Web Mining for Multimedia Data-A Soft Computing Framework

*Reshma P.K., Lajish V.L.

Abstract— The World Wide Web (WWW) is a system of interlinked hypertext documents that are accessed through the Internet. Web browsers provide easy access to numerous sources of text and multimedia very easily. Today billions of pages are indexed by search engines and finding the desired information is not so simple. This has prompted the need for developing automatic mining techniques on the WWW, by the new term "Web mining". In web mining data can be collected at the server side, client side, proxy servers or obtained from the organization's database. Depending on the source of data, the type may vary. Now most of the mining techniques used are text centric and the algorithms are oriented towards text mining framework. But in the present scenario, web is gaining a multimedia character with pages containing images, audios, videos etc. Web mining algorithms for multimedia data are developed recently, but it has a long way to go. This paper gives an overview of the existing techniques used for web mining of multimedia data and a better solution using soft computing for web mining of multimedia data.

Index Terms—World Wide Web, Web mining, Content mining, Structure mining, Usage mining, Softcomputing, Multimedia, Soft web mining, Document Object Model (DOM), Fuzzy Logic, Genetic Algorithm, Artificial Neural Network, Rough set theory

1 INTRODUCTION

THE web is huge collection of uncontrolled heterogeneous documents which makes the web a fertile area of data mining research with the huge amount of information available online. Web mining can be defined as the discovery and analysis of useful information from the WWW. In web mining data can be collected at the server side, client side, proxy servers or obtained from the organization's database. Depending on the source of data, the mining type may vary.

The problem of developing automated tools in order to find, extract, filter and evaluate the users desired information from unlabeled, distributed and heterogeneous web data required human intervention which can otherwise be incorporated using soft computing techniques.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data are raw facts such as numbers, text or images that can be processed by a computer. Today, organizations are accumulating infinite and growing amounts of data in different formats and different databases. This includes:

- Operational or Transactional data such as, sales, cost, inventory, payroll, and accounting.
- Nonoperational data, such as industry sales, forecast data, and macro economic data
- Meta data - Data about the data.

- *Reshma P.K ,Assistant Professor,Department of Computer Science,Mahatma Gandhi College, Iritty,Kannur, Kerala, India. E-mail: pkreshma@gmail.com*
- *Lajish V.L, Assistant Professor, Department of Computer Science, University Of Calicut, Kerala-673635, India. E-mail: lajish@uoc.ac.in*

Data mining mainly aims to determine the relationship between internal factors to the external factors. For example, in the case of commercial organizations, these are used to determine the relationships among internal factors such as price, product, or staff skills, and external factors such as economic indicators, competition etc. And, it is also used to determine the impact on sales, customer satisfaction and profits.

2 WEB MINING

The process of applying the techniques for extracting the data from the World Wide Web is known as Web mining. Data is the collection of facts a web page is designed to contain. It may include texts, images, audios, videos, or some structured records such as lists and tables. There are two different approaches in defining web mining. First one is a process-centric view, which defined web mining as a sequence of tasks (Etzioni 1996) [1]. The second and the most accepted definition is a data-centric view, which defined web mining in terms of the types of web data that was being used in the mining process (Cooley, Srivastava, and Mobasher 1997)[2]. Web mining can be categorized into three types. They are,

1. *Web Content Mining*: Retrieval of useful information from the web is referred to as Web Content Mining. Information can be anything like text, image, audio and video.
2. *Web Strcture Mining*: It is the method of discovering the information from the web. Based on the structure information it can be sub categorized as hyperlinks and document structure. Hyperlinks are used to connect a Web page with other Webpages or other portions of the same 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*. Document structure refers to the tree structured arrangement based on HTML and XML tags within the page. This type of mining is focused on automatically extracting Document Object Model (DOM) structures out of documents

3. *Web Usage Mining*: Web contains enormous collection of different patterns. The application of Data mining techniques in order to obtain the useful patterns from the Web is known as Web Usage Mining. Usage data collects the identity or origin of web users with their browsing behavior at a web site. Web usage mining is further divided based on the kind of usage data considered. In *Web Server Data,* user logs are collected by the web server and typically include IP address, page reference and access time. Commercial application servers used for e-Commerce applications can be tracked to get various kinds of business events and log them in application server logs. This comes under the category of *Application Server Data.* To generate histories upon the above mentioned logging and applications require a combination of these techniques. This is termed as *Application Level Data.*

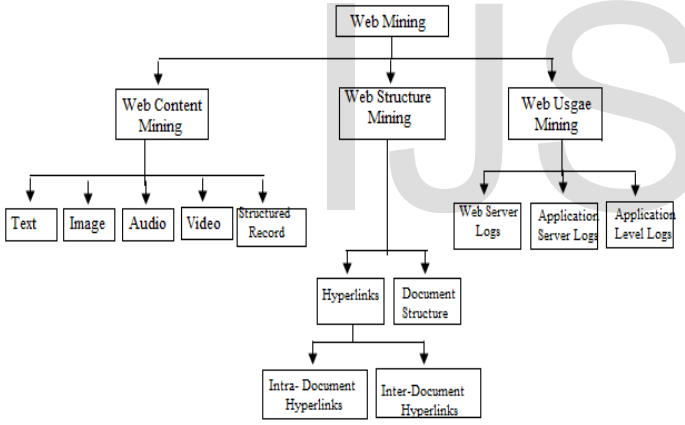

Figure 1: Web mining classification

2.1 Applications of Web Mining

Web applications being developed rapidly in the industry are based on the use of web mining concepts, even though the organizations that developed these applications would not consider it as such. Some of the notable applications are

1. ***B2C e-Commerce*:** In a traditional 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 because the cost of going to another store is high and thus the marketing budget much higher than the instore customer experience. But in the case of an on-line store, getting in or out requires just one click, and thus the main focus must be on the in-store customer experience. Hence an appraoch used is a personalized store for every customer. Web mining technique to find out the associations between pages visited and click-path analysis are used to improve the customer's experience during a store visit.
2. ***Web Search:*** Google is the most popular and promising search engine that provides its users access to information from over billions of web pages indexed on its server. The quality and speed of the search facility makes it the most successful search engine. Page rank which measures the importance of a page suing index, is the underlying technology in all Google search products. The Google toolbar is another service that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. And it also provides some advanced search features to make the search easy and effective.
3. ***Personalised Portal for the Web***: A web site can be designed to have the look-and-feel and content personalized to the needs of an individual end user. This has been first introduced by Yahoo and has lead to the development of other personalized portals.
4. ***CiteSeer****-Digital Library and Autonomous Citation Indexing*: CiteSeer was a public search engine and digital library for scientific and academic papers, mainly for Computer and Information Science. The important features provided by this were
   - Autonomous Citation Indexing automatically created
   - Citation statistics and related documents were computed for all articles cited in the database.
   - Reference linking allowing browsing of the database using citation links.
   - Citation context showed the context of citations to a given paper, allowing a researcher to quickly and easily see what other researchers have to say about an article of interest.
   - Related documents were shown using citation and word based measures and an active and continuously updated bibliography is shown for each document.

3 MULTIMEDIA DATA

Multimedia Data referes to the elements used to build a generalized multimedia environments, platforms, or integrating tools. The basic types can be categorized as:

- **Text**: The text can be stored in a variety of forms. In addition to ASCII based files, text is typically stored in processor files, spreadsheets, databases and annotations on more general multimedia objects. With the availability and abundance of GUIs that allow special effects such as color and styles, the task of storing texts becomes more complex.

- **Images**: There is great variation in the quality and size of storage for still images. Digitalized images are sequence of pixels that represents a region in the user's graphical display. The factors affecting the space overhead for still images are resolution, size, complexity, and compression scheme used to store image. The popular image formats are jpg, png, bmp and gif.
- **Audio**: Audio is an increasingly popular datatype being integrated in most of applications. Its quite space intensive. One minute of sound can take up to 2-3 MBs of space. Several techniques are used to compress it in suitable format.
- **Video**: One on the most space consuming multimedia data type is digitalized video. The digitalized videos are stored as sequence of frames. Depending upon its resolution and size a single frame can consume upto 1 MB. And to have realistic video playback, the transmission, compression, and decompression of digitalized require continuous transfer rate.
- **Graphic Objects:** Graphic Objects consists of special data structures used to define 2D and 3D shapes through which we can define multimedia objects. These include various formats used by image, video editing applicationssuch as CAD / CAM objects.

3.1 Multimedia Mining

Mining a Multimedia data is used to retrieve different types of data. The process of applying multimedia mining consists of differnt steps. Data collection is the first point of a learning system, as the quality of raw data is the factor which determines the overall achievable performance. The main goal of data pre-processing is to discover the important patterns from the raw data, which includes the concepts of data cleaning, normalization, transformation, feature selection etc.

Learning can be simple, if informative features can be identified at pre-processing stage. Detailed procedure depends highly on the nature of raw data and problem's domain. The product of data pre-processing is the training set. Given a training set, a learning model has to be chosen to learn from it and make multimedia mining model more iterative.

The process of Multimedia mining is shown in figure 2.

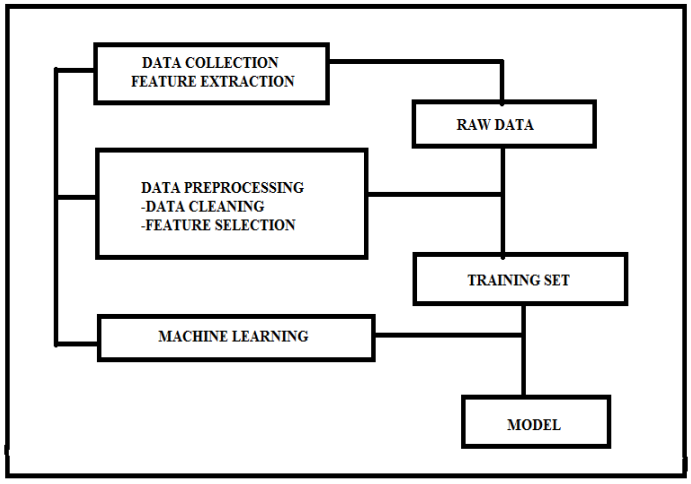

Figure 2: Multimedia Mining Process

3.1 Web Mining of Multimedia Data

A Multimedia system includes a multimedia database management system (MMDBMS),that can manage and provide support for storing, manipulating and retrieving multimedia data from a multimedia database, which is a large collection of multimedia objects, such as image, video, audio and hypertext data [3]. Multimedia gives a lot of information on each entity but the information may not be the same for each entity. One of the main characteristics of Multimedia Mining is the sequence or time element. Multimedia usually captures an entity changing over time. Video, audio and text are ordered, and are meaningless without sequence. Understanding and representing changes in the mining process is necessary to mine multimedia data [4].

Multimedia mining includes the construction of a multimedia data cube which provides multiple dimensional analyses of multimedia data, primarily based on visual content, and the mining of multiple kinds of knowledge, including summarization, comparison, classification, association and clustering [5]. The multimedia files from a database are first pre-processed to improve their quality and followed by feature extraction. With the help of generated features, information models can be devised using data mining techniques such as pattern discovery, rule extraction and knowledge acquisition to discover significant patterns from multimedia database [6]. A model of the Multimedia Database Management is shown in the figure 3.

IJSER

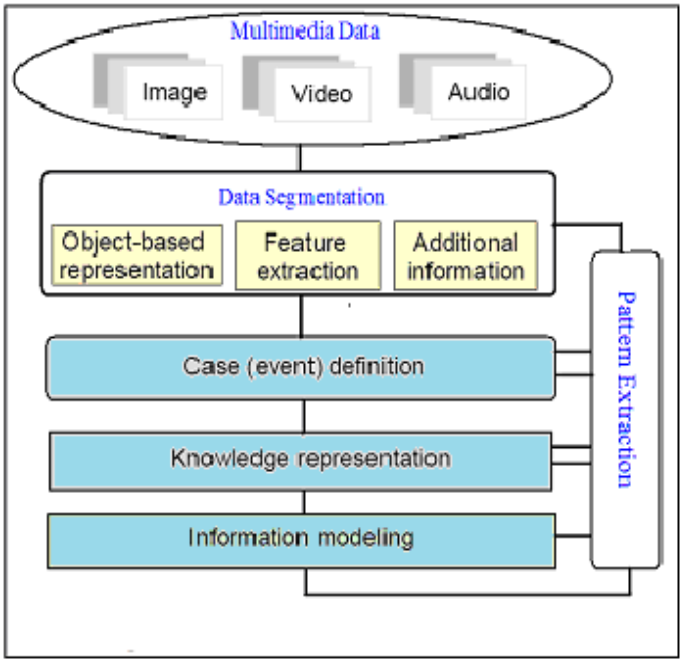

Figure 3: Multimdeia Database Management

It is difficult to fit the multimedia in any typical data mining models. Hence the heterogeneous databases could be first integrated and then mined or can apply mining tools on the individual databases and then combine the results of the various data miners with the Multimedia Distributed Processor (MDP). The process of mining and then integrating through MDPs is depicted in figure 4. Mining individual data types such as text, images, video, and audio, itself is a complex procedure, mining combinations of data types is still a challenge [4].

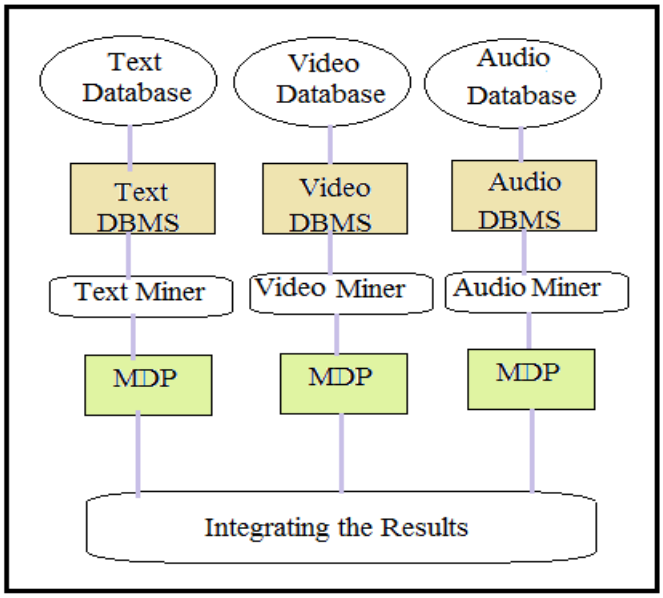

Figure 4: Mining with the help of MDPs

3.2 Issuses of the Current systems

The web creates new challenges to different component tasks of web mining like Information Retrieval, Information Extraction, generalization analysis and validation because the amount of information on the web is increasing and changing rapidly without any control. As aresult, the existing systems find difficulty in handling the newly emerged problems during these phases.

The aim of an IR system is to estimate the relevance of documents to users' information needs, expressed with the help of queries. It becomes complex due to the inherent subjectivity, imprecision, and uncertainty related to user queries. Query processing in search engines, which are an important part of IR systems, is simple keyword matching. This will not consider the context and relevance of queries with respect to documents, while these are important for efficient machine learning. Also the current search engines have no deductive capability. Determining the relevance of a retrieved document with respect to a query is a gradual property and not a crisp one [7]. It is necessary that IR systems modify the retrieved document set as per users' history or nature. Though some of the existing systems do so for a few limited problem domains, no definite general methodology is available. IR systems find difficulty in dealing with the problem of dynamism, scaling, and heterogeneity of web documents. The heterogeneity nature of web documents requires a separate mining method for each type of data.

The techniques used in an IE system are customized for a particular site and is not universally applicable. Also, source documents are designed for people and few sites provide machine readable specifications of their formatting conventions. The conventions used in one site may not be relevant elsewhere.

In the generalization phase, the tasks of clustering and association rule mining find some difficulty. Existing clustring methods used to organize the retrieved data are not so efficient because the data available on the web is distributed, heterogeneous, imprecise, very high dimensional and overlapping. In association rule mining, the current techniques are not able to appropriately mine for linguistic association rules which are more human understandable. Also the use of sharp boundary intervals is not intuitive with respect to human perception. For example, an interval method may classify a person with age less 35 as young and as old if it is not. But this may not always always correspond to the human perception of young and old, which considers the boundaries of these imprecise concepts, not hard/crisp.

The biggest problem faced in analysis is from the point of view of knowledge discovery and modeling [8]. Discovering knowledge out of the information available on the web has always been a challenge to the analysts, as the output of knowledge mining algorithms is often not suitable for direct

human interpretation because the patterns discovered are mainly in mathematical form.

4 SOFT COMPUTING

Soft coputing is a collection of tools including fuzzy sets, artificial neural network, genetic algorithms and rough set theory. Fuzzy sets provide a natural framework for the process in dealing with uncertainty. Neural networks are used for modeling complex functions, and provide learning and generalization capabilities. Genetic Algorithms are used for searching and optimization and Rough Sets are used in granular computation and knowledge discovery.

Soft computing is a collection of methodologies that provides flexible information processing capability for handling real life ambiguous situations. It is used to exploit the tolerance for imprecision, uncertainty, approximate reasoning and partial truth to achieve tractability, robustness, low-cost solutions and close resemblance to human-like decision making [9].

4.1 Soft Computing in Web Mining

Fuzzy Logic (FL) is used for handling issues related to incomplete or imprecise data or query, approximate solution, human interaction with linguistic information, understandability of patterns and deduction, and mixed media information. Artificial Neural Networks (ANNs) are used for modeling highly nonlinear decision boundaries, generalization and learning (adaptivity), self organization, rule generation, and pattern discovery. Genetic Algorithms (GAs) are seen to be useful for prediction and description, efficient search, and adaptive and evolutionary optimization of complex objective functions in dynamic environments. Rough Set (RS) theory is used to obtain approximate description of objects in a granular universe in terms of its core attributes. It provides fast algorithms for extraction of domain knowledge in the form of logical rules. Recently, various combinations of these tools have been made in soft computing paradigm, among which neuro-fuzzy integration is the most visible one [10].

- *FL for Web mining*:

  The application of FL contributes greatly to the IR and generalization tasks of Web mining. Yager describes in [11] a framework for formulating linguistic and hierarchical queries. It describes an IR language which enables users to specify the interrelationships between desired attributes of documents sought using linguistic quantifiers. Examples of linguistic quantifiers include most, at least, about half etc. Fuzzy Boolean IR models are more flexible in representing both document contents and information needs.

  The key requirements of web document clustering are measure of relevance, browsable summaries, ability to handle overlapping data, snippet tolerance, speed and incremental characteristics [12].

  A fuzzy clustering technique for web log data mining is described in [13]. Here, an algorithm called competitive agglomeration of relational data (CARD) for clustering user sessions is described, which considers the structure of the site and the URLs for computing the similarity between two user sessions.

  Some of the commonly used areas of FL are search engines, similarity measures, ontologies, summarization, e-Commerce, customization and profiling etc.

- *ANN for Web mining:*

  ANN can be defined as a massively parallel interconnected network of simple, adaptive processing elements which is intended to interact with the objects of the real world in the same way as biological systems do. ANNs are designated by the network topology, connection strength between pairs of neurons (called weights), node characteristics, and the status updating rules. They have been applied to the tasks like IR, IE, and clustering (self organization) of web mining, and for personalization. ANNs provide a convenient method of knowledge representation for IR applications. Also their learning ability helps to achieve the goal of implementing adaptive systems. Most Informatin Extraction (IE) systems that use learning fall into two groups, the one that uses relational learning [14], [15] to learn extracted patterns, and the other group learns parameters of Hidden Markov Models (HMMs) and uses them to extract information[16].

  InWEBSOM [17], the self-organizing map (SOM) algorithm is used to automatically organize very large and high-dimensional collections of text documents onto two-dimensional map displays. The map forms a document landscape where similar documents appear close to each other at different points of the regular map grid. The landscape can be labeled with automatically identified descriptive words that convey properties of each area and also act as landmarks during exploration.

  Personalization means that the content and search results are personalized as per user's interests and habits. ANNs may be used for learning user profiles with training data collected from users or systems as in [18]. Since user profiles are highly nonlinear functions, ANNs seem to be an effective tool to learn them.

- *GA for Web mining*

  GA is a biologically inspired technology which consists of randomized search and optimization techniques guided by the principles of evolution and natural genetics. They are efficient, adaptive and robust search processes, producing near optimal solutions, and have a large amount of implicit parallelism. GAs are executed iteratively on a set of coded solutions (genes), called population, with three basic operators: selection/reproduction,

crossover, and mutation. GAs are mainly used in search, optimization, and description. Web document retrieval by genetic learning of importance factors of HTML tags has been described in [19]. In [20], Boughanem *et al.* developed a query reformulation technique using GAs, in which a GA generates several queries that explore different areas of the document space and determines the optimal one.

- *RS for Web mining*:

To handle heterogeneous data and in clustering and association, RS can effectively be used. RS theory can also be used for the purpose of approximate information retrieval, where the set of relevant documents may be rough and represented by its upper and lower approximations. Uses of variable precision RSs [21] and tolerance relations are important in this context.

5 Conclusion

Web mining is growing rapidly since its development and new methodologies are being developed both using classical and soft computing approaches concurrently. Now most of the mining techniques used are text centric and the algorithms are oriented towards text mining framework. But in the present scenario, web is gaining a multimedia character with pages containing images, videos etc. This paper considers the immense potential of application of soft computing to web mining. Here, we have summarized the different types of web mining and its basic components, along with the limitations of the existing web mining methods/tools. The relevance of soft computing, including integration of its constituting tools, are briefly mentioned. An overview of promising research results achieved in Web mining for multimedia data based on the soft computing framework is also summarized

Acknowledgment

We would like to thank staff members and research scholars of the Department of Computer Science, University of Calicut and the Department of Computer Science, Mahatma Gandhi College, Iritty for their support and enlightning discussions.

References

[1] O. Etzioni, "The World Wide Web: quagmire or gold mine?" Communications of ACM, Vol. 39, p65-68,1998

[2] R.Cooley, B. Mobasher, J.Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web". IEEE Transactions, 1997

[3] Yoshitaka A. and Ichikawa T., "A Survey on Content-based Retrieval for Multimedia Databases".IEEE Transaction on Knowledge & Data Engineering, Vol 11, pp. 81-93,1999

[4] Bhavani Thuraisingham, "Managing and Mining Multimedia Databases" at International Journal on Artificial Intelligence Tools Vol. 13, No. 3 739-759,2004

[5] Osmar R. Za _ane Jiawei Han Ze-Nian Li Sonny H. Chee Jenny Y. Chiang, "MultiMediaMiner: A System Prototype for MultiMedia Data Mining," Intelligent Database Systems Research Laboratory and Vision and Media Laboratory report, 2009.

[6] Dianhui Wang, Yong-Soo Kim, Seok Cheon Park, Chul Soo Lee and Yoon Kyung Han, "Learning Based Neural Similarity Metrics for Multimedia Data Mining" Soft Computing, Volume 11, Number 4, , pp. 335- 340,2007

[7] C. V. Negotia, "On the notion of relevance in information retrieval,"*Kybernetes*, vol. 2, no. 3, pp. 161–165, 1973.

[8] S. Brin and L. Page, "The anatomy of a large scale hypertextual web search engine," in Proc. 8th Int. WWW Conf., Brisbane, Australia, pp. 107–117, 1998

[9] L. A. Zadeh, "Fuzzy logic, neural networks, and soft computing,"*Commun. AGM*, vol. 37, pp. 77–84, 1994.

[10] S. K. Pal and S. Mitra, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*. New York: Wiley, 1999.

[11] R. Yager, "A framework for linguistic and hierarchical queries for Document retrieval," in Soft Computing in Information Retrieval: Tech niques and Applications, F. Crestani and G. Pasi, Eds, Heidel berg: PhysicaVerlag,, vol. 50, pp. 3–20,2000

[12] O. Etzioni and O. Zamir, "Web document clustering: A feasibility demonstration," in *Proc. 21st Annu. Int. ACM SIGIR Conf.*, pp. 46–54,1998

[13] A. Joshi and R. Krishnapuram, "Robust fuzzy clustering methods to support web mining," in *Proc. Workshop in Data Mining and Knowledge Discovery, SIGMOD*, pp. 15-1–15-8,1998

[14] S. Soderland, "Learning information extraction rules for semistructured and free text," *Machine Learning (Special Issue Natural Language Learning)*, vol. 34, no. 1/3, pp. 233–272, 1999.

[15] D. Freitag and A. McCallum, "Information extraction from hmm's and shrinkage," presented at the Proc. AAAI-99 Workshop Machine Learning Inform. Extraction, Orlando, FL, 1999.

[16] D. Bikel, R. Schwartz, and R. Weischedel, "An algorithm that learns what's in a name," *Machine Learning (Special Issue on Natural Language*

[17] T. Kohonen, "Self organizing maps for large documents," *IEEE Trans. Neural Networks (Special Issue on Data Mining)*, vol. 11, pp. 574–589, June 2000.

[18] J. Shavlik and T. Eliassi, "A system for building intelligent agents that learn to retrieve and extract information," *Int. J. User Modeling User Adapted Interaction (Special Issue on User Modeling and Intelligent Agents)*, Apr. 2001.

[19] S. Kim and B. T. Zhang, "Web document retrieval by genetic learning of importance factors for html tags," in *Proc. Int. Workshop Text Web Mining*, Melbourne, Australia, pp. 13–23, 2000

[20] M. Boughanem, C. Chrisment, J. Mothe, C. S. Dupuy, and L. Tamine, "Connectionist and genetic approaches for information retrieval," in *Soft Computing in Information Retrieval: Techniques and Applications*, F. Crestani and G. Pasi, Eds. Heidelberg, Germany: Physica-Verlag, vol. 50, pp. 102–121,2000

[21] V. U. Maheswari, A. Siromoney, and K. M. Mehata, "The variable precision rough set model for web usage mining," presented at the Proc. 1st Asia-Pacific Conf.Web Intell. (WI-2001, Maebashi, Japan, Oct. 2001

[22] Sankar.K.Pal, Varun Talwar and Pabitra Mitra, "Web mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE Transactions on Neural Networks, September 2002

.