International Journal of Enterprise Computing and Business Systems

ISSN (Online) : 2230-8849

http://www.ijecbs.com

Vol. 1 Issue 2 July 2011

"E-SERVICE INTELLIGENCE IN WEB MINING"

Prof. Ms. S. P. Shinde
Bharati Vidyapeeth University, Pune
Yashwantrao Mohite institute of Management, Karad -Maharashtra, India

Prof. Ms. S. S. Kapase
Bharati Vidyapeeth, Institute of Management and Information Technology
Navi Mumbai, Maharashtra, India

Prof. Ms. S. R. Mulik
Bharati Vidyapeeth University, Pune
Yashwantrao Mohite institute of Management, Karad -Maharashtra, India

ABSTRACT

The World Wide Web is a popular and interactive medium to disseminate information today .The web is huge, diverse, dynamic, widely distributed global information service centre. We are familiar with the terms like e-commerce, e-governance, e-market, e-finance, e-learning,

e-banking etc. These terms come under online services called e-service applications. E-services involve various types of delivery systems, advanced information technologies, methodologies and applications of online services. The keyword intelligence will be the next paradigm shift in the e-services, thanks to internet technological advances. Intelligence is closely related with Artificial Intelligence. Web Mining is the technique used to crawl through various web resources to collect required information, which enables an individual to promote business, understanding marketing dynamics, and new promotions floating on the Internet etc. The taxonomy of web mining can be broadly divided into three distinct categories; according to the kinds of data to be mined they are Web Content Mining, Web Structure Mining and Web Usage Mining. There are a lot of techniques of web mining however; artificial intelligence techniques and algorithms are being used by almost all web mining tasks for their efficiency. This paper discusses the two main AI techniques; the Multi-Agent Systems and Swarm Intelligence, with some of their applications in web mining. Web mining intelligent techniques can be combined with traditional web mining approaches to improve the quality of mining.

Key Words*: Artificial Intelligence, E-services, Multi-Agent Systems, Swarm Intelligence, Web mining*

[I] **Introduction**

The World Wide Web is a popular and interactive medium to disseminate information today .The web is huge, diverse, dynamic, widely distributed global information service centre. We are familiar with the terms like e-commerce, e-governance, e-market, e-finance, e-learning, e-banking etc. These terms come under online services called e-service applications. E-services involve various types of delivery systems, advanced information technologies, methodologies and applications of online services. The keyword intelligence will be the next paradigm shift in the e-services, thanks to internet technological advances. Intelligence is closely related with Artificial Intelligence. Artificial Intelligence is a branch of computer science, which deals with study of developing intelligent systems imitating, extending and augmenting human intelligence through artificial means and techniques to realize intelligent

behaviors. E-service intelligence is a new research field that deals fundamental roles, social impacts and practical applications of various intelligent technologies on the internet based e-service applications. To provide intelligence for e-services various technologies including fuzzy logic, Expert Systems, Machine learning, Neural Network, Bayesian Network, Game theory, Optimization, Rough Sets, Data Mining, Web Data Mining, Multi Agents and Evolutionary Algorithms etc are being applied in various e-service approaches systems and applications. Intelligent techniques and methodologies are additional useful tools that have nevertheless been successfully applied to some of the most interesting e-service areas. Among all the intelligent techniques specified above this paper focuses more on introducing web mining concepts and states various AI applications in Web mining.

Web Mining

Web Mining is the technique used to crawl through various web resources to collect required information, which enables an individual to promote business, understanding marketing dynamics, new promotions floating on the internet etc. It is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining should be decomposed into these subtasks:

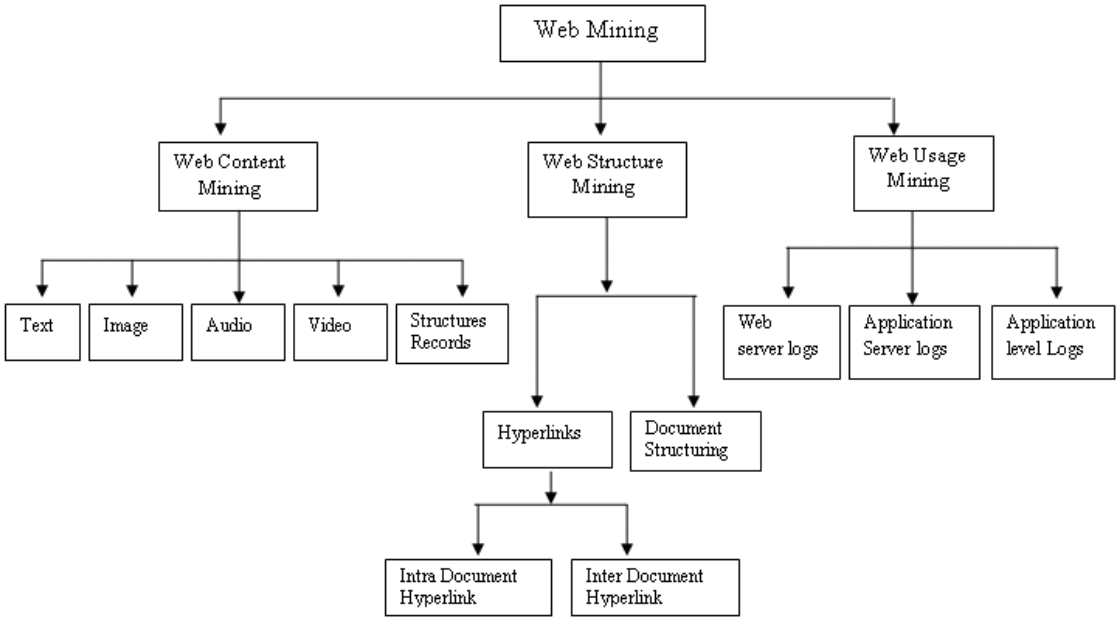
- **Resource finding:** The task of retrieving intended Web documents.
- **Information selection and preprocessing:** Automatically selecting and preprocessing specific information from retrieved Web resources. This step is transformation process retrieved in Information Retrieval [IR] process from original data. These transformations cover removing stop words, finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc.
- **Generalization:** Automatically discovers general patterns at individual Web sites as well as across multiple sites. Data mining techniques and machine learning are often used for generalization.
- **Analysis:** Validation and/or interpretation of the mined patterns. In information and knowledge discovery process, people play very important role. This is important for validation and/or interpretation in last step.

Intelligent web mining

The most important tasks addressed in the web mining literature are Classification & Association-rule discovery The two tasks involve mining web data, web customers, web documents, web-based transactions and various web based applications and each has developed a set of approaches. Also, extensive research has been carried out to integrate approaches from both sides of the tasks to conduct high performance web mining. In the meantime, some intelligence techniques including Neural Networks, Fuzzy Logic, Machine Learning, Genetic Algorithms, Bayesian Network, and Case-Based Reasoning have also been applied into web mining .Some successful examples include web user profile clustering, web document classification, web document clustering, online concept retrieval, web key information finding .One of the main aims of web mining is to find out web consumers' customer behavior patterns. By these patterns, businesses and governments can target marketing, e.g. input patterns to a web server combining with web pages so that when pattern-related customers access the website corresponding ways of marketing can be created .In general, web mining can help establish marketing patterns and customize marketing to bring right products and services to right customers. It can also establish potential customers' list to help make decisions in customer relationship management. Intelligent techniques can be combined with traditional web mining approaches to improve the quality of mining.

[II] **Web mining taxonomy**

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Following figure explains it best.



(i) *Web Content Mining:*

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Text mining and its application to Web content has been the most widely researched. Research activities in this field also involve using techniques from AI such as Information Retrieval [IR] , Natural Language Processing [NLP],Image processing and computer vision.

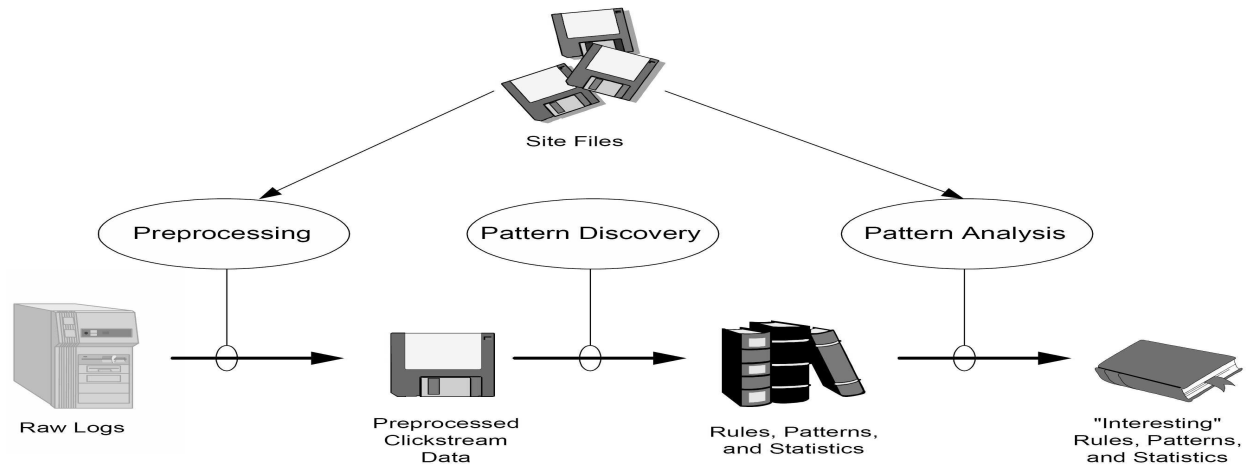
(ii) *Web Structure Mining*:

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

(a) Hyperlinks: A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink.
(b) Document Structure: The content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model [DOM] structures out of documents

(iii) *Web Usage Mining:*
Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Capturing, Modeling and analyzing of behavioral patterns of users is the goal of this web mining category. Web usage mining process can be divided into three independent tasks: Preprocessing , Pattern discovery and pattern analysis. The following figure shows this process.

Figure 3: Web usage mining process.



Web usage mining itself can be classified further depending on the kind of usage data considered:

***Web Server Data*:** They correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses, page references, and access time of the users.

***Application Server Data*:** Commercial application servers, e.g. Web logic, Broad Vision, Story Server etc. have significant features in the framework to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

***Application Level Data*:** Finally, new kinds of events can always be defined in an application, and logging can be turned on for them – generating histories of these specially defined events. The usage data can also be split into three different kinds on the basis of the source of its collection: [i] on the server side,[ii] the client side, and [iii] the proxy side. The key issue is that on the server side there is an aggregate picture of the usage of a service by all users, while on the client side there is complete picture of usage of all services by a particular client, with the proxy side being somewhere in the middle

[III] Web mining techniques

There are a lot of techniques of web mining; some few to name are as follows:

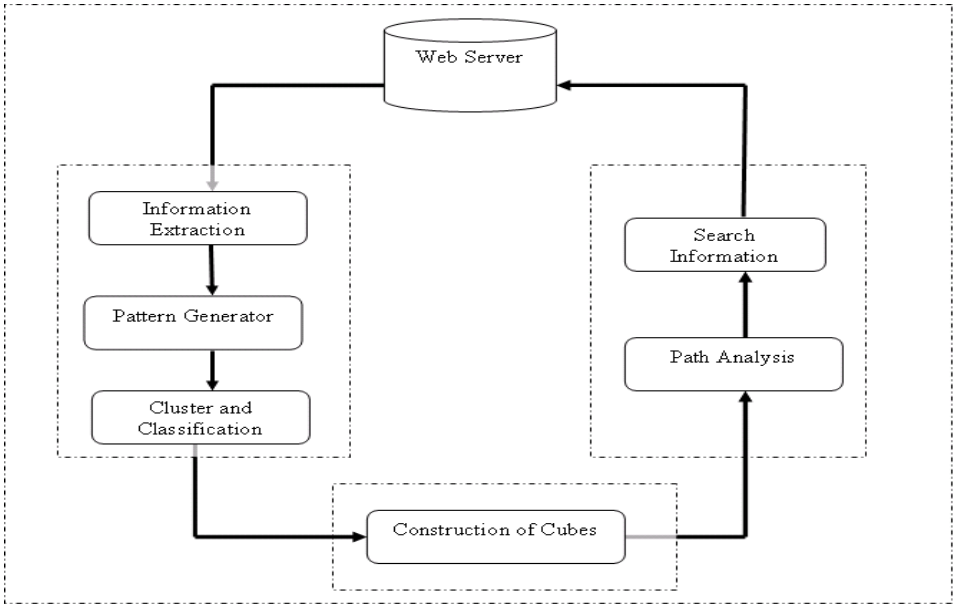
• *Cluster and Classification*: The cluster techniques identify and distribute similar Individual's behaviors in homogeneous groups. Once discoveries the profiles of each group, the characteristics of each one of them can be used to carry out and classified

• *Rules of Association*: The association rules can be seen as the identification of actions or facts that, being initially independent, they happen in a combined or associate way. The considered facts can be characteristics or behaviors observed in the individuals.

• *Path Analysis:* This technique supposes the generation of directed graphs, which represent the relationships among the web pages. The web pages are the nodes of the graphs, and the connections among the pages are the directed arcs among nodes. They can also be defined like graphs with arcs which represent the similarity among pages, or arcs that show the number of users that go from a page to another.

• *Sequential patterns*: It is a historical of transactions in a web server, where the visit of a client is stored for a period of time. The problem of discovering sequential patterns of access is based on identifying the group of more frequent accesses in a group of transactions or visits in a giving time.

• *Cubes:* A cube of data is a type of multidimensional array that allows the users to explore and to analyze a collection of data from different perspectives. From a structural perspective, the cubes of data are composed of two elements: dimensions and measures. The dimensions are categories that describe the studied factors for their analysis, and the measures are the values of the data stored in that structure.



Advanced artificial intelligence techniques for web mining

Almost all web mining tasks are using artificial intelligence techniques and algorithms in order to perform efficiently. In the following we will briefly describe some of the most valuable AI techniques; some few to name are as follows Multi-Agent Technology and Swarm Intelligence Algorithms.

***Multi-Agent Technology*:**

Agent/multi-agent systems have become an important field within artificial intelligence research. They have found a number of applications, including web mining.

(a) ***An agent*** is a computer system that is capable of independent action on behalf of its user or owner in order to satisfy design objectives. An intelligent software agent has to be autonomous, reactive, proactive, and social i.e. capable to interact with other agents, to communicate, and to negotiate. Intelligent agents can learn and adapt to new situations. Some other characteristics are also valuable: the possibility to move on an electronic network, veracity, benevolence and rationality. A multi-agent system consists of a number of agents that interact with one-another, cooperate for realizing the different tasks and are able

to negotiate and solve conflicts. Multi-agent systems or simple agents are used in almost all content mining tasks.

(b) ***A web crawler*** is a program that automatically downloads web pages. A crawler can collect information to be then analyzed and mined online or offline. Crawlers are universal, topical and focused. Adaptive topical crawlers are the most sophisticated and they are designed using different machine learning techniques, in particular classifiers to guide them through the web. Intelligent crawlers adapt to the web content and hyperlink structure. It can use a statistical model to learn to assign priorities to URLs in the considered neighborhood, based on the Bayesian interest factors derived from features An adaptive crawling algorithm that uses reinforcement learning when crawling online, without any supervised learning, is Info Spider. This crawler is inspired from artificial life models, where a population of agents live, learns, evolve, reproduce and die. The agents learn from experience. They can be rewarded or punished for their actions. Using this model, the Web is the world where agents live. Agents' actions consist in following links and visiting pages. They receive signals from the environment that are texts and link Characteristics of the pages and they learn from these signals. Each action has an energy cost that can be, for example, the size of the fetched page. Energy is gained from visiting new pages that are relevant to the topic of the query.

(c)***The wrappers*** is a programs designed to extract structured data from the web.

Swarm Intelligence

Swarm intelligence [SI] is a term introduced in 1989 in the context of cellular robotic systems and representing the collective behavior of decentralized, self-organized artificial systems, by analogy to the real world where the collective behavior of a swarm can lead to the emergence of an apparent intelligent behavior. Collective intelligence is defined as the ability of a group to solve more problems than its individual members. SI systems are in fact simple agents that are interrelated, being able to communicate one with another and to interact with their environment. The community of agents carries out a distributed problem solving. They follow simple rules and there is no centralized control. Examples from nature of SI include ant colonies, bird flocking, animal herding, bacterial growth, and fish schooling. One of the

applications of swarm intelligence algorithms to web mining is Ant Colony Optimizer. The algorithm was inspired by the behavior of ants in finding paths from the nest to food, when they create a network of pheromone trails. Two kinds of algorithms are stated. In the first one, the co-occurrence of links in web pages [user selections] was used to compute a matrix of link strengths. The second type of algorithms extracted information from a user sequential path through the web through learning rules in order to change link strengths and create new links.

[IV] Applications of artificial intelligence theories in e-services: web mining

There is much research in integrating the fields like e-services, Artificial Intelligence and web mining. Intelligent techniques can be combined with traditional web mining approaches to improve the quality of mining. Some of the existing examples are as follows

By applying fuzzy set techniques to set up corresponding marketing patterns, businesses can better predict which kind of customers are loyal for a particular time being, and therefore can help businesses hold best customers and filter potential customers precisely.

- Fuzzy Adaptive Resonance [ART] has also been used for clustering customers in groups for targeting [Jain and Krishnapuram, 2001].
- An important aspect of web mining is the clustering of customer similar profiles to create customer “segments” [Mobasher et al., 2000].Clustered user profiles are a good option when there exist insufficient data to build individual profiles. Thus, fuzzy set theory can play a major role in customer profile representations and clustering [Jain and Krishnapuram, 2001].
- Bautista et al. [2000] used a genetic algorithm to build an adaptive consumer profile based on documents retrieved by users.
- A fuzzy classification and a genetic term-selection process together provide a better utilization of valuable knowledge to learn the current and future interests of users.
- Granular computing techniques have been used in web mining applications to enhance the intelligent functionality of mining systems.

- For example, Zhang et al. [2003] used both fuzzy computing and interval computing techniques to design a fuzzy-interval data mining system for credit card companies with actual large data sets.
- Park [2000] also introduced a neural network-based data mining method to a company's internal customer data for target marketing.
- A fuzzy ART NN proposed in this study takes customer's purchasing history as input values and cluster similar customers into groups.
- Web document and web pages classification as another important web mining aspect has used NN architectures [Kohonen et al., 2000].
- In the meantime, the use of evolution-based genetic algorithms, and the utilization of fuzzy function approximation, has also been presented as possible solutions for the classification problems [Rialle et al., 1997, Petridis and Kaburlasos, 2001, Haruechaiyasak et al., 2002].
- Anagnostopoulos et al. [2004] described a probabilistic NN that classifies web pages under the concepts of business media framework. The classification is performed by estimating the likelihood of an input feature vector according to Bayesposterior probabilities.
- Web usage mining has become very critical for effective website management, creating adaptive websites, business and support services, personalization, network traffic flow analysis and so on [Abraham and Ramos, 2003].
- Krishnapuram et al. [2001] introduced the notion of uncertainty in web usage mining, discovering clusters of user session profiles using robust fuzzy algorithms.
- In the approach, a user or a page can be assigned to more than one cluster. A dissimilarity matrix is created that is used by fuzzy algorithms presented in order to cluster typical user sessions.
- The study of ant colonies behavior and their self-organizing capabilities is of interest to information/knowledge retrieval, decision support systems sciences and also web mining, because it provides models of distributed adaptive organization, which are useful to solve difficult optimization, classification, and distributed control problems.

- Abraham and Ramos [2003] proposed an ant clustering algorithm to discover web usage patterns [data clusters] and a linear genetic programming approach to analyze the visitor trends.
- Robot Detection and Filtering—Separating Human and Non human Web Behavior: Web robots are software programs that automatically traverse the hyperlink structure of the web to locate and retrieve information. The importance of separating robot behavior from human behavior prior to building user behavior models has been illustrated by Kohavi [2001].
- Tan and Kumar [2002] proposed a classification based approach that uses the navigational patterns in click-stream data to determine if it is due to a robot. Experimental results have shown that highly accurate classification models can be built using this approach. Furthermore, these models are able to discover many camouflaged and previously unidentified robots.
- Preprocessing—Making Web Data Suitable for Mining. Preprocessing of web structure data, especially link information, has been carried out for some applications, the most notable being Google style web search [Brin and Page 1998]. An up-to-date survey of structure preprocessing is provided by Desikan, Srivastava, Kumar, and Tan [2002].
- Flake, Lawrence, and Giles [2000] applied the “maximum-flow minimum cut model” to the web graph for identifying “web communities.” and comparisons were made to detect the strengths and weakness of the two methods.
- Reddyand Kitsuregawa [2002] proposes a dense bipartite graph method, a relaxation onto the complete bipartite method followed by HITS approach, to find web communities.

CONCLUSION

The World Wide Web is today the major source of data and information for all domains.Web Mining is a new area of pursuit in Computer Science. It integrates concepts and techniques from many existing areas such as Artificial Intelligence and Networking, and applies these to the Web. This paper describes Web mining an important and challenging activity that aims to

discover new, relevant and reliable information and knowledge by investigating the web structure, its content and its usage. It is the application of data mining techniques to extract knowledge from the content, structure, and usage of Web data sources. There are many applications of these techniques, for example search engines, Web analysis, Web agents, personalization services etc. This paper focuses on the two main AI techniques: the multi-agent systems and swarm intelligence, with some of their applications in web mining. The mining tasks are so complex that they cannot be efficiently performed without the support of appropriate advanced AI techniques Web Mining Intelligent techniques can be combined with traditional web mining approaches to improve the quality of mining. Fuzzy set theory, User Profile Clustering, Genetic Algorithm, Neural Networking are some of the examples of integration of intelligent and traditional web mining techniques.

Acknowledgments

The researchers are grateful to the authors, writers and editors of books and articles which have been referred for preparing the presented research paper. It is the duty of the researchers to thank all friends and remember their parents whose blessings are always with us .

References

[1] Dr. Jennifer Seitzer CPS 499/592: Introduction to Web Mining: http://homepages.udayton.edu/~jseitzer1/cps499/lec1_introduction.htm 2/3/2011 10.00am

[2] Jose Aguilar 2009 INFORMATION SCIENCE and APPLICATIONS, Issue 9, Volume 6, page 1523-1532

[3] A Web Mining System : JOSE AGUILAR,CEMISID, Departamento de Computación

[4] Han, J., Kamber, and M.: Data Mining: Concepts and Techniques, Second edition, p. 628-648.Morgan Kaufmann Publishers, 2006.

[5] Liu, B., Chang, K. Ch.: Editorial: Special Issue on Web Content Mining. SIGKDD Explorations,2004.WWW:http://delivery.acm.org/10.1145/1050000/1046457/p1-liu.pdf [January 2007].

[6] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000. Paper available on WWW: http://www.acm.org/sigs/sigkdd/explorations/issue1-2/srivastava.pdf [January 2007].
[7] Vakali, A., Pallis, G.: Web Data Management Practices: Emerging Techniques and Technologies. Idea Group Publishing, 2007.
[8] A Web Mining System:JOSE AGUILAR,VENEZUELA
[9] Advanced AI Techniques for Web Mining:IOAN DZITAC IT Department Agora University Piata Tineretului, 8, 410526 Oradea, ROMANIA