Web Mining: Prominent Applications and Future Directions

Subashini S,
Research Scholar,
Singhania University, Rajasthan
Email: suba.anandan@gmail.com

Mahesh T. R
Professor, Dept. of CSE,
TJIT, Bangalore
Email: dr.maheshtr@gmail.com

Abstract-- The WWW can be considered as a huge semi structured database, presenting all the problems implicit in semi-structured data. Extracting the structure of every HTML document is a challenging issue given the absence of predefined standard and schema. Often the schema can be derived only after the existence of data as compared to conventional databases where the schema is defined before the database is populated even though the schema can be very large and constantly evolving. The World-Wide Web provides every internet citizen with access to an abundance of information, but it becomes increasingly difficult to identify the relevant pieces of information. Research in web mining tries to address this problem by applying techniques from data mining and machine learning to Web data and documents. The Web Mining is an application of Data Mining. Without the internet, life would have been almost impossible. The data available on the web is so voluminous and heterogeneous that it becomes an essential factor to mine this available data to make it presentable, useful, pertinent to a particular problem. Web mining deals with extracting these interesting patterns developing useful abstracts from diversified sources. The present paper deals with a preliminary discussion of WEB mining, few key computer science contributions in the field of web mining , the prominent successful applications and outlines some promising areas of future research.

Keywords *web mining, content mining, structure mining, usage mining*

I. INTRODUCTION

Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks [7]. Second was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process [8]. In this paper we follow the data-centric view, and refine the definition of Web mining as, Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of the structures (hyperlink) or the usage (Web log) data is used in the mining process (with or without other types of Web data).

Web mining is a new research issue under dispute which draws great interest from many communities. Currently, there is no agreement about Web mining yet. It needs more discussion among researchers in order to define what it is exactly. The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. It is our attempt in this paper to capture them in a systematic manner, and identify directions for future research.

In this paper we present a preliminary discussion about Web mining, key accomplishments, applications and future directions. The rest of the paper is organized as follows. Section II describes the various categories of web mining. Section III describes some of the computer science contributions in the field of web mining. Section IV discusses some prominent applications. Finally Section V outlines some promising areas of future research and Section VI concludes the paper.

II. TAXONOMY

Web mining is the application of data mining Techniques to extract knowledge from Web data, in which at least one of structure or usage (Web log) data is used in the mining process (with or without other types of Web). Researchers have identified three broad categories of Web mining:

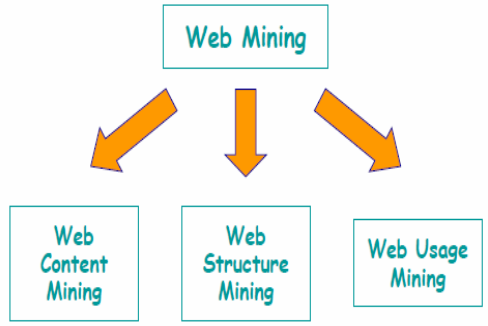

Fig 1. Web Mining Classification

A. Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the 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. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

B. 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.

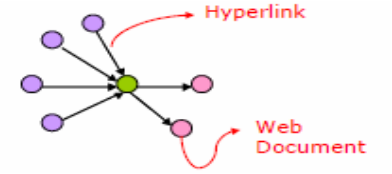

Fig 2. Web Graph Structure

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.

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*. There has been a significant body of work on hyperlink analysis, of which an up-to-date survey. There has been a significant body of work on hyperlink analysis, of which [9] provides an up-to-date survey.

Document Structure: In addition, 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 [10].

C. 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 [11]. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site Web Usage mining results from user interactions with a Web server, including Web logs, click streams and database transaction at a Web site or a group of related sites. Web usage mining introduces privacy concern and is currently the topic of extensive debate.

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 [BEA], Broad Vision [BV], Story Server [VIGN], 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: on the server side, the client side, and 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 [11].

III. ACHIEVEMENTS

This section briefly describes the key new concepts introduced by the Web mining research community.

A *Pre-processing – making Web data suitable for mining*

Pre-processing of Web data makes data suitable for mining was identified as one of the key issues for Web mining. A significant amount of work has been done in this area for Web usage data, including user identification [18], session creation [19, 20], robot detection and filtering [15], extracting usage path patterns [21], etc.

B *Detection and Filtering - Separating human and non human Web behavior*

Web robots are software programs that automatically traverse the hyperlink structure of the World Wide Web in order to locate and retrieve information. The importance of separating robot behavior from human behavior prior to extracting user behavior knowledge from usage data has been illustrated by [14]. First of all, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their Web sites. In addition, Web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to Web robots also make it more difficult to perform click-stream analysis effectively on the Web data. Conventional techniques for detecting Web robots are often based on identifying the IP address and the user agent of the Web clients. While these techniques are applicable to many well-known robots, they may not be sufficient to detect camouflaging and previously unknown robots. Experimental results have shown that highly accurate classification models can be built using this approach [15].

C *User profiles - Understanding how users behave*

The Web has taken user profiling to completely new levels. For example, in a 'brick and mortar' store, data collection happens only at the checkout counter, usually called the 'point-of-sale'. This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every single action taken by the user - which can provide much more

detailed insight into the decision making process. Adding such behavioral information to other kinds of information about users, e.g. demographic, psychographic, etc. allows a comprehensive user profile to be built, which can be used for many different applications [16].

D Maximum-Flow models - Web community identification

The idea of a maximal flow models has been used to identify communities, which can be described as a collection of Web pages such that each member node has more hyperlinks (in either direction) within the community than outside of the community. The s - t maximal flow problem can be described thus: Given a graph G = (V, E) whose edges are assigned positive flow capacities, and with a pair of distinguished nodes s and t, the problem is to find the maximum flow that can be routed from s to t. s is known as the source node and t as the sink node. Of course, the flow must strictly adhere to the constraints that arise due to the edge capacities. Ford and Fulkerson [17] proposed that the maximal flow is equivalent to a ”minimal cut” - that is the minimum number of edges that need to be cut from the graph to separate the source s from sink

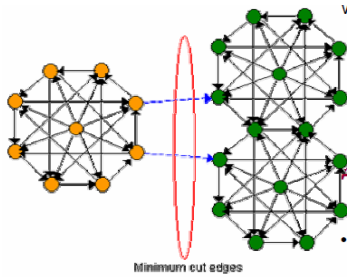

Fig 3. .Maximum Flow Model for web Communities

IV. PROMINENT APPLICATIONS

This section describes some of the most successful applications in this section. Clearly, realizing that these applications use Web mining is largely a retrospective exercise.

A Personalized Customer Experience in B2C E-commerce - Amazon.com

Early on in the life of Amazon.com, its visionary CEO Jeff Bezos observed, ’In a traditional (brick-and-mortar) store, the main effort is in getting a customer to the store. Once a customer is in the store they are likely to make a purchase - since the cost of going to another store is high – and thus the marketing budget (focused on getting the customer to the store) is in general much higher than the in-store customer experience budget (which keeps the customer in the store). In the case of an on-line store, getting in or out requires exactly one click, and thus the main focus must be on customer experience in the store.’ This fundamental observation has been the driving force behind Amazon’s comprehensive approach to personalized customer experience, based on the mantra ’a personalized store for every customer’ [22]. A host of Web mining techniques, e.g.associations between pages visited, click-path analysis, etc., are used to improve the customer’s experience during a ’store visit’. Knowledge gained from Web mining is the key intelligence behind Amazon’s features such as ’instant recommendations’, ’purchase circles’, ’wish-lists’, etc.

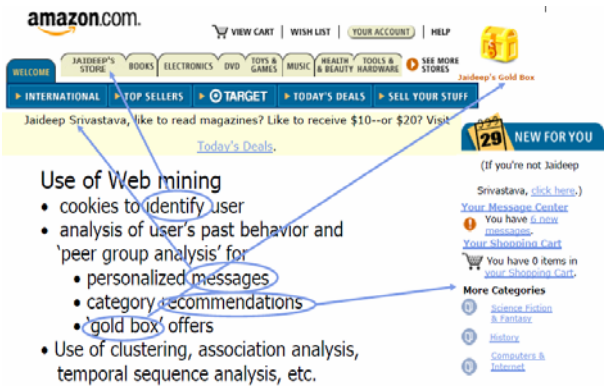

FIG 4. AMAZON.COM’S PERSONALIZED WEB PAGE

B Web Search—Google

Google has successfully used the data available from the Web content (the actual text and the hyper-text) and the Web graph to enhance its search capabilities and provide best results to the users. Google has expanded its search technology to provide site-specific search to enable users to search for information within a specific website. The Google Toolbar’ is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. The full version of the toolbar, if installed, also sends the click-stream information of the user to Google. The usage statistics thus obtained would be used by Google to enhance the quality of its results. Google also provides advanced search capabilities to search images and look for pages that have been updated within a specific date range. Built on top of Netscape’s Open Directory project, Google’s web directory provides a fast and easy way to search within a certain topic or related topics. The Advertising Programs introduced by Google targets users by providing advertisements that are relevant to search query. This does not bother users with irrelevant ads and has increased the clicks for the advertising companies by four or five times.

One of the latest services offered by Google is,’ Google News’. It integrates news from the online versions of all newspapers and organizes them categorically to make it easier for users to read ”the most relevant news”. It seeks to provide information that is the latest by constantly retrieving pages that are being updated on a regular basis.

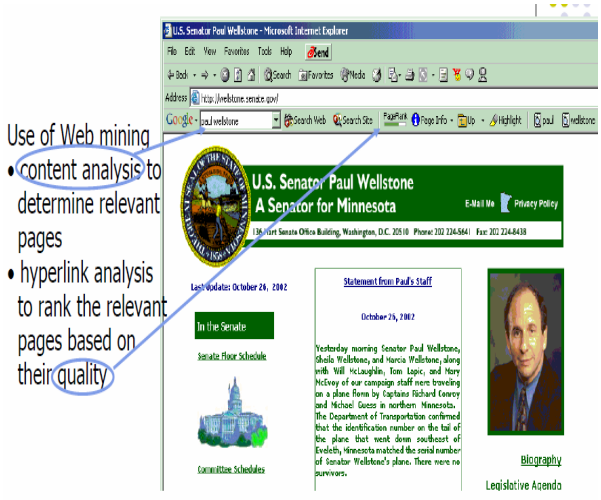

Fig 5. Web page returned by Google for query "paul Wellstone"

C *Understanding Web communities – AOL*

One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base. A large portion of this customer base participates in various 'AOL communities', which are collections of users with similar interests. In addition to providing a forum for each such community to interact amongst themselves, AOL provides useful information, etc. as well. Over time, these communities have grown to be well-visited 'waterholes' for AOL users with shared interests. Applying Web mining to the data collected from community interactions provides AOL with a very good understanding of its communities, which it has used for targeted marketing through ads and e-mail solicitations. Recently, it has started the concept of 'community sponsorship', whereby an organization like Nike may sponsor a community called 'Young Athletic Twenty Somethings'.

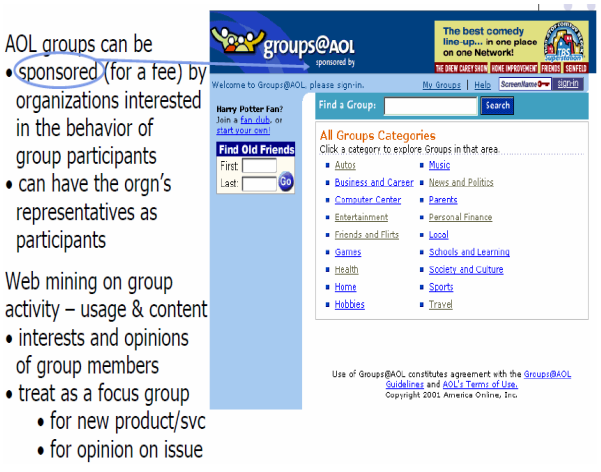

Fig 6. Understanding user communities—AOL

D *Understanding auction behavior – eBay*

As individuals in a society where we have many more things than we need, the allure of exchanging our 'useless stuff' for some cash - no matter how small - is quite powerful. This is evident from the success of flea markets, garage sales and estate sales. The genius of eBay's founders was to create an infrastructure that gave this urge a global reach, with the convenience of doing it from one's home PC. In addition, it popularized auctions as a product selling/buying mechanism, which provides the thrill of gambling without the trouble of having to go to Las Vegas. All of this has made eBay as one of the most successful businesses of the Internet era. Unfortunately, the anonymity of the Web has also created a significant problem for eBay auctions, as it is impossible to distinguish real bids from fake ones. eBay is now using Web mining techniques to analyze bidding behavior to determine if a bid is fraudulent [25].

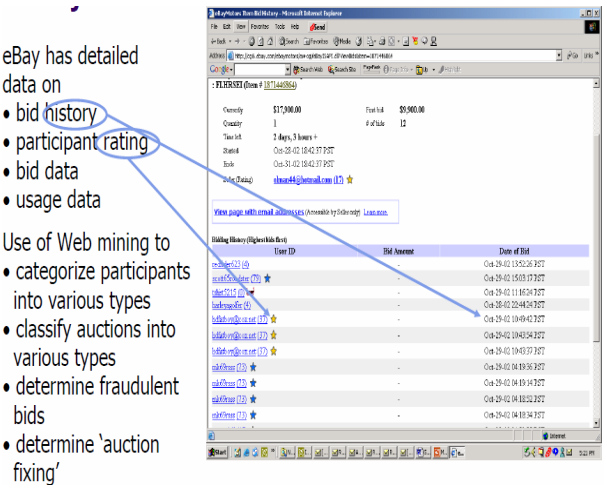

Fig 7. Understanding auction behavior-eBay

E *Personalized Portal for the Web – My Yahoo*

Yahoo was the first to introduce the concept of a 'personalized portal', i.e. a Web site designed to have the look-and-feel as well as content personalized to the needs of an individual end-user. This has been an extremely popular concept and has led to the creation of other personalized portals, e.g. Yodlee for private information. Mining My Yahoo usage logs provides Yahoo valuable insight into an individual's Web usage habits, enabling Yahoo to provide compelling personalized content, which in turn has led to the tremendous popularity of the Yahoo Web site.

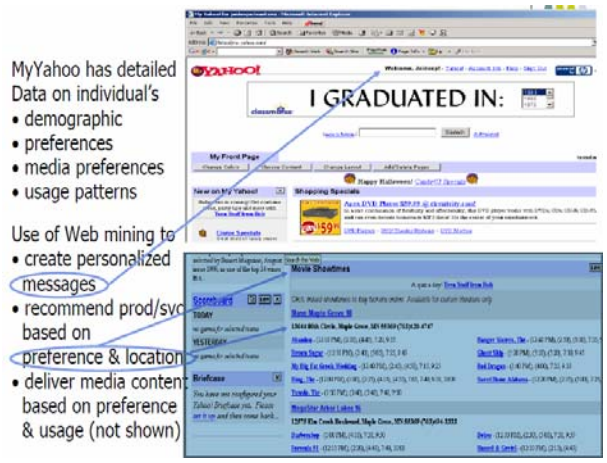

Fig 7. Personalized web portal – My Yahoo

V. FUTURE RESEARCH DIRECTIONS

As the Web and its usage grows, it will continue to generate ever more content, structure, and usage data, and the value of Web mining will keep increasing. Outlined here are some research directions that must be pursued to ensure that we continue to develop Web mining technologies that will enable this value to be realized.

A *Temporal evolution of the Web*

Society's interaction with the Web is changing the Web as well as the way the society interacts. While storing the history of all of this interaction in one place is clearly too staggering a task, at least the changes to the Web are being recorded by the pioneering Internet Archive project [IA]. Research needs to be carried out in extracting temporal models of how Web content, Web structures, Web communities, authorities, hubs, etc. are evolving. Large organizations generally archive (at least portions of) usage data from there Web sites. With these sources of data available, there is a large scope of research to develop techniques for analyzing of how the Web evolves over time.

The temporal behavior of the three kinds of Web data: Web Content, Web Structure and Web Usage. The methodology suggested for Hyperlink Analysis in [9] can be extended here and the research can be classified based on Knowledge Models, Metrics, Analysis Scope and Algorithms. For example, the analysis scope of the temporal behavior could be restricted to the behavior of a single document, multiple documents or the whole Web graph. The other factor that has to be studied is the effect of Web Content, Web Structure and Web Usage on each other over time.

B *Web services optimization*

As services over the Web continue to grow [28], there will be a need to make them robust, scalable, efficient, etc. Web mining can be applied to better understand the behavior of these services, and the knowledge extracted can be useful for various kinds of optimizations. The successful application of Web mining for predictive pre-fetching of pages by a browser has been demonstrated in [29]. Research is needed in developing Web mining techniques to improve various other aspects of Web services.

C Fraud and threat analysis

The anonymity provided by the Web has led to a significant increase in attempted frauds, from unauthorized use of individual credit cards to hacking into credit card database for blackmail purposes [30]. Yet another example is auction fraud, which has been increasing on popular sites like eBay. Since all these frauds are being perpetrated through the Internet, Web mining is the perfect analysis technique for detecting and preventing them. Research issues include developing techniques to recognize known frauds, and characterize and then recognize unknown or novel frauds, etc. The issues in cyber threat analysis and intrusion detection are quite similar in nature

D Web mining and privacy

While there are many benefits to be gained from Web mining, a clear drawback is the potential for severe violations of privacy. Public attitude towards privacy seems to be almost schizophrenic - i.e. people say one thing and do quite the opposite. For example, famous case like [23] seems to indicate that people value their privacy, while experience at major e-commerce portals shows that over 97can be provided based on it. Explicitly bringing attention information privacy policies had practically no effect. One explanation of this seemingly contradictory attitude towards privacy may be that we have a bi-modal view of privacy, namely that "I'd be willing to share information about myself as long as I get some (tangible or intangible) benefits from it, as long as there is an implicit guarantee that the information will not be abused". The research issue generated by this attitude is the need to develop approaches, methodologies and tools that can be used to verify and validate that a Web service is indeed using an end-user's information in a manner consistent with its stated policies.

VI. CONCLUSION

As the Web and its usage continues to grow, so does the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years has seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing. In this paper we have briefly described the key computer science achievements made in this field, the prominent successful applications, and outlined some promising areas of future research in the area of web mining.

REFERENCES

[1] T. Berners-Lee, R. Cailliau, A. Loutonen, H. Nielsen, and A. Secret. The World- Wide Web. *Communications of the ACM*, 37(8):76- 82, 1994.

[2] R.Albert, H. Jeong, and A.-L. Barabasi.Diameter of the World-wide web. *Nature, 401:13Q-131,* September 1999.

[3] I.Androutsopoulos, G. Paliouras, and E.Michelakis. *Learning to filter unsolicited commercial e-mail*. Technical Report NCSR Demokritos, March 2004.

[4] B. Berendt. Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*,37-59,2002.

[5] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In I. Horrocks and 1. Hendler, editors, *Proceedings of the1st International Semantic Web Conference (ISWC-02)*, pages 264-278. Springer-Verlag, 2002.

[6] J. Srivastava, B. Mobasher, Panel discussion on "Web Mining: Hype or Reality?" at *the 9th IEEE International Conference on Tools With Artificial Intelligence* (ICTAI '97), Newport Beach, CA, 1997.

[7] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", in *Proceedings of the 9th IEEE International Conference on Tools With Artificial Intelligence* (ICTAI '97), Newport Beach, CA, 1997.

[8] R. Kosala, H. Blockeel, "*Web Mining Research: A Survey*", in SIGKDD Explorations 2(1), ACM, July 2000.

[9] P. Desikan, J. Srivastava, V. Kumar, P.-N. Tan, "*Hyperlink Analysis –Techniques & Applications", Army High Performance Computing Center Technical Report*, 2002.

[10] Chuang-Hue Moh, Ee-Peng Lim, Wee KeongNg, "*DTD-Miner: A Tool for Mining DTD from XML Documents*", WECWIS 2000: 144-151.

[11] J. Srivastava, R. Cooley, M. Deshpande and P-N. Tan. "*Web Usage Mining: Discovery and Applications of usage patterns from Web Data*", SIGKDD Explorations, Vol1, Issue 2, 2000.

[12] L. Page, S. Brin, R. Motwaniand T. Winograd "*The Page Rank Citation Ranking: Bringing Order to the Web*" Stanford Digital Library Technologes, 1999-0120, January 1998.

[13] S. Brin, L. Page, "The anatomy of a large-scale hyper-textual Web search engine". *In the 7th International World Wide Web Conference*, Brisbane, Australia, 1998.

[14] R. Kohavi, "*Mining E-Commerce Data*: The Good, the Bad, the Ugly", Invited Industrial presentation at the ACM SIGKDD Conference, San Francisco, CA, 2001.

[15] Pang-Ning Tan, Vipin Kumar, *Discovery of Web Robot Sessions based on their Navigational Patterns*, DMKD, 6(1): 9-35 (2002).

[16] B. Masand, M. Spiliopoulou, J. Srivastava, O. Zaiane, ed. Proceedings of "*WebKDD2002 –Web Mining for Usage Patterns and User Profiles*", Edmonton, CA, 2002.

[17] L.R. Ford Jrand D.R. Fulkerson, "*Maximal Flow through a network*." Canadian J. Math.,8:399-404, 1956.

[18] W. Dong, M.S. Thesis, University of Minnesota, Computer Science & Engineering, 1999.

[19] R. Cooley, B. Mobasher, J. Srivastava, "*Data Preparation for Mining World Wide Web Browsing Patterns*", Knowledge and Information Systems, 1(1), 1999.

[20] B. Mobasher, M. Spiliopoulou, B. Berendt, *Proceedings of the SIAM Web Analytics Workshop*, Chicago, IL, 2001.

[21] M. Spiliopoulou, "Data Mining for the Web", *Proceedings of the Symposium on Principles of Knowledge Discovery in Databases* (PKDD), 1999.

[22] E. Morphy, "*Amazon Pushes 'Personalized Store for Every Customer'* ", E-Commerce Times, September 28, 2001 .

[23] C. Dembeck, P. A. Greenberg, "*Amazon: Caught Between a Rock and a Hard Place*", E-Commerce Times, Spetember8, 2000.

[24] DoubleClick'sLawsuit, http:/www.wired.com/news/business/0,1367,36434,00.html

[25] E. Colet, "*Using Data Mining to Detect Fraud in Auctions*", DSStar, 2002.

[26] Kok-LeongOng, Wee KeongNg, Ee-Peng Lim, "*Mining Relationship Graphs for Effective Business Objectives*", PAKDD 2002: 561-566.

[27] P. Underhill, "*Why We Buy: The Science of Shopping*", Touchstone Books, New York, 2000.

[28] R.H. Katz, "*Pervasive Computing: It's All About Network Services*", Keynote Address, Pervasive 2002, Zurich, Switzerland,2002.

[29] A. Pandey, J. Srivastava, S. Shekhar, "*A Web Intelligent Prefetcherfor Dynamic Pages Using Association Rules*" –A Summary of Results, SIAM Workshop on Web Mining, 2001.

[30] D. Scarponi, "*Blackmailer Reveals Stolen Internet Credit Card Data"*, Associated Press, January 10, 2000.

Ms. Subashini S is pursuing her Ph D in the area of Web usage mining for user profile analysis from the Department of Information Technology of Singhania University, Rajasthan. Her area of interests includes Data Mining, Web Usage Mining and Networking.

Dr. Mahesh T.R has done B.E.,in Computer Science & Engineering from Mysore University , MTech in Computer Science & Engineering from DR. MGR Educational & Research Institute, Chennai and PhD from University of Allahabad in the area of Data Mining for IDS. He has published several articles in national as well as International journals. His area of interests includes Data Mining, Web Usage Mining, Image processing, High performance systems and Networking.