends. Particularly applicable and useful in domains where the analysis of content is very expensive or difficult.	of questions, derived from a decision tree and arrives at results desired by user.	cascade, meta level, switching, feature augmentation and feature combination are some combination approaches used.
Yields better performance than CF at finding locally similar objects and thus widely used in recommending news, TV programs etc.,	Very effective for forecasting customer precedence in choice of objects and thus widely used in e-commerce.	Used to filter unnecessary and unimportant information as well as to block unwanted messages.	Combines advantages of other approaches and thus overcomes some of their limitations.
Can suggest objects only from a restrict theme scope and serendipitous recommends are very difficult to achieve. Cons: over-specialization, limited content analysis and new user problems.	Designed to work on enormous database. Limitations include: Sparsity in ratings, cold start, gray sheep, synonymy, shilling attacks, scalability and efficiency problems.	Difficult to make recommendations that are out of pre-defined association rules. Bias is caused by the subjective description of users or their interests as input. Time consuming, prone to false positives and the false positives are not equally distributed even.	Increased complexity and difficult to implement.

E. Literature Review

Building of Web recommendation system by Web access log is an accepted method and several related significant research works exist in the data mining literature for Web Recommendations. A review of some related works are furnished below:

Pei et al., [3] proposed PrefixSpan (i.e., Prefix-projected Sequential pattern mining) that explores prefix projection in sequential pattern mining. The main idea is to examine only the prefix subsequences and project only their corresponding postfix subsequences into projected databases. In each projected database, sequential patterns are grown by exploring only local frequent patterns. To improve the mining efficiency further, two kinds of database projections (level-by-level projection and bi-level projection) are explored along with an optimization technique that explores pseudo-projection. PrefixSpan mines the complete set of patterns and greatly reduces the efforts of candidate subsequence generation. It is fast, focused and elegant. The performance study reveals that PrefixSpan outperforms both the Apriori-based GSP algorithm and FreeSpan, in mining large sequence databases. Its extension is possible by mining sequential patterns with time constraints, time windows and/or taxonomy, and other kinds of time-related knowledge.

Niranjan et al., [7] designed a Web recommendation system based on mining closed sequential Web access patterns. The input to the proposed system is the Web server log data that describes the users visiting behavior. The Web server log data is preprocessed and then sequential Web access patterns are mined from it by using Prefix Span algorithm. Then, closed sequential Web access patterns are discovered from the mined sequential Web access patterns using post-pruning strategy. Subsequently, a pattern tree i.e., Patricia Trie based data structure is constructed from the mined closed sequential Web access patterns. For a given user's Web access sequence, the proposed system provides recommendations on the basis of the constructed pattern tree. The experimentation is performed using synthetic dataset and the performance is evaluated with precision, applicability and hit ratio.

Suneetha K et al., [10] developed a Web page recommendation algorithm using weighted sequential patterns and Markov model. The PrefixSpan algorithm is modified by incorporating the weightage constraints such as, spending time and recent visit. The weighted sequential patterns are then utilized to construct the recommendation model using the Patricia trie-based tree structure. Finally, the recommendation of the current users is done with the help of Markov model by matching the visiting path with the tree and Markov model. The synthetic dataset is utilized to analyze the performance of W-PrefixSpan algorithm as well as Web page recommendation algorithm by using precision, applicability and hit-ratio measures. From the results, the memory required for the W-PrefixSpan algorithm is less than 50% of memory needed for PrefixSpan algorithm.

Wanaskar et al., [12] proposed a Web recommendation system based on weighted association rule mining and text mining. In this approach, weight is assigned to each page to show its importance depending on the time spent by each user on a particular page or visiting frequency of each page. To add semantic knowledge to the data to be recommended, text mining is used. First, the pages are clustered based on user's usage pattern. Then,

Vol 05, Article 04437; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

the seed recommendation set is generated based on weighted association rules to include even rarely visited pages or newly added pages. Subsequently, the seed set is extended to generate candidate set and HITS algorithm is used to rank this candidate set. Finally, text mining is applied using TF-IDF algorithm by taking the above results and the page results for the current user session as the input. The so obtained results and the results from the previous are sorted for generating the final recommendation set. The performance is evaluated with precision metric. This method can be applied under cloud computing environment in the near future.
Sugana et al., [11] proposed a system for efficient Web page recommendation based on Web usage mining and Markov model, coupled with the pattern discovery algorithms - clustering and association rule mining. The traditional Apriori algorithm is improved by adding the time duration spent on each Web page. Markov model is used for recommending the Web pages based on users past history. This recommendation system consists of the following four processes: (i) Data preparation, (ii) Clustering, (iii) Finding associative patterns, and (iv) Web page recommendation. The performance of the result is evaluated based on precision, applicability and hit ratio. In future, FP-Tree algorithm with more improvements in using the minimum support value will be applied for finding the associative patterns for more accuracy.

III. STUDY OF WEB DATA EXTRACTION AND WEB DOCUMENT CLUSTERING

Today, Web is a source of rich and varied information. Information on the Web is growing and changing at an exponential rate. Even though information access has been limited to browsing and searching, information is not being effectively mined. Current search engines are providing relatively low professional services and majority of the users have to browse a lot to get the desired information. Due to the heterogeneity and lack of structure of Web information sources, data/information extraction from the Web has become a difficult task, now-a-days.
'Web Data Extraction Systems' are software applications or methods targeting at extracting information from Web sources in an efficient manner. The task of Web information extraction differs from traditional ones as traditional IE aims at extracting data from totally unstructured free texts written in natural language and Web IE processes online documents that are semi-structured and generated automatically by a server-side application program. Traditional IE takes advantage of NLP techniques whereas Web IE applies machine learning and pattern mining techniques to exploit the syntactical patterns or layout structures of the template-based documents.
Web data extraction mainly deals with unstructured or semi-structured form of data and transforming those Web pages in to program-friendly structures. Currently available data extraction approaches and tools are dependent on Web programming or design language and thus there is a great need for an appropriate information extraction methodology for providing fast, accurate and hidden information as per user's demand and to improve the efficiency of search engines as well.

A. Web Data Extraction.

Formally, an information extraction task is defined by its input and its extraction target. Web data extraction system is a software extracting data/information automatically and repeatedly from Web pages (dynamic or static) with changing contents, and then delivering extracted data to a database, email or some other application, after transforming it in to a structured format to ease further processing or storage. Thus, the main phases associated with a Web Data Extraction System are: Automation and Scheduling, Data Transformation and Use of the extracted data by interacting with Web pages and generating wrapper.

B. Applications and Challenges in Web Data Extraction

The spectrum of applications of Web data extraction is huge. Web Data Extraction has profound applications in several areas such as, Business, Bio-Informatics, Medicine, Social Web, Real-time systems and Search engines. Web data extraction techniques are used for data analysis in business and competitive intelligent systems as well as for business process re-engineering, at the enterprise level. At the social Web level, they are used to group structured data generated and disseminated continuously by Web, social media and online social network users, so as to analyse human behavior.
Context-aware advertising, customer care, database building, software engineering, business and competitive intelligence, Web process integration and channel management, functional Web application testing, main content extraction, Web experience archiving are some applications at the enterprise level. Social Media is the engine of Web 2.0. Web links have to be created between people, to share thoughts, opinions, photos, travel tips, etc. At the social level, Web data extraction techniques are used in applications such as, social networks, social bookmarks, comparison shopping, mash up scenarios, opinion mining, citation databases and Web accessibility [2].

- Context-aware advertising techniques present Web site user with commercial thematized advertisements along with the Web page content.

Vol 05, Article 04437; December 2015 International Journal of Engineering Sciences Research-IJESR
http://ijesr.in/ ISSN: 2230-8504; e-ISSN-2230-8512

- Web data extraction systems ease classifying, inferring, populating unstructured information related to customer.
- Building a database of information about a particular domain is a key concept in Web marketing.
- Extracting relevant information from social bookmarks will be easier and faster with Web data extraction systems.
- A mash up is a Web site or application that combines a number of Web sites into an integrated view. It is possible by wrapper technology.
- Citation databases help users to perform searches, comparisons, count citations, cross-references etc. Cite Seer, Google Scholar and Publish or Perish are some examples.
- Techniques for automatic data extraction are extremely helpful in making Web pages more accessible to blind and partial-sighted users.
- Web Harvesting is the most attractive future application and bio-informatics is a growing field of Web data extraction.

Web data extraction is a challenging task [6] as it has to deal with the following issues:

- The main challenge is to create an automated system for extracting data from the Web sources with high performance, accuracy, speed of information retrieval.
- It also requires a high degree of human expertise for the process of automation. Web data extraction techniques must process large volumes of data in very short time.
- Dealing with privacy or security concerns in social Web, banking and other fields is another challenge.
- Machine learning data extraction techniques require large training set of manually labeled Web pages and the task of labeling pages is time-expensive and error-prone.
- Web sources are continuously changing and these changes are unpredictable. Handling these unpredictable changes through flexibility and ease of modifications is a challenge.
- Removing noise (unwanted information) provides efficient Web data extraction and it is a challenge again.
- Extracting data from semantic Web in machine understandable form and in the form of ontologies is also a research challenge.
- Integrating Web content and usage mining to get useful and more efficient results is an additional challenge.
- Removing limitations of page ranking algorithms in Web structure mining is an added challenge.

C. Taxonomy of Web Data Extraction Techniques

Normally, the following methods [4] are used for Web data extraction:

TABLE III
TAXONOMY OF WEB DATA EXTRACTION TECHNIQUES

Tree-based Techniques	Web Wrappers	Machine Learning Approaches	Web Mining
Used to extract semi-structured or unstructured content of Web pages.	A program that 'wraps' information source so that information can be accessed without changing its core query answering mechanism.	Used to extract domain-specific information from the Web.	Used for searching hidden information and patterns from the Web pages.
Document Object Model (DOM) is used to represent the Web page.	Characterized by life cycle with phases - wrapper generation, wrapper execution and wrapper maintenance.	Rely on training sessions for acquiring domain expertise.	Deals with even unstructured data.
Easy and cheap to implement	May be semi-automated or fully automated. Faster.	May be fully automated. Statistics based techniques are also available.	Used to find information about anything.
Plagued by high computational costs.	Expensive as different program has to be developed for every Web site or script.	Requires a large training set of manually labelled Web pages that is time-expensive and error prone. Also requires boot strap period for top-level human management in the system.	Difficult to deal with abundant and varied information

Vol 05, Article 04437; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

D. Web Document Clustering.

Presentation of information located on the Web, to the user is a major part of Web data extraction. Semantically, similar objects resemble each other in the sense of human perception and so clustered together usually. Clustering is a technique in which the data objects are grouped into a set of disjoint groups called clusters, such that, objects in each cluster are analogous to each other than the objects from different clusters.

Clusters can be formed for the information extracted from the Web. Then, the user's query is matched against each document and each cluster separately to remove unnecessary and unimportant information. Thus, clustering increases the efficiency and effectiveness of information retrieval. If clusters are formed for the user behavior records in Web logs, searching time will be reduced as searching is performed in clusters rather than in the complete Web log.

Clustering techniques are used in several application areas such as, pattern recognition, data mining and machine learning and so on. Methods available for document clustering are: Text-based Clustering, Link-based clustering and Hybrid clustering. Clustering algorithms can be generally classified as, Hard, Fuzzy, Possibilistic and Probabilistic.

E. Literature Review

In data mining literature, a handful of researches are available for Web data extraction and document clustering. Reviews of few researches are summarized below:

Wei Liu et al., [14] proposed a vision-based approach to extract structured data from deep Web pages, including data record extraction and data item extraction. This approach is Web-page programming language independent and primarily utilizes the visual features on the deep Web pages. It consists of four primary steps: visual block tree building, data record extraction, data item extraction, and visual wrapper generation. A new evaluation measure, revision to capture the amount of human effort needed to produce perfect extraction is also proposed and the experimental results showed that this approach yields better performance. Still issues remained. This approach can only process deep Web pages containing one data region, though there are multi data region deep Web pages. Zhao et al., [15] have attempted to address this problem, and their solution is HTML-dependent. But, its performance has scope for improvement. In this approach, the visual information of Web pages is obtained by calling the programming APIs of IE, which is a time-consuming process. Its efficiency can be improved by obtaining visual information directly from the Web pages though new APIs.

To resolve the issue of multiple data regions of deep Web pages in Wei Liu's [13] work, Vijay et al., proposed a framework that efficiently uses the Web database. The algorithm makes use of enhanced co-citation algorithm instead of developing a new set of APIs for the extraction of visual information. The proposed algorithm retrieves visual information of the deep Web pages directly from the Web database. Empirical studies with large set of database for Web data extraction demonstrate that the performance of the proposed VBEC exhibit high precision while enabling efficient and accurate recall value with better time consumption.

Raghu et al., [8] proposed a dynamic vision-based approach for extracting data from the deep Web that consists of two phases. The first phase includes query analysis and query translation and the second one covers vision-based extraction of data from the dynamically created deep Web pages. Dynamic Deep with visual features has two components: the deep data record extractor and the deep data item extractor. It employs four steps for data extraction from deep Web page. First, it takes sample deep Web page from a particular Web database and obtains its visual representation. Later, this is converted to a visual block tree. Second, data records are extracted from visual block tree. Third, this extracted data records are further partitioned into data items and the similar data items (semantically related) are clustered together. In the fourth step, visual wrappers are generated to extract the data items and data records from the other deep Web pages which are dynamically created by the same Web database. Precision and Recall measures were used to evaluate the performance of the Deep Web. The database of deep Web pages is created for different domains, and is to be updated frequently. This process of update will require an effective algorithm to maintain the efficiency of the system. The future works may proceed to overcome this drawback.

Lavanya et al., [4] proposed a document clustering method based on vision-based deep Web data extraction. It consists of 2 steps: Vision-based Web data extraction and Web Document Clustering. In this method, Web page information is segmented into various chunks. Initially, irrelevant data such as, advertisements, images etc. are removed using chunk segmentation. A set of chunks will be obtained and from which duplicate chunks are removed using three parameters – hyperlink percentage, noise score and cosine similarity. For each chunk, three parameters - title word relevancy, frequency-based chunk selection and position features are computed to identify the main chunks. Then, a set of keywords are extracted from these main chunks. Finally, the extracted keywords are subjected to Web document clustering using Fuzzy C-Means clustering approach. Experimental results demonstrated that the proposed VDEC method can achieve stable and good results for both synthetic datasets and general data sets. The performance of the proposed VDEC method is evaluated with precision, recall and revision ratio.

Vol 05, Article 04437; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

Reka et al. [9] proposed a Web document clustering approach that uses semantic relation between documents for enhancing Web document clustering. It consists of three phases: Preprocessing, computing document relation score and clustering. The key concepts are identified from the Web documents by preprocessing, stemming, and stop word removal. Identified concepts are used to compute the document relation score and cluster relation score. The semantic ontology was used to compute both DRS and CRS. Based on the DRS and CRS, the Web document cluster is identified and the Web page is assigned to that cluster. ClueWeb09 data set was used for the evaluation and the results revealed that this algorithm reduces the overlap and increases clustering efficiency. Moreover, it reduces the overall clustering time, time complexity and false positive indexing. This work relies on text documents alone. The work can be enhanced by incorporating images along with the text.

IV. CONCLUSION

The volume of the Web is very huge and is still fabricating today. Recommendation Systems aim at directing users through this information space, towards the resources that best meet their needs. Endowing Web systems with efficient and reliable recommendations is the target of intensive research in the recent years. Current works in the data mining literature address the existing problems in their individual way and going on towards providing better recommendation capabilities. This paper reviews the basic concepts, general challenges and some works on Web Recommendation Systems, Web Document Clustering and Web Data Extraction.

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. The basic idea of Web document clustering is arrive to at a conclusion by examining vast number of Web documents But, Web documents are complex and heterogeneous and thus automated methods for document clustering are of great importance. Web data extraction is the backbone of applications that facilitate Web Document Clustering. 'Web Data Extraction' is retrieving target information items from the Web pages. Generally, a Web page comprises of both main content blocks and noise blocks. The user is interested only in the informative main blocks but the remaining blocks reduce the precision of the mined results and speed of processing. Web data extraction has become more demanding and challenging due to the difficulty of assortment of Web structures and representation. The application of clustering techniques on Web documents, using Web data extraction, makes information retrieval much easier and less time consuming for the Web users.

V. FUTURE WORK

Web contents are expanding rapidly in support of elegant technologies and digitization to form a huge information source. Recommendation systems constitute a problem-rich research area because of the abundance of practical applications that help users to deal with information overloads and provide personalized recommendations, content and services to them. Both the industry and academia are interested in developing new approaches to recommendation systems. Despite the advances, the existing recommendation systems still suffer from inherent limitations and require further improvements to make recommendations more effectual for a broader range of applications. These issues will advance the research towards the next generation of recommendation technologies. Knowing more about user can improve the quality of recommendations. The future works may concentrate on improved modeling of users and items, incorporation of the contextual information into the recommendation process, support for multi criteria ratings, and provision of a more flexible and less intrusive recommendation process.

Dynamic Web has abundant information in it. Obtaining more précised desirable information in less time is of an immense advantage. Majority of the Web contents are either unstructured or semi structured. The intrinsic need for structured information urged researchers to concentrate on various strategies for automatic extraction of information from the available Web sources. The structured data that is extracted can be used for processing Web-based applications in real time. Moreover, to tap the available data resources, an efficient method that automatically discovers the main content in the Web page and allots substantial measures for different areas in the Web page, is desirable. Techniques proposed till now have inherent limitations and the research should proceed towards overcoming them and make the information retrieval tasks much better. 'Web Harvesting' is the process of gathering and integrating data from various heterogeneous Web sources. It remains an open problem with large margin of improvement as the amount of gathered data is many times greater than the extracted data.

REFERENCES

[1] Amir Hossein Nabizadeh Rafsanjani, Naomie Salim, Atae Rezaei Aghdam, Karamollah Bagheri Fard", "Recommendation Systems: a Review", IJCER Vol. 3, Issue 5, May 2013 pp. 47-52

[2] Emilio Ferrara, Giacomo Fiumara and Robert Baumgartner, "Web Data Extraction, Applications and Techniques: A Survey", ACM Transactions on Computational Logic, Vol. V, No. N, June 2010, Pages 1-20

Vol 05, Article 04437; December 2015
http://ijesr.in/
International Journal of Engineering Sciences Research-IJESR
ISSN: 2230-8504; e-ISSN-2230-8512

[3] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl and Helen Pinto, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth"
[4] Lavanya M, Dr. Usha Rani M, "Vision-Based Deep Web Data Extraction for Web Document Clustering", Global Journal of Computer Science & Technology, Vol. XII Issue V, Version I, Mar 2012
[5] Mani Madhukar, "Challenges & Limitations in Recommender Systems", IJLTET Vol.4 Issue 3 September 2014
[6] Neeraj Raheja and Dr. Katiyar V K, "A Survey on Data Extraction in Web Based Environment", IJSWS 13-275; © 2013
[7] Niranjan U, Dr. R.B.V. Subramanyam, Dr. Khanaa V "An Efficient System Based On Closed Sequential Patterns for Web Recommendations", IJCSI Vol. 7, Issue 3, May 2010
[8] Raghu D, Sridhar Reddy, Raja Jacob, "Dynamic Vision- Based Approach in Web Data Extraction", IJCSIT Vol. 2, No. 6, 2011, pp. 2734-2736
[9] Reka M & Dr.Shanthi N, "Relation Based Mining Model for Enhancing Web Document Clustering", IJET ISSN: 0975-4024 Vol. 6 No 2 Apr-May 2014
[10] Suneetha K & Dr.Usha Rani M, "Web Page Recommendation Approach Using Weighted Sequential Patterns and Markov Model", Global Journal of Computer Science and Technology Vol.12 Issue 9 Version 1.0 Apr 2012
[11] Suguna R and Sharmila D, "An Efficient Web Recommendation System using Collaborative Filtering and Pattern Discovery Algorithms" International Journal of Computer Applications (0975 – 8887) Volume 70– No.3, May 2013
[12] Ujwala H.Wanaskar, Sheetal R.Vij, Debajyoti Mukhopadhyay, "A Hybrid Web Recommendation System based on the Improved Association Rule Mining Algorithm"
[13] Vijay R, Dr. Prasadh K, "A Vision Based Approach for Web Data Extraction Using Enhanced Co citation Algorithm", IJCSI Vol. 10, Issue 5, No 2, Sep 2013
[14] Wei Liu, Xiaofeng Meng, Weiyi Meng," ViDE: A Vision- Based Approach for Deep Web Data Extraction", Transactions on Knowledge and Data Engineering, Vol. 22, 2010
[15] Zhao H, Meng W, and Yu C T, "Automatic Extraction of Dynamic Record Sections from Search Engine Result Pages," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 989-1000, 2006

AUTHORS

Ms. S SUNEETHA received her Bachelor's Degree in Science and in Education, Master's Degree in Computer Applications (MCA) from SVU, Tirupati and M.Phil. in Computer Science from SPMVV, Tirupati. Currently, she is pursuing her Ph.D. in SPMVV, Tirupati. She is a life time member of ISTE.
Her areas of interest are Data Mining, Software Engineering, Big Data and Cloud Computing. She has 21 papers in National/ International Conferences/ Journals to her credit. She also attended several workshops in varied areas. She served Narayana Engineering College, Nellore, Andhra Pradesh as Sr. Asst. Professor, heading the departments of IT and MCA.

Dr. M Usha Rani is Professor & BOS Chair Person in the Department of Computer Science, Sri Padmavati Mahila Viswa Vidyalayam (Women's' University), Tirupati. She did her Ph.D. in Computer Science in the area of Artificial Intelligence & Expert Systems.
She is in teaching since 1992. She presented many papers at National and International Conferences and published articles in National & International Journals. She has also written 4 books like Data Mining - Applications: Opportunities and Challenges, Superficial Overview of Data Mining Tools, Data Warehousing & Data Mining and Intelligent Systems & Communications. She is guiding M.Phil. and Ph.D. in areas like Data Warehousing and Data Mining, Computer Networks and Network Security and Cloud Computing.