Studies on Research and Development in Web Mining

Pawan Singh
Department of Computer Science,
IIMT Institute of Engineering and Technology,
Meerut, India

Amit Kumar
Department of Computer Science,
IIMT Institute of Engineering and Technology,
Meerut, India

Prashast
Department of Information Systems,
French School of Electronics and Computer Sciences, Paris, France.

ABSTRACT
World Wide Web has changing into one amongst the foremost comprehensive data resources. It most likely, if not perpetually, covers the data requirement for any user. However the net demonstrates several radical variations to traditional information containers such as databases in schema, volume, topic coherence etc. Web mining techniques could be applied to fully use web information in an effective and efficient manner, partially or completely. However, mining techniques are not the only tools to use web information efficiently but the mining techniques are the best solution. In this paper we study Web mining, Web mining categories and overview of various research issues and development efforts in web mining.

General Terms
Web Mining, Web Structure Mining, Web Content Mining, Web Usage Mining, Agent, Search Engine, Preprocessing, Pattern, Clustering.

Key words
World Wide Web, Web Mining, Data Mining.

1. INTRODUCTION
Web mining is a data mining technique to the World Wide Web. It has been the main focus of several research projects. The web mining field encompasses a large array of problems, primarily geared towards deriving actionable knowledge from the web [1]. "The application of Data Mining Techniques (shown in Fig-1) is to extracts fascinating (non trivial, implicit, previously, unknown and doubtless useful) information from massive web data, in which at least one of structure, usage or content data is used in the mining process". The World Wide Web consists of billions of web pages and huge amount of information available within web pages. To retrieve the required information the search engines process goes from crawling, indexing searching till filtering and sorting of information [2].The general architecture of search engine is shown in Fig-2.This paper is structured as follows: section 2 describes the categories of web mining. In section 3 different aspects of web content mining is detailed .Section 4 brief the details of web usage mining and section 5 shows the concepts of different web structure mining methods.

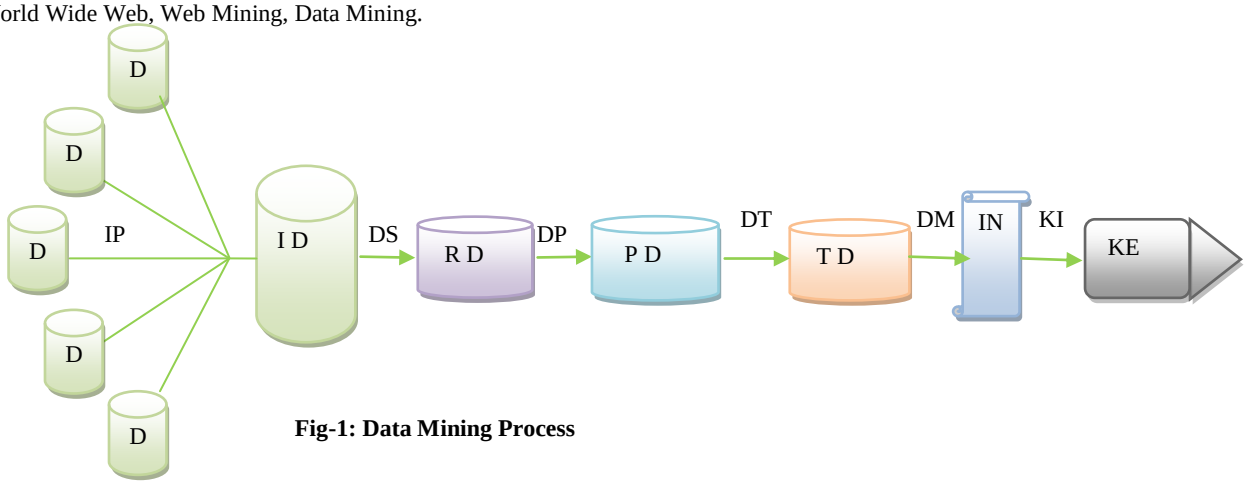

Fig-1: Data Mining Process

IP	Integration Process
DS	Data Selection
DP	Data Preprocessing
D	Data
ID	Integrated Data
RD	Require Data
PD	Preprocessed Data
TD	Transformed Data
IN	Information
KE	Knowledge
DT	Data Transformation
DM	Data Mining
KI	Knowledge Interpretation

2. CLASSIFICATION OF WEB MINING

Initially the researchers have their working in the individual fields of web mining, structure mining, content mining and usage mining. Some researchers combine content and structure mining, some of them combine structure and usage mining to collect the advantages of the mining techniques into one. Web mining can be categorized into three types as shown in Fig-3. Several research communities as database, IR and AI etc are converged into web mining research. This paper is an attempt to describe the research and development in web mining.

3. WEB CONTENT MINING

It means mining the content of web pages. It may be applied on web pages or search result pages. Content mining having two categories –Information Retrieval and Database. Unstructured text can be represented by collection of words whereas HTML structure inside the document can be used for the semi structured data. Also a web site database and web mining tries to extract the structure of website from this database.

3.1 Agent based Approach

Agent based web mining can be classified into the following three categories.

3.1.1 Intelligent Search Agent

Several intelligent web agents have been developed to search the relevant information by using domain characteristics and user profiles. Afterwards the organization and interpretation is done for discovered information. Some of the intelligent search agents are as follows Harvest [3] , FAQ Finder [4] , Information Manitold [5] , OCCAM [6] , Parasite[7], Shopbot[8] ,ILA [9].

3.1.2 Information Filtering /Categorization

Number of web agents use different information retrieval techniques and characteristics of hypertext web data. By using these techniques and characteristics automatic retrieval , filtering ,and categorization is done [10], [11], [12], [13], [14].Some agents uses semantic information embedded in link structure ,document content ,hierarchical clustering and user interaction from web.

3.1.3 Personalized Web Agents

This sort of web agents recognizes the user preferences and accordingly discovers the web information sources. Some uses the user profile and rate the web pages accordingly. Few example of such agents includes WebWatcher [15] ,Grouplens[16], PAINT [17] ,Firefly [18].

3.2 Database approach

Data Base approach has focused on the two categories: Multilevel Database and Web Query System.

3.2.1 Multilevel databases

In this approach at the bottom level database the semi structured data is stored in web repositories as hyper text .The generalizations are extracted at the higher level so as to organize in structured collections as relational and Object Oriented databases. Hen et al .[19] , Khosla, et.al.[20] , King and Novak [21] uses the multilevel database approach.

3.2.2 Web Query System

Natural language processing ,web structural information and standard database query language as SQL are used by different Web based query systems and languages for the queries used in WWW searches W3QL[22] ,Web log[23] , Lorel[24] extracts the the data using web query system.

4. WEB USAGE MINING

These mining techniques predict the user behavior of interacting WWW and collect the data from Web log records, these records can be analyze for web search, system improvement site modification, business intelligence and usage characterization.

4.1 Preprocessing

Due to the dearth of obtainable data Usage preprocessing is alleged to be most troublesome task in web usage mining. The mined data on the web which is generated owing to the user interaction to web. These data might range very widely however we can classify them as they reside in the web clients, proxy servers and servers [25]. Preprocessing technique uses directly the log data. The most important downside is identification among distinctive users, server sessions, episodes etc. Within the presence of caching and proxy servers [26],[25].

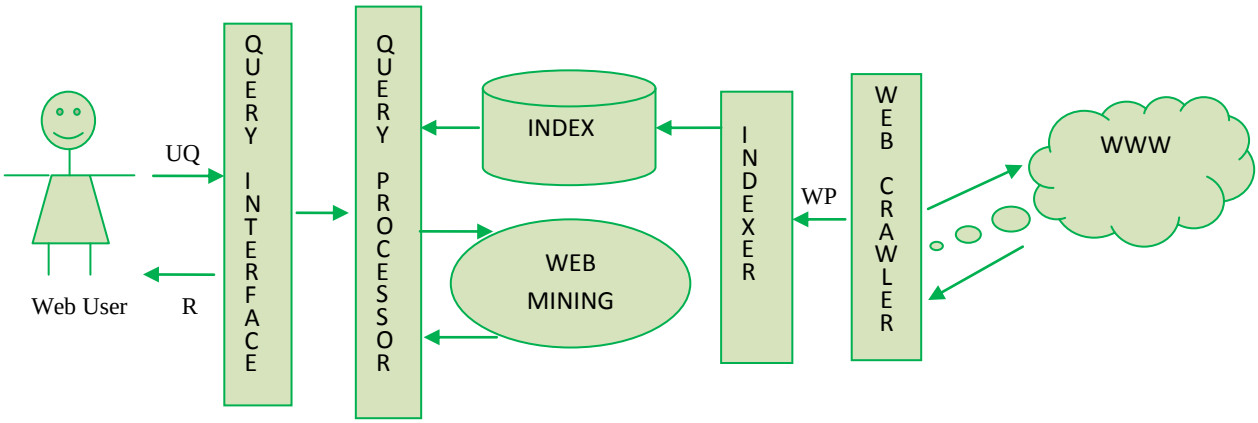

Fig-2: Architecture of Search Engine

UQ: User Query
R: Rank
WP: Web Pages

4.2 Pattern Discovery

Pattern Discovery stress upon the methods and algorithms developed from several fields like statistics, data mining, machine learning and pattern recognition [27], [28]. It describes the varieties of mining activities that have been applied to the web domain .In the web usage mining, a server session is an ordered sequence of web pages requested by some users. There are some difficulties in identifying the sessions uniquely because of which additional knowledge is needed such as imposing a default timeout period, etc.

4.2.1 Statistical Analysis

With the assistance of statistical techniques one can extract knowledge concerning web site visitors. Different variables such as page views, viewing time and length of a navigational paths etc. can be calculated from the session file .On the calculated values different statistical analysis can be performed .These records can be used to improve performance, security, modification task, marketing decision etc.

4.2.2 Association Rules

Pages referenced often together in a single server session can be related using association rules. The support values must exceed a threshold value .The single session pages may or may not be connected to one another via hyperlinks.

4.2.3 Clustering

Clustering is technique to group users exhibiting similar browsing patterns and pages having related content .Such information may help search engines and web assistances providers. With the help of past history of information need or user query, related hyperlinks can be suggested to the user by creating HTML pages.

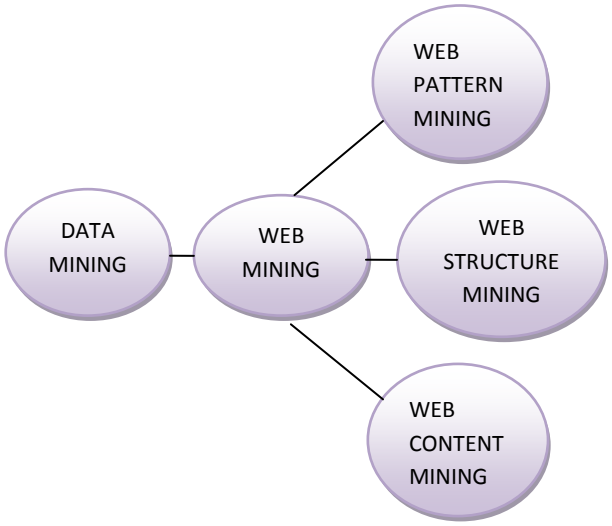

Fig-3: Categories of Web Mining

4.2.4 Classification

Classification is a method by which one can categorized the data item from the predefined categories .Each category has some features which describe the properties of a given class. All users belonging is been categorized to a particular class .Classification can be accomplished using supervised or un supervised learning methods.

4.3. Pattern Analysis

After discovering different pattern, pattern analysis is done to get filter out uninteresting patterns form the set found. Some of the analysis techniques are knowledge query mechanism (SQL), OLAP operations, Visualization techniques etc.

5. WEB STRUCTURE MINING

Web structure mining deals with the hyper link structure within the web itself. The WWW contains web pages and links .The web contains different objects with almost no unique structure. With the growing interest in the web mining research a new area has emerged called link mining. The link analysis algorithms can be classified into two categories .The first category is of algebraic methods and second of probabilistic methods.

5.1 HITS

Kleinberg [29] creates HITS as the first link analysis algorithm used for web mining .HITS was used for the first time within the Clever [30] search engine from IBM.As per Kleinberg "Hubs and authorities exhibit what may well be referred to as a mutually reinforcing relationship: a good hub could be a page that points to several good authorities: a good authority could be a page that is pointed to by many good hubs."HITS construct a Root Set of pages consisting of a brief list of web pages returned by search engine .Further more this Root Set is get added the pages pointed to by pages in the Root Set and pages that point to pages in the Root Set. This collective set is called Base sets HITS construct the adjacency matrix A where Aij =1 if there is a link from i to j and otherwise Aij =0. It assigns to each page i an authority weight a_i and hub weight h_i calculate as $a_i^{(t+1)} = \sum_{j \to i} h_j^t$, $h_j^t = \sum_{i \to k} a_k^t$ using iteration method to get converged value a_i^* , h_i^*. a_i^* is taken to be a value of page i & hub value h_i^* .

5.2 Page rank

L. Page & S. Brin proposed Page Rank algorithm [31], it was implemented in Google Search engine. In page ranking probability to choose a page randomly is ε and any of the forward links on the current page is 1- ε. The transition probability matrix of this Markov chain is given by (εU + (1-ε) Arow) where Arow is constructed by renormalizing each row of the adjacency matrix A to sum to 1^4 and U is the transition matrix of uniform transition probabilities [32]. In page rank algorithm a single value is assigned to each page, unlike HITS where hub values and authority values exist.

5.3 SALSA

Salsa is proposed by Lempel and Morgan [33].Salsa performs two random walks on web pages; a random walk by following a backward link and then a forward link alternatively, and another by following a forward – link and then a backward link alternatively. Stationary distribution of the former random walk defines the authority weight where as the stationary distribution of the later random walk defines the hubs weight.

Table-1: Overview of Web Mining [34]

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR View	DB View		
View of Data	- Unstructured - Semi structured	- Links structure - Interactivity	- Semi structured	- Web site as DB
Main Data	- Text documents - Hypertext documents	- Links structure	- Server logs	- Hypertext documents - Browser logs
Representation	- Bag of words, n-grams - Terms, phrases - Concepts or ontology - Relational	- Edge-labeled graph (OEM) - Relational	- Graph	- Relational table - Graph
Method	- TFIDF and variants - Machine learning - Statistical (including NLP)	- Proprietary algorithms - ILP - (Modified) association rules	- Proprietary algorithms	- Machine Learning - Statistical - (Modified) association Rules
Application Categories	- Categorization - Clustering schema discovery -Finding extraction rules - Finding patterns in text - User modeling	- Finding frequent substructures - Web site	- Categorization - Clustering	- Site construction, adaptation, and management - Marketing - User modeling

6. CONCLUSION

In this paper we survey the research area of web mining, the types of web mining (Table-1 gives an overview of the web mining categories) and working of different algorithms of different mining techniques. It seems that still there is a scope of improvement and enhancement in mining techniques to be discovered. There are still several interesting research issues not fully covered in this short abstract .We hope this paper could be a useful starting point for identifying opportunities for further research.

7. REFERENCES

[1] Web Mining: Research and Practices, PRANAM KOLARI and ANUPAM JOSHI, IEEE, 2004.

[2] A Novel page ranking method based on Link –Visits of Web Pages , A.k Sharma ,Neelam Duhan,,Gyanendra kumar. International journal of trends in engineering and technology vol 4, no 1 nov 2010.

[3] CM Brown ,BB Danzing ,D Hordy ,U.Manber and MF Schwartz ;The harvest information discovery and access system. In Proc. 2nd International WWW Conference 1994 .

[4] K.Hammond ,R.Burke ,C.Martin and S.Lytinen .FAQ Finder :A case based approach to knowledge navigation .In Working Notes of the AAAI Press ,1995 .

[5] T. KIRK ,A K Levy ,Y Sagiv and D.Srivastava .The information manifold. In Working Notes of the AAAI Spring Symposium: Information gathering from heterogeneous ,Distributed Environments ,AAAI press ,1995.

[6] C.Kwok and D .Weld .Planning to gather information. In Proc .11th National Conference on AI,1996.

[7] E.Spertus. Parasite: mining structural information on the web .In Proc. Of 6th International World Wide Web Conference,1997.

[8] R.B.Doorenbos ,O.Etzioni,and D.S. Weld .A scalable comparison shopping agent for the world wide web. Technical Report 96-01-03, University of Washington,Dept. Of Computer Science and Engineering ,1996.

[9] M.Perkowitz and O.Etzioni .Category translation: learning to understanding information on the internet. In Proc. 15th International Joint Conference on AI, pages 930-936, Montral, Canada, 1995.

[10] A.Z.Broder ,S.C. Glassman ,M.S. Manasse, and G Zweig. Syntactic clustering of the web .In Proc. Of 6th International World Wide Web Conference, 1997.

[11] C.Chang and C.Hsu. Customizable multi -engine search tool with clustering. In Proc. Of 6th International World Wide Web Conference, 1997.

[12] Y.S Maarek and I.Z. Ben Shaul .Automatically organizing bookmarks per content . In Proc. Of 5th International World Wide Web Conference,1996.

[13] M.R.Wulfekuhler and W.F.Punch. Finding salient features for personal web pages categorization. In Proc. Of 6th International World Wide Web Conference,1997.

[14] R.Weiss, B.Velez ,M.A. Sheldon ,C .Namprempre ,P.Szilagyi ,A. Duda, and D.K.Gifford .Hypursuit : a hierarchical network search engine that exploits content –link hypertext clustering .In Hypertext '96: The Seventh ACM Conference on Hypertext ,1996.

[15] R. Armstrong ,D. Freitag, T.Joachims , and T. Mitchell. Webwatcher :A learning apprentice for the world wide web .I Proc. AAAI Spring Symposium on Information Gathering from Heterogeneous ,Distributed Environments ,1995.

[16] W.B. Frakes and R. Baeza- Yates .Information Retrieval Data Structures and Algorithms .Prentice Hall, Englewood Cliffs ,NJ , 1992.

[17] K.A. Oostendorp ,W.F. Punch ,and R.W. Wiggins .A tool for individualizing the web. In Proc. Of 2nd International World Wide Web Conference,1994.

[18] U .Shardanand and P.Maes .Social information filtering: Algorithm for automating "word of mouth". In Proc. Of 1995 Conference on Human Factors in Computing Systems (CHI- 95), pages 210- 217, 1995.

[19] O.R. Zaiane and J. Han. Resource and knowledge discovery in global information systems: A preliminary design and experiment. In Proc. Of the First Int'l Conference on knowledge Discovery and Data Mining, pages 331-336 ,Montreal ,Quebec ,1995.

[20] I. Khosla ,B. Kuhn, and N .Soparkar .Database search using information mining . In Proc. Of 1996 ACMSIGMOD Int. Conf. On Management of Data, 1996.

[21] R. King and M.Novak. Supporting information infrastructure for distributed, heterogeneous knowledge discovery. In Proc. SIGMOD 96 Workshop on Research Issues on Data Mining and Knowledge Discovery ,Montreal , Canada, 1996.

[22] D. Konopnicki and O. Shumeli. W3qs: A query system for the World Wide Web. In Proc. Of 21th VLDB Conference, pages 54-65 ,Ziruch ,1995.

[23] L. Lakshmanan ,F. Sadri , and I.N. Subramanian. A declarative language for querying and restructuring the web. In Proc. Of 6th International Workshop on Research Issues in Data Engineering: Interoperability on Nontraditional Database Systems (RIDE-NDS'96),1996.

[24] D. Quass , A. Rajaraman ,Y. Sagiv ,J. Ullman, and J. Widom. Querying semistructured heterogeneous information. In International Conference on Deductive and Object Oriented Databases, 1995.

[25] J. Srivastav ,R. Cooley ,M. Deshpande, and P. N. Tan. Web usage mining: Discovery and applications of usage patterns from web data .SIGKDD Explorations, 1(2) , 2000.

[26] B. Masand and M. Spiliopoulou. Webkdd-99: Workshop on web usage analysis and user profiling. ACMSIGKDD Explorations, 1(2) , 2000.

[27] U. Fayyad , G. Piatetsky- Shapiro, and P. Smyth. From data minig to knowledge discovery: An overview .In Proc. ACM KDD, 1994.

[28] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective .IEEE Transactions on Knowledge and Data Engineering, 8(6): 866-833 ,1996.

[29] J. Klienberg, "Authoritive zsources in a hyperlinked environment" , Journal of the ACM,46 1999.

[30] Av. Padre Tomás, S.J., Taipa, Macao S.A.R.," Web Structure Mining: An Introduction", Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China

[31] S. Brin and L. Page , "The natomy of a large -scale hypertext (Web) search engine" , Proc. 7th International World Wide Web Conference,1998.

[32] SPIRE 2004; H. C. Lee, "Metasearch via the co-citation graph, in Proc. IC 2003; H. C. Lee and A. Borodin,"Perturbation of the hyperlinked-environment"

[33] R. Lempel and S. Moran, The stochastic approach for link –structure analysis (SALSA) and the TKC effect. Proc. 9th International World Wide Web Conference May 2000.

[34] R.Khosala,H.Blockeel,SIGKDD Exploration , ACM SIGKDD,July 2000.