International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 1, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Web Mining in Soft Computing Relevance and Future Directions

Amandeep Kour
Deptt. of CSE,
Lovely Professional University, Punjab

Vimal Kishore Yadav
Deptt. of Energy Science and Engineering,
Indian Institute of Technology, Bombay, India

Vikas Maheshwari
Deptt. of ECE,
School of Engg. and Technology
Apeejay Stya University, Sohna, Gurgaon, Haryana, India

Deepak Prashar
Deptt. of ECE,
Lovely Professional University, Punjab

Abstract **- This paper summarizes the different characteristics of web data, the basic components of web mining and its different types. Web mining combines two of the activated research areas: Data Mining and World Wide Web. The Web mining research relates to several researches communities such as Database, Knowledge Discovery, Information Retrieval and Artificial Intelligence. The limitations of some of the existing web mining and knowledge discovery methods and tools are enunciated, and the significance of soft computing (comprising fuzzy logic (FL), artificial neural networks (ANNs), genetic algorithms (GAs), and rough sets (RSs)) highlighted. A survey of the existing literature on "soft web mining" is provided along with the commercially available systems. The prospective areas of web mining where the application of soft computing needs immediate attention are outlined with justification. Scope for future research in developing "soft web mining" systems is explained. An extensive bibliography is also provided.**

Keywords **- Artificial Neural Networks (ANNs), Data Mining, Fuzzy Logic (FL), Genetic Algorithms (GAs), Information Retrieval (IR), Knowledge Discovery, Pattern Recognition, Rough Sets (RSs), Search Engines.**

I. INTRODUCTION

World Wide Web (WWW) continues to grow at a very high speed containing a huge amount of information available online as an information gateway and also a medium for conducting business. Due to its convenience and rich information, the WWW is a fertile area for data mining research [1]. Along with the amount of information available on the WWW, the number of people connected to the Internet and the number of web pages accessed have also increased exponentially over the past years. There are a great variety of resources and information available on the web for people with the most diverse background and interests. However, one of the major problems is the poor quality of information retrieved. Thus, in most of the cases, users have to search for and filter out the relevant information by themselves. Even qualified users, such as students, researchers and lecturers, do spend time searching and filtering the information retrieved from the Web. This is because of the lacking for performing incorporate and embeds artificial intelligence into web tools. The necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the internet and in particular web localities is drawing the attention of researchers from the domains of information retrieval, knowledge discovery, machine learning, and artificial intelligence (AI), among others additional computation over their results in standard search engines. To proceed toward web intelligence, obviating the need for human intervention, we need to However, 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 is far from being solved. To handle these characteristics and overcome some of the limitations of existing methodologies, soft computing seems to be a good candidate; the research area combining the two may be termed as "soft web mining."

Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability for handling real-life ambiguous situations. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, low-cost solutions, and close resemblance to human-like decision making. The guiding principle is to devise methods of computation that lead to an acceptable solution at low cost by seeking for an approximate solution to an imprecisely/precisely formulated problem. At present, the principal soft computing tools include fuzzy sets, artificial neural networks (ANNs), genetic algorithms (GAs), and rough set (RS) theory. Fuzzy sets provide a natural framework for the process in dealing with uncertainty. Neural networks (NNs) are widely used for modeling complex functions, and provide learning and generalization capabilities. GAs are an efficient search and optimization tool. RSs help in granular computation and knowledge discovery.

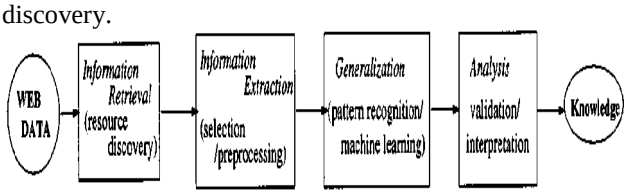

Fig.1. Web mining subtasks includes information retrieval, information extraction, generalization and analysis processes.

II. WEB MINING

As many believe, it is Oren Etzioni first proposed the term of Web mining in his paper [2] 1996. In this paper, he claimed the Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. Many of the following researchers cited this

explanation in their works. In the same paper, Etzioni came up with the question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has become the focus of quite a few research projects and papers. Some of the commercial consideration has presented on the schedule. They suggested a similar way to decompose Web mining into the following subtasks as shown in figure 1:

a) Resource Discovery: the task of retrieving the intended information from Web.
b) Information Extraction: automatically selecting and pre-processing specific information from the retrieved Web resources.
c) Generalization: automatically discovers general patters at the both individual Web sites and across multiple sites.
d) Analysis: analyzing the mined pattern.

Web mining can be broadly 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 an organization's database. Depending on the location of the source, the type of collected data differs. It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and meta information, that might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are:

1) Unlabeled;
2) Distributed;
3) Heterogeneous (mixed media);
4) Semi structured;
5) Time varying;
6) High dimensional.

Therefore, web mining basically deals with mining large and hyper-linked information base having the aforesaid characteristics. Also, being an interactive medium, human interface is a key component of most web applications. Some of the issues which have come to light, as a result, concern

1) Need for handling context sensitive and imprecise queries;
2) Need for summarization and deduction;
3) Need for personalization and learning.

Thus, web mining, though considered to be a particular application of data mining, warrants a separate field of research, mainly because of the aforesaid characteristics of the data and human related issues.

III. Web Mining Categories

Web mining may be of three types, namely, Web Centered Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM). Let us now describe them.

3.1 WCM

WCM deals with the discovery of useful information from the web contents/ data/ documents/ services. However, web contents are not only text, but encompass a very broad range of data such as audio, video, symbolic, metadata, and hyperlinked data. Out of these, research at present is mostly centered on text and hypertext contents. The web text data can be of three types:

1) Unstructured data such as free text;
2) Semi structured data such as HTML;
3) Fully structured data such as in tables or databases.

3.2 WSM

WSM pertains to mining the structure of hyperlinks within the web itself (inter document structure unlike WCM, which pertains to intra document structure). Here, structure represents the graph of the links in a site or between sites. WSM reveals more information than just the information contained in documents. The link topology of the web has also been exploited to develop a notion of hyper linked communities. The analysis shows that communities can be viewed as containing a core of central authoritative pages linked together by hub pages and they exhibit a natural type of hierarchical topic generalization that can be inferred directly from the pattern linkage. It shows that the notion of community provides a surprisingly clear perspective from which to view the seemingly haphazard development of web infrastructure [3]. The Page Rank [4] and CLEVER [5] methods take advantage of this information conveyed by the links to find pertinent web pages. Focused Crawling [6] is a further enhancement in the field of hypertext resource discovery system. The goal of a focused crawler is to selectively seek out pages that are relevant to a predefined set of topics. The topics are specified not using keywords, but using exemplary documents. Rather than collecting and indexing all accessible web documents to be able to answer all possible ad hoc queries, a focused crawler analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the Web. This leads to significant savings in hardware and network resources, and helps to keep the crawl more up-to-date.

3.3 WUM

While content mining and structure mining utilize the real or primary data on the web, usage mining mines secondary data generated by the users' interaction with the web. Web usage data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks and scrolls, and any other data generated by the interaction of users and the web.WUM works on user profiles, user access patterns, and mining navigation paths which are being heavily used by e-commerce companies for tracking customer behavior on their sites. WUM plays a key role in personalizing space, which is the need of the hour. To satisfy all the users with the same tool is extremely difficult, and, instead, we need to learn user access patterns, their path patterns at an individual level, and also as a whole at web sites. Besides learning access patterns, one needs to use "collaborative filtering" for listing other users with similar interests. Collaborative recommender systems allow personalization for e-commerce by exploiting similarities

and dissimilarities in users' preferences. A new algorithm is suggested in [7] for specifically catering to association rule mining in collaborative recommendation systems. In [8], a framework is given for applying machine learning algorithms along with feature reduction techniques, such as singular value decomposition (SVD), for collaborative recommendation. It uses feature reduction techniques to reduce the dimension of the rating data and then NNs are applied on the simplified data to make a model for collaborative recommendation. However, the discovery of patterns from usage data by itself is not sufficient for performing personalization tasks. A way of deriving well quality and useful "aggregate user profiles" from patterns is suggested in [9]. It evaluates two techniques based on clustering of user transactions and clustering of page views, in order to discover overlapping aggregate profiles that can be effectively used by recommender systems for real-time personalization. A framework for web mining has been proposed in WEBMINER [10] for pattern discovery from WWW transactions.

IV. Limitations of Existing Web Mining Methods

The web creates new challenges to different component tasks of web mining (Figure 1) as the amount of information on the web is increasing and changing rapidly without any control. As a result, the existing systems find difficulty in handling the newly emerged problems during IR, IE, generalization (clustering and association), and analysis. Some of these are described below.

4.1 Information Retrieval

The following difficulties may be encountered during this task.

4.1.1 Subjectivity, Imprecision, and Uncertainty

The aim of an IR system is to estimate the relevance of documents to users' information needs, expressed by means of queries. This is a hard and complex task which most of the existing IR systems find difficult to handle due to the inherent subjectivity, imprecision, and uncertainty related to user queries. Most of the existing IR systems offer a very simple modeling of retrieval, which privileges the efficiency at the expense of accuracy. Query processing in search engines, which are an important part of IR systems, is simple blind keyword matching. This does not take into account the context and relevance of queries with respect to documents, while these are important for efficient machine learning.

4.1.2 Deduction

The current search engines have no deductive capability. For example, none of them gives a satisfactory response to a query like: How many computer science graduates were produced by European universities in 1999?

4.1.3 Soft Decision

Current query processing techniques follow the principle of hard rejection while determining the relevance of a retrieved document with respect to a query. This is not correct since relevance, itself, is a "gradual" property of the documents [10], not a crisp one.

4.1.4 Page Ranking

Page ranks are important since human beings find it difficult to scan through the entire list of documents returned by the search engine in response to his/her query. Rather, one sifts through only the first few pages, say less than 20, to get the desired documents. Therefore, it is desirable, for convenience, to get the pages ranked with respect to "relevance" to user queries.

However, there is no definite formula which truly reflects such relevance in top-ranked documents. The scheme for determining page ranks should incorporate: 1) weights given to various parameters of the hit like location, proximity, and frequency; 2) weight given to reputation of a source, i.e., a link from yahoo.com should carry a much higher weight than a link from any other not so popular site; and 3) ranks relative to the user.

4.1.5 Personalization

It is necessary that IR systems tailor 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. Although efforts in this direction have been made by clustering logged data, the similarity metric used in clustering is not meaningful and the principle on which it is derived is not clear.

4.1.6 Dynamism, Scale, and Heterogeneity

IR systems find difficulty in dealing with the problem of dynamism, scaling, and heterogeneity of web documents. Because of the time-varying nature of web data many of the documents returned by the search engines are outdated, irrelevant, and unavailable in the future, and, hence the user has to try his queries across different indexes several times before getting a satisfactory response. Regarding the scaling problem, Etzioni [2] has studied the effect of data size on precision of the results obtained by the search engine. Current IR systems are not able to index all the documents present on the web and this leads to the problem of "low recall." The heterogeneity nature of web documents demands a separate mining method for each type of data.

4.2 IE

Most of the IE algorithms used by different tools are based on the "wrapper" technique. Wrappers are procedures for extracting particular information from web resources. Its biggest limitation is that each wrapper is an IE system 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. Here, ad hoc formatting conventions, used in one site, are rarely relevant elsewhere.

4.3 Generalization

The following difficulties may arise during this task:

4.3.1 Clustering

IR community has explored document clustering as an alternative method of organizing retrieved results, but clustering is yet to be deployed on the major search

engines. Google [4], which seems to be the most effective search engine to date, currently supports simple hostname-based clustering. Besides, there are some problems in efficient clustering arising out of the nature of web data itself.

4.3.2 Outliers

The web server, which logs the data of all users and of their transactions, has many outliers (bad observations), including incomplete, noisy, and vague data due to various reasons inherent in web browsing and logging. These outliers are not a very small percentage of the database since many users just follow links, which are easily visible, big, and prominent. These outliers, in web log server data during WUM, mainly arise because users end up traversing paths which are not in accordance with their interests. Since information on the web is distributed widely, spotting outliers is difficult without clustering the data.

4.3.3 Association Rule Mining

In association rule mining, the current techniques are not able to appropriately mine for linguistic association rules which are more human understandable. Some algorithms which convert linguistic rules to numeric ones suffer from the problem of "hard" rejection. Also, the use of sharp boundary intervals is not intuitive with respect to human perception. For example, an interval method may classify a person young if the age is less than 35, and old if it is greater than 35 years. This obviously does not always correspond to the human perception of "young" and "old," which considers the boundaries of these imprecise concepts, not hard/crisp.

4.4 Analysis

The biggest problem faced in this step is from the point of view of knowledge discovery and modeling. 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. This is so, because the patterns discovered are mainly in mathematical form.

V. Soft Computing and its Relevance

Soft computing is a consortium of methodologies which work synergistically and provides in one form or another flexible information processing capabilities for handling real-life ambiguous situations. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, low-cost solutions, and close resemblance to human-like decision making [11]. In other words, it provides the foundation for the conception and design of high machine IQ (MIQ) systems, and, therefore, forms the basis of future generation computing systems. At this juncture, FL, RSs, ANNs, and GAs are the principal components, where FL provides algorithms for dealing with imprecision and uncertainty arising from vagueness rather than randomness, RS for handling uncertainty arising from limited objects, ANN the machinery for learning and adaptation, and GA for optimization and searching. Relevance of soft computing to pattern recognition and image processing is extensively established in the literature [12] [13]. Recently, the application of soft computing to data mining problems has also drawn the attention of researchers. A recent review [14] is a testimony in this regard. Here, FL is used for handling issues related to incomplete/imprecise data/query, approximate solution, human interaction (linguistic information), and understandability of patterns and deduction, and mixed media information (fusion). NNs are used for modeling highly nonlinear decision boundaries, generalization and learning (adaptively), self organization, rule generation, and pattern discovery. Gas is seen to be useful for prediction and description, efficient search, and adaptive and evolutionary optimization of complex objective functions in dynamic environments. 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 [13]. In this context, we mention about the computational theory of perception which is explained recently by Zadeh [15] as the basic theory behind performing the tasks like driving a car in a city, cooking a meal, and summarizing a story, in our day to day life. Here, computation may be done with perception, which is fuzzy-granular in nature. Web data, being inherently unlabeled, precise/incomplete, heterogeneous, and dynamic, appears to be a very good candidate for its mining in the soft computing framework.

VI. NNs and Learning Systems for Web Mining

An NN can formally be defined as: a massively parallel interconnected network of simple (usually adaptive) processing elements which is intended to interact with the objects of the real world in the same way as biological systems do. NNs are designated by the network topology, connection strength between pairs of neurons (called weights), node characteristics, and the status updating rules. Normally, an objective function is defined which represents the complete status of the network and the set of minima of it corresponds to the set of stable states of the network. NN-based systems are usually reputed to enjoy the following major characteristics: generalization capability, adaptability to new data/information, speed due to massively parallel architecture, robustness to missing, confusing, ill-defined/noisy data, and capability for modeling nonlinear decision boundaries. NNs have been applied, so far, to the tasks like IR, IE, and clustering (self organization) of web mining, and for personalization. We summarize the existing literature on these lines as follows. Some of the prospective areas which need immediate attention are also discussed.

6.1 IR

ANNs provide a convenient method of knowledge representation for IR applications. Also their learning ability helps to achieve the goal of implementing adaptive systems. Shavlik [16] suggests an agent, the Wisconsin Adaptive Web Assistant (WAWA-IE IR) system, using NNs with reinforcement learning, which uses two network modules namely, ScorePage and ScoreLink. ScoreLink uses unsupervised learning, while ScorePage uses supervised learning in the form of advice from the users. The system uses knowledge-based NNs (KBNNs) as its knowledge base to encode the initial knowledge of users which is then refined. This has the following advantages: 1) the agent is able to perform reasonably well initially because it can utilize the users' prior knowledge and 2) users' prior knowledge does not have to be correct as it is refined through learning. Information is derived by extracting rules from KBNNs [17]. In order to map large sized web pages into fixed-sized NNs, a concept of sliding window is used. This parses each page considering three words at a time, and the html tags like act as window breakers. Using self generated training examples it can act also as a self tuning agent. Rules of the type: when "precondition" then "action" are extracted where actions could be of the type strength, followed by "show page" or "follow link," or "avoid showing page." Here, strength could be weakly, moderately, strongly, or definitely, which are determined by the weight of the links between layers of the NN.

6.2 Self-Organization (WEBSOM)

The emerging field of text mining applies methods of data mining and exploratory data analysis to analyze text collections and to convey information to the users in an intuitive manner. Visual map-like displays provide a powerful and fast medium for portraying information about large collections of text. Relationships between text items and collections, such as similarity, clusters, gaps, and outliers, can be communicated naturally using spatial relationships, shading, and colors. In WEBSOM [18], 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. With the help of an HTML-based interactive tool the ordered landscape can be used in browsing the document collection and in performing searches on the map. An organized map offers an overview of an unknown document collection helping the user in familiarizing oneself with the domain. Map displays that are already familiar can be used as visual frames of reference for conveying properties of unknown text items.

6.3 Personalization

Personalization means that the content and search results are tailored as per users' interests and habits. NNs may be used for learning user profiles with training data collected from users or systems. Since user profiles are highly nonlinear functions, NNs seem to be an effective tool to learn them. An agent which learns user profiles using Bayesian classifier is "Syskill and Webert" [19]. Once the user profiles have been learned, it can be used to determine whether the users would be interested in another page. However, this decision is made by analyzing the HTML source of a page, and it requires the page to be retrieved first. To avoid network delays, we allow the user to prefetch all pages accessible from the index page and store them locally. Once this is done, Syskill and Webert can learn a new profile and make suggestions about pages to visit quickly. Once the HTML is analyzed, it annotates each link on the page with an icon indicating the user's rating or its prediction of the user's rating together with the estimated probability that a user would like the page. Note that these ratings and predictions are specific to only one user and do not reflect on how other users might rate the pages.

VII. GAs for Web Mining

7.1 GAs

A biologically inspired technology is 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 is executed iteratively on a set of coded solutions (genes), called population, with three basic operators: Selection/reproduction, crossover, and mutation. They use only the payoff (fitness function) information and probabilistic transition rules for moving to the next iteration. The literature explaining the use of GAs to web mining seems to be even poorer than that of FL and NNs. GAs are used, mainly in search, optimization, and description. Here we describe some of the attempts.

Search and Retrieval

A GA-based search to find other relevant homepages, given some user-supplied homepages, has been implemented in G-Search [20].Web document retrieval by genetic learning of importance factors of HTML tags has been described in [21].

7.2 Query Optimization

In [22], 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. Yang *et al.* [23] presented an evolutionary algorithm for query optimization by reweighting the document indexing without query expansion. Kraft *et al.* [24] apply genetic programming in order to improve weighted Boolean query formulation.

7.3 Document Representation

Gordon [25] adopted a GA to derive better descriptions of documents. Here each document is assigned descriptions where each description is a set of indexing terms. Then genetic operators and relevance judgments are applied to these descriptions in order to determine the best one in terms of classification performance in response to a

specific query. Automatic web page categorization and updating can also be performed using GAs [26].

7.4 Distributed Mining

Gene expression messy GA (GEMGA) [27] which is a sub quadratic highly parallel evolutionary search algorithm, is specially found suitable for distributed data mining applications including the web. Foundation of GEMGA is laid on the principle of both decomposing black box search into iterative construction of partial ordering and performing selection operation in the relation, class, and sample spaces.

VIII. Result

We have seen how the web mining is capable in handling sensitive and imprecise queries, able to summarize and deduction, also provides human interface key, which is useful factor in handling tremendous growth of the data sources on the Web due to the dramatic growth in the e-commerce in the business community web mining, though it considered to be a particular application of data mining.

IX. Conclusion and Discussion

Web mining is growing rapidly since its inception in or around 1996, and new methodologies are being developed both using classical and soft computing approaches concurrently. Considering the immense potential of application of soft computing to web mining, this paper is timely and appropriate. In this paper, we have summarized the different types of web mining and its basic components, along with their current states of art. The limitations of the existing web mining methods/tools are explained. The relevance of soft computing, including integration of its constituting tools, is illustrated through examples and diagrams. Their applications to each web mining task along with the commercially available systems are described. Last, the possible future directions of using FL, ANNs, GAs, and RSs for some of these tasks are given in detail. In addition, to those discussed in the article, some aspects of web mining, in general, where soft computing is likely to play a key role, in future, are as follows.

1) At present web content is mainly text-centric, and most mining algorithms are oriented toward and developed from text mining framework. However, web is increasingly gaining a multimedia character with pages containing images, videos, etc. Web mining algorithms having capabilities for handling multimedia data need to be developed in near future.

2) Currently, queries are in the form of keywords, advanced search engines may support visual queries. In this regard, the research on CBIR in soft computing framework has potential significance.

3) Most search engines perform search on English text only, not across languages. With these becoming increasingly common, multilingual search engines and IR systems which can identify languages, translate, perform thematic classification, and can provide summaries automatically are recently being developed. Soft computing may be used to increase the efficiency of such systems.

4) Collaborative mining and automatic interaction among sites and constitute a recent research area (e.g., the NET paradigm of Microsoft). Here, web query is not the only means to obtain required documents, and answers to queries can be automatically obtained from distributed web resources. Thus, text sources could be used for planning and problem solving tasks (e.g., an agent on the web could be used to make ones' travel plans automatically). Significance of applying soft computing for the above tasks may therefore be explored.

References

[1] R. Kosala and H. Blockeel, "Web Mining Research: A Survey", *Journal of ACM SIGKDD Explorations*, Vol. 2, Issue 1, 2000.

[2] O. Etzioni, "The world wide web: Quagmire or gold mine," *Commun. ACM*, vol. 39, no. 11, pp. 65–68, 1996.

[3] D. Gibson, "Inferring web communities from link topologies," presented at the U.K. Conf. Hypertext, 1998.

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

[5] D. Konopnicki and O. Shmulei, "W3qs: A query system for the world wide web," in *Proc. 21st VLDB Conf.*, Zurich, Switzerland, 1995, pp. 54–65.

[6] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," presented at the 8th World Wide Web Conf., Toronto, ON, Canada, May 1999.

[7] W. Y. Lin, S. A. Alvarez, and C. Ruiz. Collaborative recommendation via adaptive association rule mining. presented at Int. Workshop Web Mining for E-Commerce

[8] J. Pazzani and D. Billsus, "Learning collaborative information filters," presented at the Proc. 15th Int. Conf. Machine Learning, Madison, WI, 1998, pp. 46–54.

[9] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Y. Sun, and J. Wiltshire, "Discovery of aggregate usage profiles for web personalization," presented at the Proc. KDD-2000 Workshop Web Mining E-Commerce, Boston, MA, Aug. 2000.

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

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

[12] .S. K. Pal, A. Ghosh, and M. K. Kundu, Eds., *Soft Computing for Image Processing*. Heidelberg, Germany: Physica-Verlag, 2000.

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

[14] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: A survey," *IEEE Trans. Neural Networks*, vol. 13, pp. 3–14, Jan.2001.

[15] .L. A. Zadeh, "A new direction in AI: Toward a computational theory of perceptions," *AI Mag.*, vol. 22, pp. 73–84, 2001.

[16] .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.

[17] .J. Shavlik and G. G. Towell, "Knowledge-based artificial neural networks,"*Artificial Intell.*, vol. 70, no. 1/2, pp. 119–165, 1994.

[18] .H. Chen, M. Ramsay, and P. Li, "The Java search agent workshop," in *Soft Computing in Information Retrieval: Techniques and Applications*, F. Crestani and G. Pasi, Eds. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50, pp. 122–140.

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

[20] .M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill and webert: Identifying interesting web sites," in *Proc. 13th Nat. Conf. AI*, 1996, pp. 54–61.
[21] .F. Crestani and G. Pasi, Eds., *Soft Computing in Information Retrieval: Techniques and Application*. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50.
[22] .S. Kim and B. T. Zhang, "Web document retrieval by genetic learning of importance factors for html tags," in *Proc. Int. Workshop Text WebMining*, Melbourne, Australia, Aug. 2000, pp. 13–23.
[23] .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, 2000, vol. 50, pp. 102–121.
[24] .J. J. Yang and R. Korfhage, "Query Modification Using Genetic Algorithms in Vector Space Models," Dept. IS, Univ. Pittsburgh, Pittsburgh, PA, TRLIS045/1 592 001, 1992 D. H. Kraft, F. E. Petry, B. P. Buckles, and T. Sadasivan, "The use of genetic programming to build queries for information retrieval," *Proc. IEEE Symp. Evol. Comput.*, 1994.
[25] M. D. Gordon, "Probabilistic and genetic algorithms for document retrieval," *Commun. ACM*, vol. 31, no. 10, pp. 208–218, 1988.
[26] V. Loia and P. Luongo, "An evolutionary approach to automatic web page categorization and updating," in *Web Intelligence: Research and Development*, N. Zhong, Y. Yab, J. Liu, and S. Oshuga, Eds, Singapore: Springer-Verlag, 2001, vol. LNCS 2198, pp. 292–302.
[27] .H. Kargupta, "The gene expression messy genetic algorithm," *Proc.IEEE Int. Conf. Evol. Comput.*, pp. 631–636, 1996.

AUTHOR'S PROFILE

Amandeep Kour

passed B.E degree in information technology from Mahant Bachittar Singh College of Engineering and Technology, Jammu University,India in the year 2009. She received the diploma in Intrnational cross cultural research and human resource management from The Bussiness School, Jammu university, India in the Year 2010. Presently she is persuing her M.Tech in computer Science from Lovely Professional University , Jalandhar,india.Her research intrest includes Face recognition using Template Matching.Her 2 paper is accepted in International Journals and 1 paper in International Conference(IEEE).

Vimal Kishore Yadav

passed B.Tech degree in Electronics and engineering from Anand Engineering college,Agra,U.P Technical University Lucknow. He served as an Assistant Proffesor in ECE department at Hindustan College of Science and Technology Mathura. He has published 2 Papers in International conference in IEEE and 4 papers in International Journal. Presently he is persuing his M.Tech in Energy Science and Engineering at IIT Bombay, Mumbai, India.