# Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions

Sankar K. Pal, *Fellow, IEEE*, Varun Talwar, *Student Member, IEEE*, and Pabitra Mitra, *Student Member, IEEE*

***Abstract*—This paper summarizes the different characteristics of web data, the basic components of web mining and its different types, and their current states of the art. The reason for considering web mining, a separate field from data mining, is explained. The limitations of some of the existing web mining 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.**

***Index Terms*—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

OVER the last decade, we have witnessed an explosive growth in the information available on the World Wide Web (WWW). Today, web browsers provide easy access to myriad sources of text and multimedia data. More than 1 000 000 000 pages are indexed by search engines, and finding the desired information is not an easy task. This profusion of resources has prompted the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "web mining."

To proceed toward web intelligence, obviating the need for human intervention, we need to incorporate and embed 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. 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.

The objective of this article is to provide an outline of web mining, its various classifications, its subtasks, and to give a perspective to the research community about the potential of applying soft computing techniques to its different components. The article, besides reviewing the existing techniques/tools and their limitations, lays emphasis on possible enhancements of these tools using soft computing framework. In this regard, the relevance of fuzzy logic (FL), ANNs, GAs, and RSs is illustrated through examples and diagrams, along with the mention of some commercially available systems. Broad guidelines for future research, and on web mining in general, are outlined. It should be noted that the use of soft computing in "web mining" is a field in its genesis, and thus the significance of this paper at this juncture is evident.

The rest of this paper is organized as follows: Section II deals with the characteristics of web data, and the different components and types of web mining. The limitations of existing web mining methods are discussed in Section III. Section IV provides an introduction to soft computing and its relevance. Sections V and VI cover, in detail, the use of FL and NNs in the different phases of web mining along with some prospective areas of applications. Sections VII and VIII give brief outline of possible applications of GAs and RSs in web mining. Section IX provides the conclusion and scope of future research in the area of soft web mining.

Manuscript received July 30, 2001; revised January 30, 2002.

S. K. Pal and P. Mitra are with the Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700035, India (e-mail: sankar@isical.ac.in; pabitra_r@isical.ac.in).

V. Talwar is with the Department of Computer Science, Netaji Subhas Institute of Technology, New Delhi 110045, India (e-mail: varun.talwar@ieee.org).

Publisher Item Identifier S 1045-9227(02)05562-5.

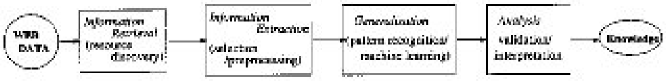
WEB DATA
Information Retrieval (resource discovery)
Information Extraction (selection /preprocessing)
Generalization (pattern recognition/ machine learning)
Analysis validation/ interpretation
Knowledge

Fig. 1. Web mining subtasks.

## II. Web Mining

The web is a vast collection of completely uncontrolled heterogeneous documents. Thus, it is huge, diverse, and dynamic, and raises the issues of scalability, heterogeneity, and dynamism, respectively. Due to these characteristics, we are currently drowning in information, but starving for knowledge; thereby making the web a fertile area of data mining research with the huge amount of information available online. Data mining refers to the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

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) semistructured;
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.

### A. Web Mining Components and the Methodologies

Web mining can be viewed as consisting of four tasks, shown in Fig. 1, according to Etzioni [1]. Each task is described below along with a survey of the existing methodologies/tools for the task.

*1) Information Retrieval (IR) (Resource Discovery):* Resource discovery or IR deals with automatic retrieval of all relevant documents, while at the same time ensuring that the nonrelevant ones are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.

Table I [2] shows the different techniques used by various authors for document representation in IR of web content mining (to be explained in detail in Section III) for semistructured documents. An index is, basically, a collection of terms with pointers to places where the information about documents can be found. However, indexing of web pages to facilitate retrieval is quite a complex and challenging problem as compared to the corresponding one associated with classical databases where straightforward techniques suffice. The enormous number of pages on the web, their dynamism, and frequent updating make the indexing techniques seemingly impossible. At present, four approaches to index documents on the web are human or manual indexing, automatic indexing, intelligent or agent-based indexing, and metadata-based indexing.

Search engines are programs written to query and retrieve information stored in databases (fully structured), HTML pages (semistructured), and free text (unstructured) on the web. The most popular indexes have been created by web robots such as AltaVista and WebCrawler which scan millions of web documents and store an index of the words in the documents. There are over a dozen different indexes currently in active use, each with a unique interface and a database covering a different fraction of the web. MetaCrawler presents the next level of information food chain by providing a single unified interface for web document searching [3]. It submits the query to nine indexes in parallel, and then collates the results and prunes them. Thus, instead of tackling the web directly, MetaCrawler mines robot-created searchable indexes. Future resource discovery systems will make use of automatic text categorization technology to classify web documents into categories. This technology could facilitate the automatic construction of web directories such as Yahoo by discovering documents that fit Yahoo categories. Alternatively, the technology could be used to filter the results of queries to searchable indexes. For example, in response to a query such as "find me product reviews of Encarta," a discovery system could take documents containing the word "Encarta" found by querying searchable indexes and then identify the subset that corresponds to product reviews. Other agents such as W3QL [4] combine structure queries, based on organization of hypertext documents and content queries, based on IR techniques. A number of IR agents use various characteristics of open hypertext web documents to automatically retrieve, filter, and categorize them [5], [6]. Bookmark Organizer (BO) [7] combines hierarchical clustering techniques and user interaction to organize a collection of web documents based on conceptual information. Research in IR also includes modeling, developing user interfaces, data visualization, and filtering [8]. A more detailed survey of IR on the web is available in [9].

*2) Information Selection/Extraction and Preprocessing:* Once the documents have been retrieved the challenge is to automatically extract knowledge and other required information without human interaction. Information extraction (IE) is the task of identifying specific fragments of a single document that constitute its core semantic content. Until now, the major methods of IE involve writing *wrappers* (hand coding) which map the documents to some data model. Information integration systems operate by interpreting the various sites as knowledge sources and extracting information from them. To

TABLE I
IR AND GENERALIZATION TECHNIQUES OF SOME WCM ALGORITHMS

| Author | Document representation (for IR) | Method used (for generalization) | Task performed (for generalization) |
|---|---|---|---|
| Craven et al [34] | Relational and Ontology | Modified Naive Bayes<br>Inductive Logic Programming | Hypertext Classification<br>Learning web page relation<br>Learning extraction rules |
| Crimmins et.al. [35] | Phrase,URLs and meta information | Unsupervised and Supervised<br>Classification algorithms | Hierarchical and Graphical classification<br>Clustering |
| Furnkranz [36] | Bag of words and hyperlinks information | Rule Learning | Hypertext classification |
| Joachims et.al. [37] | Bag of words and hyperlinks information | TFIDF<br>Reinforcement learning | Hypertext prediction |
| Muslea et.al. [38] | Bag of words,tags and word positions | rule learning | Learning extraction rules |
| Shavlik and Elliasi-Rad | Localised bag of words | Neural networks with reinforcement learning | Hypertext(homepage)classification |
| Singh et.al. [39] | Concepts and named entity | Modified association rule<br>Classification algorithm | Finding patterns in semi structured texts |
| Soderland [40] | Sentences, phrases and named entity | Rule learning | Learning extraction rules |

do so the system processes site documents to extract relevant text fragments and a library of wrappers is used wherein each wrapper is an IE system customized for a particular internet site [10]. Another method for IE from hypertext is given in [11] where each page is approached with a set of standard questions. The problem, therefore, reduces to identifying the text fragments which answer those specific questions. In other words, slots are to be filled by text fragments in the document, and these are, thus, called slottfills. Therefore, IE aims to extract new knowledge from the retrieved documents by capitalizing on the document structure and representation of the document, whereas IR experts view the document text just as a bag of words and do not pay attention to the structure of the document. Scalability is the biggest challenge to IE experts; it is not feasible to build IE systems which are scalable to the size and dynamism of the web. Therefore, most IE systems extract from specific sites and focus on some defined areas.

We now give a survey of all systems which aim to dynamically extract information from unfamiliar resources. Some of the intelligent web agents have been developed to search for relevant information using characteristics of a particular domain (and possibly a user profile) to organize and interpret the discovered information. Agents such as Harvest [12], FAQ-Finder [13], Information Manifold [14], OCCAM [15], and Parasite [16] rely either on prespecified domain specific information about particular types of documents, or on hard-coded models of the information sources to retrieve and interpret documents. The Harvest system [12] relies on semistructured documents to improve its ability to extract information. For example, it knows how to find author and title information in Latex documents and how to strip position information from postscript files. Harvest neither discovers new documents nor learns new models of document structure. Similarly, FAQ-Finder [13] extracts answers to frequently asked questions (FAQs) from FAQ files available on the web. ShopBot [17] and Internet Learning Agent (ILA) [18] attempt to interact with and learn the structure of unfamiliar information sources. ShopBot retrieves product information from a variety of vendor sites using only general information about the product domain. ILA, on the other hand, learns models of various information sources and translates these into its own concept hierarchy. ILA learns to extract information from unfamiliar resources by querying them with familiar objects and then matching the returned output with the knowledge about the query objects. An IE system, built at Carnegie Mellon University (CMU), Pittsburgh, PA, extracts the course, instructor's name, his/her email address, research interests, from the homepages of the faculty members of the computer science department [19].

A robust preprocessing system is required in order to extract any kind of knowledge from even medium sized databases. When a user requests a web page, a variety of files like images, sound, video, executable cgi and html, are accessed. As a result, the server log contains many entries that are redundant or irrelevant for mining tasks. Therefore, these need to be removed through preprocessng. One of the preprocessing techniques used for IE is latent semantic indexing (LSI) that seeks to transform the original document vectors to a lower dimensional space by analyzing the correlational structure of terms in that document collection such that similar documents that do not share the same terms are placed in the same category (topic). "Stemming" is yet another preprocessing technique, which reduces the input feature size by stemming words like *informed*, *information*, *informing* to their root *inform*.

*3) Generalization:* In this phase, pattern recognition and machine learning techniques are usually used on the extracted information. Most of the machine learning systems, deployed on the web, learn more about the user's interest than the web itself. A major obstacle when learning about the web is the labeling problem: data is abundant on the web but it is unlabeled. Many data mining techniques require inputs labeled as positive (yes) or negative (no) examples with respect to some concept. For example, if we are given a large set of web pages labeled as positive and negative examples of the concept *homepage*, then it is easy to design a classifier that predicts whether any unknown web page is a home page or not; unfortunately web pages are unlabeled. Techniques such as uncertainty sampling reduce the amount of unlabeled data needed, but do not eliminate the labeling problem. An approach to solve this problem is based on the fact that the web is much more than just a linked collection of documents, it is an interactive medium. For example, "Ahoy" [20] takes as input a person's name and affiliation, and attempts to locate the person's homepage; hence it asks the users to label its answers as correct or incorrect. Clustering techniques do not require labeled inputs and have been applied successfully to large collections of documents [21]. Indeed, the web offers a fertile ground for document clustering research. Table I shows the different tasks performed and methods used for generalization in web content mining (WCM). Association rule mining is also an integral part of this phase. Basically, association rules are expressions of the type $X \Rightarrow Y$ where $X$ and $Y$ are sets of items. $X \Rightarrow Y$ expresses that whenever a transaction $T$ contains $X$ then $T$ probably contains $Y$ also. The probability or rule confidence is defined as the percentage of transactions containing $Y$ in addition to $X$ as compared to overall transactions containing $X$. The idea of mining association rules originates from the market-based data where rules like "A customer who buys product $x1$ and $x2$ will also buy product $y$ with probability $c\%$" are found [22].

Machine processable documents have led to the development of the concept of the "semantic web," which is inspired by the fact that most information on the web is designed for human consumption, and even if it was derived from a database with well-defined meanings (in at least some terms) for its columns, the structure of the data is not evident to a robot browsing the web. Leaving aside the AI problem of training machines to behave like people, the semantic web approach, instead, develops languages for expressing information in a machine processable form.

*4) Analysis:* Analysis is a data-driven problem which presumes that there is sufficient data available so that potentially useful information can be extracted and analyzed. Humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and or interpretation of the mined patterns which take place in this phase. Once the patterns have been discovered, analysts need appropriate tools like, Webviz system [23], to understand, visualize, and interpret these patterns. Some others use Online analytical processing (OLAP) techniques such as data cubes for the purpose of simplifying the analysis of usage statistics from sever access logs. The WEBMINER [24] system proposes a structured query language (SQL)-like querying mechanism for querying the discovered knowledge (in the form of association rules and sequential patterns). Cohen [25] focuses on the nature of knowledge that one can derive from the web.

Based on the aforesaid four phases (Fig. 1), *web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. Here, evaluation includes both "generalization" and "analysis."*

### *B. Web Mining Categories*

Web mining may be of three types, namely, WCM, web structure mining (WSM), and web usage mining (WUM). Let us now describe them.

*1) WCM:* WCM deals with the discovery of useful information from the web contents/data/documents/services. As mentioned earlier, Table I describes various kinds of document representation (used for IR) and the tasks performed (for generalization) [2] in WCM. 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 around text and hypertext contents. The web text data can be of three types: 1) unstructured data such as free text; 2) semistructured data such as HTML; 3) fully structured data such as in tables or databases. Mining 1) is termed as knowledge discovery in texts (KDT) or text data mining or text mining. Text mining is a well-developed subject and full coverage of it is beyond the scope of this survey. For details, one may refer to [26]. For mining from symbolic knowledge extracted from the web, one may refer to [27]. In this context, a concept-based knowledge discovery method from texts, extracted from the web, is given in [28]. Here, instead of analyzing words or attribute values, concepts are extracted. Since the content of a text document presents no machine-readable semantic, some approaches have suggested to restructure the document content in a representation that could be exploited by machines. Techniques using lexicons for content interpretation are yet to come. Hypertext mining involves mining semistructured HTML pages which have hyperlinks, besides text. In hypertext mining, at times, supervised learning or classification plays a key role, like in email, newsgroup management, and maintaining web directories. For a good tutorial on hypertext mining, one may refer to [29].

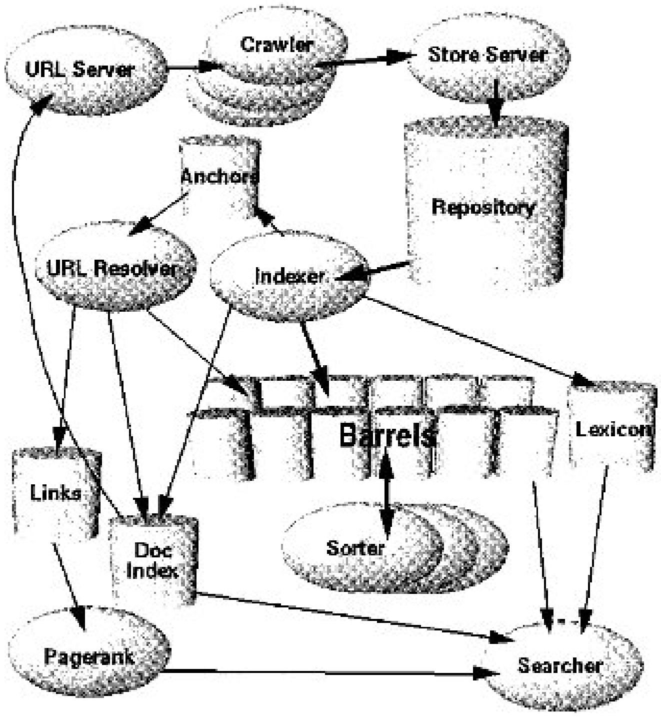

Fig. 2. Google © architecture.

Due to a large number of services like usenet, newsgroups, digital libraries, and mailing lists coming up, mining from services is also gaining importance. Work in mining from digital libraries can be found in [30].

It should be noted that web content mining and IR are separated by a thinline, as some claim IR on the web as an instance of WCM, whereas others associate WCM with intelligent IR. There are two strategies for WCM: those that directly mine the content of documents (web page content mining) and those that improve on the content search of other tools like search engines (serach result mining). Thus, the task of search engines is closely associated with WCM. Fig. 2 shows, as an example, the architecture of the popular Google [31] search engine.

WCM can take two approaches: agent-based approach and database approach. The agent-based approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi autonomously on behalf of a partic-

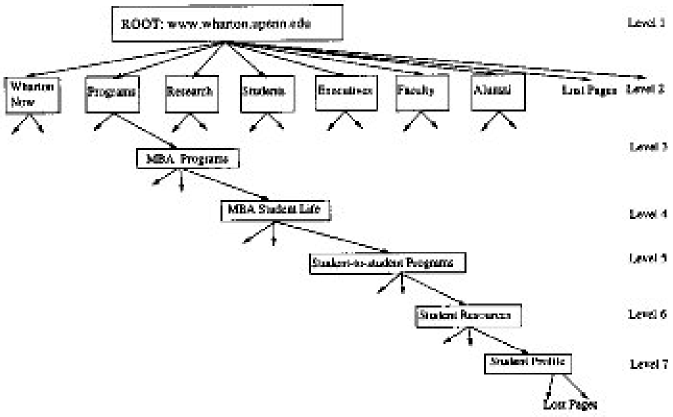

Fig. 3. Web structure mining.

ular user, to discover and organize web-based information. Generally, agent-based web mining approach can be further organized into three categories: intelligent search agents, information filtering/categorization, and personalized web agents [32]. The database approach focuses on techniques for organizing semistructured data on the web into more structured collection of resources, and uses database querying mechanisms and data mining techniques to analyze it. Levy *et al.* [33] discuss general intelligent internet systems with respect to user modeling, discovery, and analysis of remote information sources, information integration, and web-site management.

*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. For example, Fig. 3 shows a sample link structure and levels of the site wharton.upenn.edu. WSM reveals more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This was first highlighted and used in the HITS [41] algorithm. This is analogous to bibliographical citations. When a paper is cited often, it ought to be important. 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 [41]. The Page Rank [31] and CLEVER [4] methods take advantage of this information conveyed by the links to find pertinent web pages. Focused Crawling [42] 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.

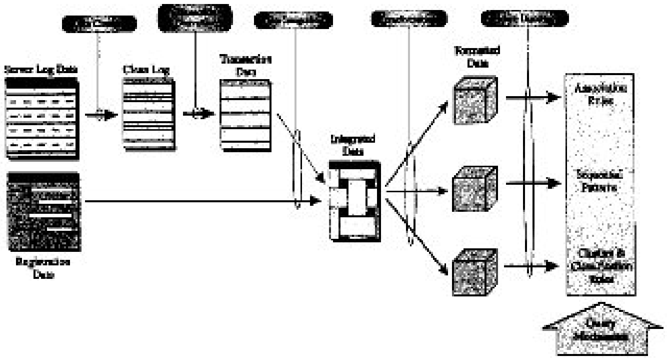

Fig. 4. Webminer architecture.

*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 [43] for specifically catering to association rule mining in collaborative recommendation systems. In [44], 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 good quality and useful "aggregate user profiles" from patterns is suggested in [45]. 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 [24] for pattern discovery from WWW transactions. Fig. 4 shows the WEBMINER architecture. It clearly brings out that, after integrating registration data and "cleaned" transactions data, pattern discovery techniques are applied to discover association rules, sequential patterns and clusters, and classification rules.

## III. Limitations of Existing Web Mining Methods

The web creates new challenges to different component tasks of web mining (Fig. 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.

### A. Information Retrieval

The following difficulties may be encountered during this task.

*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.

*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?

*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 [46], not a crisp one.

*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.

*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.

*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 [47] 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.

### B. IE

Most of the IE algorithms used by different tools are based on the "wrapper" technique. Wrappers are procedures for extracting a 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. Harvest and FAQ Finder, discussed in Section II, also have two key limitations. First, both systems focus exclusively on web documents and ignore services for, e.g., web-log analysis, performance analysis, and customer relationship management. Second, both rely on a prespecified description of certain fixed classes of web documents.

### C. Generalization

The following difficulties may arise during this task:

*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 [31], 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. As mentioned in Section II, the data is not only distributed, heterogeneous, and imprecise, it is also very high dimensional and overlapping. Thus, existing conventional clustering techniques find difficulty in handling these characteristics.

*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.

*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 as 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.

### D. 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.

## IV. Soft Computing and Its Relevance

Soft computing is a consortium of methodologies which work synergestically 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* [48]. 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 discernibility of 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 [49] [50]. Recently, the application of soft computing to data mining problems has also drawn the attention of researchers. A recent review [51] 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), understandability of patterns and deduction, and mixed media information (fusion). NNs are used for modeling highly nonlinear decision boundaries, generalization and learning (adaptivity), self organization, rule generation, and pattern discovery. GAs are 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 [50]. In this context, we mention about the computational theory of perception which is explained recently by Zadeh [52] 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, imprecise/incomplete, heterogeneous, and dynamic, appears to be a very good candidate for its mining in the soft computing framework. Besides, since human interaction is a key component in web mining, as mentioned in Section II, issues such as context-sensitive and approximate queries, summarization and deduction, and personalization and learning are of utmost importance where soft computing seems to be the most appropriate paradigm for providing effective solutions. This realization has drawn the attention of soft computing community to develop "soft web mining" systems in parallel to the conventional ones since its inception in or around 1996. In the following sections, we discuss some such applications of each of the soft computing tools.

## V. FL for Web Mining

The application of FL, so far made, to web mining tasks mainly falls under IR and generalization (clustering, association). These attempts will be described here along with different commercially available systems. Some of the prospective areas which need immediate attention are also outlined.

### A. Information Retrieval

Yager describes in [53] 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." Let $Q$ be a linguistic expression corresponding to a quantifier such as "most" then it is represented as a fuzzy subset $Q$ over $I = [0, 1]$ in which, for any proportion $r$, belonging to $I, Q(r)$ indicates the degree to which $r$ satisfies the concept indicated by the quantifier $Q$. Koczky and Gedeon [54] deal with the problem of automatic indexing and retrieval of documents where it cannot be guaranteed that the user queries include the actual words that occur in the documents that should be retrieved. Fuzzy tolerance and similarity relations are presented, and the notion of "hierarchical cooccurrence" is defined that allows the introduction of two or more hierarchical categories of words in the documents.

As an example of the use of fuzzy set theory to extend boolean information retrieval, we discuss the methodology proposed by Bordonga and Pasi [55] for semistructured document (e.g., HTML) retrieval. It models the concept of graduality of "relevance" of a document to the user query. Formally, a document is represented as a fuzzy binary relation

$$R_d = \sum_{(t) \in T} \mu_d(t)/t \tag{1}$$

where $R_d$ is the representation of document $d \in D$, the set of archive documents $t \in T$, the set of index terms, and $\mu_d\ D \times T \rightarrow [0, 1]$ the membership function of $R_d$. $\mu_d$ is again a dynamic function with $\mu_d(t, s)$ expressing the significance of the term $t$ in section $s$ of document $d$. $\mu_d(t, s)$ are based upon the semantics of the section $s$. For example, in sections containing fromatted text like *author* and *keywords*, a single occurence of a term makes it fully significant $(\mu_d(t, s) = 1)$ for that section, and $\mu_d(t,s)$ is a Boolean function. On the other hand, for sections containing textual descriptions, $\mu_d(t,s)$ can be computed as a function of the normalized term frequency for that section, for example

$$\mu_d(t,s) = \mathrm{tf}_{\mathrm{dst}}\mathrm{IDF}_t \tag{2}$$

in which $\mathrm{IDF}_t$ is the inverse document frequency of the term $t$, $\mathrm{tf}_{\mathrm{dst}}$ is the normalized term frequency defined as

$$\mathrm{tf}_{\mathrm{dst}} = \frac{\mathrm{OCC}_{\mathrm{dst}}}{\mathrm{MAXOCC}_{\mathrm{sd}}} \tag{3}$$

where $\mathrm{OCC}_{\mathrm{dst}}$ is the number of occurrences of the term $t$ in section $s$ of document $d$, and $\mathrm{MAXOCC}_{\mathrm{sd}}$ is a normalization parameter depending on the sections' length. To obtain the overall degree of significance of a term in a document, computed over all the sections, an aggregation scheme ordered weighted average (OWA) is used

$$\mu(d,t) = \mathrm{OWA}_{lq}(\mu_1(d,t),\ldots,\mu_n(d,t)). \tag{4}$$

Parameters $l, q$ are determined by users specified relative weights to the sections.

A query $\langle t, w\rangle$ is represented by terms $t_i$ and the corresponding weights $w_i$. The query is evaluated by an $E$ function for a given document, and then aggregation operators are used. Thus the result of a query evaluation is represented as a fuzzy subset of the archived documents, given by

$$R_d(t) = \sum_{d\in D} \mu_W(d,t)/d. \tag{5}$$

This brings to light that fuzzy Boolean IR models are more flexible in representing both document contents and information needs.

### B. Generalization

*Clustering:* Etzioni [47] has listed the key requirements of web document clustering as measure of relevance, browsable summaries, ability to handle overlapping data, snippet tolerance, speed and incremental characteristics. In [56], fuzzy c medoids (FCNdd) and fuzzy c Trimmed medoids (FCTMdd) are used for clustering of web documents and snippets (outliers). In [57], a fuzzy clustering technique for web log data mining is described. 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. This approach requires the definition and computation of dissimilarity/similarity between all session pairs, forming a similarity or fuzzy relation matrix, prior to clustering. Since the data in a web session involves access method (GET/ POST), URL, transmission protocol (HTTP/FTP), etc., which are all nonnumeric, correlation between two user sessions and, hence, their clustering, is best handled using fuzzy set approach. Other techniques for clustering web data include those using hypergraph-based clustering [58].

*Association Rule Mining:* Some algorithms for mining association rules using FL techniques have been suggested in [59]. They deal with the problem of mining fuzzy association rules understandable to humans from a database containing both quantitative and categorical attributes. Association rules of the form, if $X$ is $A$, then $Y$ is $B$ where $X, Y$ are attributes and $A, B$ are fuzzy sets, are mined. Nauck [60] has developed a learning algorithm that creates *mixed* fuzzy rules involving both categorical and numerical attributes.

### C. Commercially Available Systems

Here, we list some commercially available systems with their characteristics.

- Nzsearch: (www.searchnz.co.nz/) is a search engine based completely on FL. It considers the entire phrase rather than individual words for the purpose of matching. It also uses a "fuzziness" parameter while searching, which can be chosen from the set: "minimal," "normal," "moderate," "very," and "extremely."
- DNS search: (www.amnesi.com/dnssearch) uses FL to find the closest DNS entry to your typed URL. For example, if by mistake one types www.macrosoft.com then the system gives suggestions on the possible close URLs.
- Finder: (www.finder.co.uk) uses "multidimensional optimization" to display the best or most suitable matches to a query unlike most existing search engines which display only the exact matches to a given query. Finder goes way beyond the simple "yes" or "no" criterion, used by most database query engines such as SQL or Btrieve. Finder uses scoring modules designed especially for each data type. If one is looking for a scarlet car, and the car in the database was cherry red, it would not ignore the entry altogether, it would just give it a lower score. The search engine looks at each element of the database and scores it using one of its knowledge-based scoring modules.
- Aptronix: (www.aptronix.com) uses the fuzzy inference data environment (FIDE) to build smart and intelligent Java applets. The FIDE software package provides a rich set of functions for building FL controls, expert systems, automated diagnosis, adaptation to environment, and self-learning. Their software can convert an application program in fuzzy inference language (FIL) into Java, C, MatLab, and assembly code.

### D. Prospective Areas of Application

*Search Engines:* There is immense scope of applying FL to improve web search from the points of view of deduction, matching, and ranking, among others. To add human-like deductive capability to search engine, the use of FL is not an option, rather it is a necessity. Regarding matching, a probable approach is to compromise sightly on precision (which is, anyway, very difficult to achieve due to millions of web pages), and retrieve most "relevant" documents from an expanded domain. The retrieved documents may then be clustered during/after search, or filtered at the client side, or both. The concept of linguistic variables and membership functions can be used for keyword matching. Similarly, for page ranking the degree of closeness of hits in a document can be used for its computation. For example, variables like "close," "far," and "nearby" may be used to represent the distance between hits in a document for a given query. Similarly fuzzy variables like "reputation" and "importance," attached to the URL which is referencing a particular page, can be used in calculating page ranks. For example, in Fig. 6, which shows a neuro-fuzzy IR system, match parameters such as "proximity" and subjectivity in queries can be found using fuzzy sets.

Let us consider the popular Google search engine, which is considered highly effective among the existing ones. In [31] a schematic diagram (Fig. 3) of the technology behind Google has been explained in which we can see that the lexicon gives wordIDs to each word of the query and wordIDs are then matched. If the query contains quantifiers like less, very less, and more, then instead of blindly rejecting/selecting pages based on their absence/presence in the document, a smoother transition based on their membership value is a better option. Considering fuzzy queries i.e., if the query text includes linguistic variables like almost, somewhat, more or less, about, we can provide more relevant documents by giving grades of membership to different results. When we consider hits, greater weight should be given to documents in which query words are closer to each other than those in which they are far apart.

*Similarity Measures:* There are certain questions like: What is the distance between two URLs? Which two URLs are always requested together? Which users have common interests and request similar documents? that appear to be better handled in a fuzzy set theoretic framework since answers to these questions need not always be crisp.

*Others:* Some other areas where FL may be applied include

- ontology;
- matching techniques;
- recognition technology;
- summarization;
- e-commerce;
- content management;
- database querying;
- information aggregation and fusion;
- customization and profiling.

## 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, adaptivity 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.

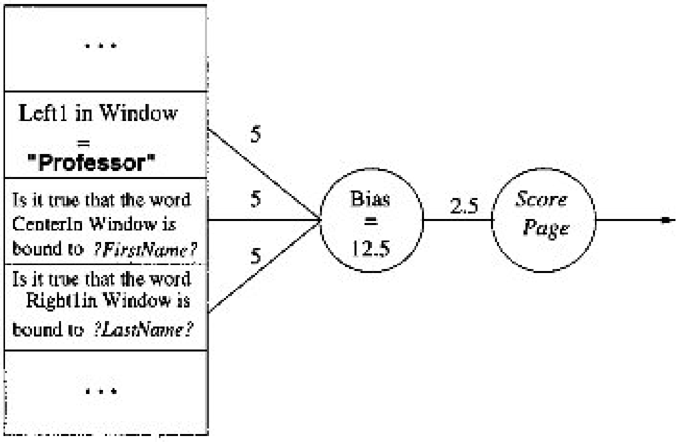
...
Left1 in Window
=
"Professor"
Is it true that the word CenterIn Window is bound to ?FirstName?
Is it true that the word Right1in Window is bound to ?LastName?
...
5
5
5
Bias
=
12.5
2.5
Score
Page

Fig. 5. Mapping advice into ScorePage network.

### A. 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 [61] 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 [62]. 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 $\langle \mathrm{p} \rangle, \langle /\mathrm{p} \rangle, \langle \mathrm{br} \rangle$ 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. For example, Fig. 5 shows how to map into ScorePage a rule like: When "Professor ?Firstname ?Lastname" then show page. If all three parts of the sliding window give a true result, their weighted sum exceeds the bias and produce an activation of the sigmoidal hidden unit near one. Some additional zero-weighted links are also added to this new hidden unit to further allow subsequent learning, as is done in KBNN.

In [63], Chen *et al.* have implemented several search methods in Java based on NNs and genetic search on databases, intranet, and internet. Mercure [64] is another IR system, based on multilayered networks, that allows document retrieval using a spreading activation process and query optimization using relevance backpropagation. This model consists of an input layer, which represents users information needs, a term neuron layer, a document neuron layer, and an output layer representing the result of query evaluation.

Lim [65] has developed a concept of visual keywords which are abstracted and extracted from visual documents using soft computing techniques. Each visual keyword is represented as an NN or a soft cluster center. Merkl and Rauber [66] have shown how to use hierarchical feature maps to represent the contents of a document archive. After producing a map of the document space using self-organizing maps, the system provides a hierarchical view of the underlying data collection in the form of an atlas. Using such a modeling the user can easily "zoom" into particular regions of interest while still having general maps for overall orientation. An ANN-based hybrid web text mining system has been described in [67].

### *B. IE*

Most IE systems that use learning fall into two groups: the one that uses relational learning [68], [40] to learn "extracted patterns," and the other group learns parameters of hidden Markov models (HMMs) and uses them to extract information [69]. In [70] wrapper induction techniques are combined with adaBoost algorithm (Schapire and Singer, 1998) called boosted wrapper induction (BWI) and the system has outperformed many of the relational learners and is competitive with WAWA-IE and HMM. For a brief and comprehensive view of the various learning systems used in web content mining, one may refer to Table I.

### *C. 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 [71], 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. Thematically arranged static document landscapes provide meaningful background for dynamic visualizations of time-related properties of the data, for example. The mathematical preliminaries, background, basic ideas, implications, and numerous applications of self-organizing maps are described in a recent book [72].

### *D. 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 as in [61]. 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" [73]. 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. As described above, the agent is limited to making suggestions about which link to follow from a single page only. This is useful when someone has collected a nearly comprehensive set of links about a topic. A similar system which assists users in browsing software libraries has been built by Drummond [74].

### *E. Prospective Areas of Application*

*Personalized Page Ranking:* As mentioned in Section IV-A, 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. Therefore, it is desirable, for convenience, to get the pages ranked with respect to "relevance" to user queries so that one can get the desired documents only by scanning the first few pages.

Let us consider here again the case of the popular search engine Google [31]. It computes the rank of a page $a$ using

$$\mathrm{Pr}(a) = 1 - d + d\sum_{i=1}^{n} \frac{\mathrm{Pr}(T_i)}{C(T_i)} \tag{6}$$

where $d$ is the damping factor, $\mathrm{Pr}(a)$ is the rank of page $a$ which has pages $T_1, T_2, \ldots, T_n$ pointing to it, and $C(a)$ is the number of outgoing links from page $a$.

Note that it takes into consideration only the popularity of a page (reputation of incoming links) and richness of information content (number of outgoing links) and does not take care of other important factors like:

- User preference: Whether the link matches with the preferences of the user, established from his/ her history?
- Validity: Whether the link is currently valid or not?
- Interestingness: Whether the page is of overall interest to the user or not?

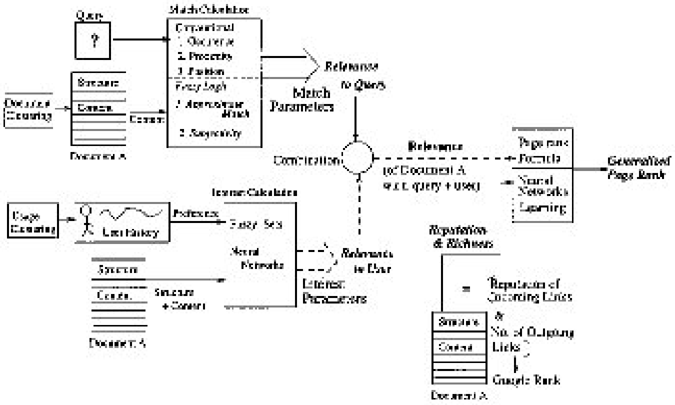

Fig. 6. Neuro fuzzy IR system.

These should also be reflected in the computation of page ranks. The learning/generalization capability of ANNs can be exploited for determining user preference and interestingness. user preference can be incorporated by training an NN-based on user history. Since ANNs can model nonlinear functions and learn from examples, they appear to be a good candidate for computing the "interestingness" of a page. Self organizing NNs can be used to filter out invalid pages dynamically.

An ANN can compute the page rank from a combination of each of the parameters like hub, authority, reputation, validity, interestingness, and user preference with weights assigned to each which the user can modify; thereby refining the network as per his personalized interest. These factors may sometimes also be characterized by fuzzy variables. For example, variables like "close," "far," and "nearby" may be used to represent the distance between hits in a document for a given query. Similarly, fuzzy variables like "reputation" and "importance" can be attached to the URL which is referencing a particular page. This is unlike in Google, where these variables are considered to be crisp. This signifies the importance of integrating synergistically ANN with FL under *neuro-fuzzy paradigm* [75] for computing page ranks. In the next paragraph we describe more details of a proposed neuro-fuzzy IR system.

*Neuro-fuzzy IR:* A schematic diagram of a proposed neuro-fuzzy IR system is shown in Fig. 6. It shows that the total relevance of a document is the combination of "relevance with respect to a query" (match parameters) and "relevance with respect to the user" (interest parameters). This total relevance, when combined with "richness and reputation" of document A (currently reflected in Google rank) give the generalized page rank. The dotted links represent areas not addressed by existing algorithms. Means of computing each of the quantities, namely, "relevance with respect to a query," "relevance with respect to the user," and "richness and reputation," are mentioned below

Relevance to a query: Here FL can be used in computing match parameters by handling subjectivity in queries and approximate matching in documents; thereby better modeling "relevance to query." Also structured documents can be handled more effectively in the framework of FL. Literature described in Section V-A address many of these tasks.

Relevance with respect to a user: Literature in this area is relatively scarce. Existing approaches belong mainly to three categories: 1) learning from user "history" or "profile;" 2) clustering of users into homogeneous groups; and 3) using relevance feedback. ANNs can be used to learn the nonlinear user profiles from their previous history and reflect "relevance to user" of a document A in interest parameters.

Richness and reputation: This parameter is reflected in most existing page ranking systems. However, efficient computation of the page rank is an open research issue where NNs may be used.

*Clustering and Classification:* NNs can be used to classify web pages as well as user patterns, in both supervised and unsupervised modes. Its ability in modeling complex nonlinear functions can also be exploited here.

*Deduction:* Another area where NNs may be used is in building deductive capabilities in web mining systems. As mentioned earlie, complex nonlinear functions may be learned using NNs and logical rules may be extracted from trained networks using rule extraction algorithms. The logical rules are human interpretable and help in generating deductions.

## VII. GAs for Web Mining

GAs, a biologically inspired technology, are 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. 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 [76]. Web document retrieval by genetic learning of importance factors of HTML tags has been described in [77]. Here, the method learns the importance of tags from a training text set. GAs have also been applied for the purpose of feature selection in text mining [78].

*Query Optimization:* In [79], 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.* [80] presented an evolutionary algorithm for query optimization by reweighting the document indexing without query expansion. Kraft *et al.* [81] apply genetic programming in order to improve weighted Boolean query formulation.

*Document Representation:* Gordon [82] adopted a GA to derive better descriptions of documents. Here each document is assigned $N$ descriptions where each description is a set of indexing terms. Then genetic operators and relevance judgements 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 [83].

*Distributed Mining:* Gene expression messy GA (GEMGA) [84] which is a subquadratic 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. The GEMGA is designed based on an alternate perspective of evolutionary computation proposed by the SEARCH framework that emphasizes the role of gene expression or intracellular flow of genetic information. The research on the use of GEMGA in distributed data mining is growing fast and it deals with the problems of finding patterns in data in an environment where both the data and the computational resource are distributed. In [85], Kar Gupta *et al.*, suggest CDM as a new approach toward distributed data mining (DDM) from heterogeneous sites. It is pointed out that naive approaches to distributed data analysis in an heterogeneous environment may lead to an ambiguous situation. CDM processes information locally and builds a global model in situations where it is not possible, as in the case of web, to collect all the data in a common data warehouse and then process.

Regarding the prospective areas of application of GAs, let us consider here the case of *Adaptive Web Sites*. These are sites which automatically improve their organization and presentation by learning from visitor access patterns [86]. It focuses on the problem of *index page synthesis* where an index page is a page consisting of a set of links that cover a particular topic. Its basic approach is to analyze web access logs and find groups of pages that often occur together in user visits (and hence represent coherent topics in users' minds) and to convert them into "index" pages. Here, GAs may be used for prediction of user preferences, dynamic optimization, and evolution of web pages.

## VIII. RSs for Web Mining

RSs are characterized by their ability for granular computation. In RS theory a concept $B$ is described by its "lower" $(\underline{B})$ and "upper" $(B)$ approximations defined with respect to some indiscernibility relation. The use of RS theory for knowledge discovery [87] and rule mining [88], [89] is widely acknowledged. However, the current literature on application of RSs to web mining, like genetic approach, is very scanty. Some web mining tasks where RS theory have been used are mentioned below.

### A. IR

*Granular IR* An important application of RSs is in "granular information retrieval." In RS theory, *Information granules* refer to homogeneous blocks/clusters of documents as described by a particular set of features which may vary over the clusters. The approach is efficient in many document retrieval tasks where the documents are clustered into homogeneous groups before they are retrieved. Documents are often represented by a large number of features or words and dimensionality reduction needs to be performed for clustering. However, instead of representing all the clusters by the same set of features (words), in granular IR using RS theory each cluster is represented by different feature sets. This is closer to reality because different word sets are important for different document classes. Wong *et al.* [90] suggests reducing the dimensionality of terms by constructing a term hierarchy in parallel to a document hierarchy.

*Handling Heterogeneous Data:* RS as well as rough-fuzzy hybrid systems [93] have been used for handling multimedia data and information fusion. A system where RSs have been used for retrieval of multimedia objects is described in [92].

### B. Association/Clustering

RSs have been used for document clustering [93] and mining of web usage patterns [94]. Uses of variable precision RSs [94] and tolerance relations are important in this regard.

Some additional areas where RSs may be applied include:

- WSM: Rough mereology [95], which has ability for handling complex objects, is a potential tool for mining multimedia objects as well as complex representations like web graphs, semantic structures.
- Multiagent systems and collaborative mining.
- Rough-neuro computing (RNC), as a means of computing with words (CWW), is likely to play an important role in natural language query formulation.
- 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. The lower approximation refers to the set which is definitely relevant and the upper approximation denotes the maximal set which may be possibly relevant. Dynamic and focused search, exploiting the above concept, may also help in developing efficient IR systems.

## 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. Some attempts in this direction are described in [96] and [97].
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.
5) Some tasks related to embedded internet systems that could be handled using soft computing include, access control, task scheduling, system configuration, priority order, device monitoring, bug and failure reporting, as well as distributed (remote) control of electronic products (devices).
6) E-commerce is an important application area of soft computing. It may be used to impart human like interaction in e-shopping portals [98]. For example, buyers' interest can be better modeled (as functions of price and quality) using fuzzy set theory.
7) The development of new knowledge visualization techniques for effective user interface may also be done with soft computing.

Finally, (soft) case-based reasoning (CBR) [99], which is a popular AI problem solving paradigm, using soft computing tools, and is recently drawing the attention of researchers worldwide, may be used for solving many of the web mining problems, stated above. In this context, the mention may be made of the computational theory of perception, recently explained by Zadeh [52], which is characterized mainly using the concept of "fuzzy-granularity" of perceptions.

## References

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

[2] R. Kosla and H. Blockeel, "Web mining research a survey," *SIG KDD Explorations*, vol. 2, pp. 1–15, July 2000.

[3] O. Etzioni, "Moving up the information food cahin: Deploying softbots on the web," in *Proc. 14th Nat. Conf. AI*, Portland, OR, 1996, pp. 1322–1326.

[4] 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.

[5] W. F. Punch and M. R. Wulfekuhler, "Finding salient features for personal web page categorization," *Comput. Networks ISDN Syst.*, vol. 29, pp. 1147–1156, 1997.

[6] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," in *Proc. 6th Int. WWW Conf.*, 1997, pp. 391–404.

[7] Y. S. Mareek and I. Z. Benshaul, "Automatically organizing bookmarks per content," in *Proc. 5th Int. WWW Conf.*, 1996.

[8] R. Baeza-Yates and B. Ribiero-Neto, *Modern Information Retrieval*. Reading, MA: Addison-Wesley Longman, 1999.

[9] M. Kobayashi and K. Takeda, "Information retrieval on the web," *ACM Comput. Surveys*, vol. 32, no. 2, pp. 144–173, 2000.

[10] N. Kushmerick, "Gleaning the web," *IEEE Intell. Syst.*, vol. 14, no. 2, pp. 20–22, 1999.

[11] D. Freitag, "Information extraction from html: Application of a general machine learning approach," in *Proc. 15th Conf. Artificial Intell. AAAAI-98*, 1998, pp. 517–523.

[12] C. M. Brown, D. Hardy, B. B. Danzig, U. Manber, and M. F. Schwartz, "The harvest information discovery and access system," in *Proc. 2nd Int. WWW Conf. Distributed Environments*, 1994, pp. 763–771.

[13] K. Hammond, R. Burke, C. Martin, and S. Lytinen, "Faq-finder: A case based approach to knowledge navigation," presented at the Working Notes of AAAI Spring Symposium on Information Gathering From Heterogeneous Distributed Environments, Stanford, CA, 1995.

[14] A. Y. Levy, T. Kirk, and Y. Sagiv, "The information manifold," presented at the AAAI Spring Symposium on Information Gathering From Heterogeneous Distributed Environments, 1995.

[15] C. Kwok and D. Weld, "Planning to gather information," in *Proc. 14th Nat. Conf. AI*, 1996.

[16] E. Spertus, "Parasite: Mining structural information on the web," presented at the Proc. 6th WWW Conf., 1997.

[17] O. Etzioni, D. S. Weld, and R. B. Doorenbos, "A Scalable Comparison Shopping Agent for the World Wide Web," Univ. Washington, Dept. Comput. Sci., Seattle, Tech. Rep. TR 96-01-03, 1996.

[18] O. Etzioni and M. Perkowitz, "Category translation: Learning to understand information on the internet," in *Proc. 15th Int. Joint Conf. Artificial Intell*, Montreal, QC, Canada, 1995, pp. 930–936.

[19] M. Craven, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, and D. DiPasquo, "Learning to extract symbolic knowledge from the world wide web," in *Proc. 15th Nat. Conf. AI (AAAI98)*, 1998, pp. 509–516.

[20] O. Etzioni, J. Shakes, and M. Langheinrich, "Ahoy! the homepage finder," presented at the Proc. 6th WWW Conf., Santa Carla, CA, Apr. 1997.

[21] D. D. Cutting, J. Karger, J. Pederson, and J. Scatter, "A cluster based approach to browsing large document collections," in *Proc. 15th Int. Conf. Res. Development Inform. Retrieval*, June 1992, pp. 318–329.

[22] J. Hipp, U. Guntzer, and J. Nakhaeizadeh, "Algorithms for assocation rule mining a general survey and comparison," *ACM SIGKDD Explorations*, vol. 2, pp. 58–65, July 2000.

[23] J. Pitkow, "In search of reliable usage data on the www," in *Proc. 6th Int. WWW Conf.*, Santa Carla, CA, 1997, pp. 451–463.

[24] B. Mobasher, N. Jain, E.-H. Han, and J. Srivastava, "Web Mining: Patterns from WWW Transactions," Dept. Comput. Sci., Univ. Minnesota, Tech. Rep. TR96-050, Mar. 1997.

[25] W. W. Cohen, "What can we learn from the web?," in *Proc. 16th Int. Conf. Machine Learning (ICML99)*, 1995, pp. 515–521.

[26] D. Mladenic and M. Grobelnik. Efficient text categorization. presented at Proc. Text Mining Workshop 10th European Conf. Machine Learning ECML98. [Online]http://www=ai.ijs.si/DunjaMladenic/papers/PWW/pwwWsheEMCL99.ps.gz

[27] R. Ghani, R. Jones, D. Mladenic, K. Nigam, and S. Slattery, "Data mining on symbolic knowledge extracted from the web," in *Proc. 6th Int. Conf. Knowledge Discovery Data Mining (KDD-2000) Workshop on Text Mining*, Boston, MA, Aug. 2000, pp. 29–36.

[28] S. Loh, L. K. Wives, and J. P. M. de, "Concept based knowledge discovery from texts extracted from the web," *ACM SIGKDD Explorations*, vol. 2, pp. 29–40, July 2000.

[29] S. Chakrabarti, "Data mining for hypertext," *ACM SIGKDD Explorations*, vol. 1, no. 2, pp. 1–11, 2000.

[30] *IEEE Computer (Special Issue on Digital Libraries)*, vol. 32, 1999.

[31] 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.

[32] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," *Proc. 9th IEEE Int. Conf. Tools with Artificial Intelligence*, pp. 558–567, Nov. 1997.

[33] A. Levy and D. Weld, "Intelligent internet systems," *Artificial Intell.*, vol. 118, no. 1–2, 2000.

[34] M. Craven and J. Shavlik, "Using neural networks for data mining," *Future Generation Comput. Syst. (Special Issue on Data Mining)*, vol. 13, pp. 211–229, 1998.

[35] F. Crimmins, A. Smeaton, T. Dkaki, and J. Mothe, "Information discovery on the internet," *IEEE Intell. Syst.*, vol. 14, pp. 55–62, 1999.

[36] J. Furnkranz, "Exploiting structural information for text classification on the WWW," in *Proc. Advances Intell. Data Anal. 3rd Int. Symp., IDA99*, 1999, pp. 487–498.

[37] T. Mitchell, D. Freitag, and T. Joachims, "Webwatcher: A tour guide for the world wide web," in *Proc. Int. Joint Conf. AIIJCAI97*, 1997, pp. 770–777.

[38] I. Muslea, S. Minton, and C. Knoblock, "Hiererchical wrapper induction for semistructured information sources," *J. Autonomous Agents Multiagent Syst.*, vol. 4, pp. 93–114, 2001.

[39] L. Singh, B. Chen, R. Haight, P. Scheu, I. Muslea, S. Minton, and C. Knoblock, "Wrapper induction for semistructured web based information sources," in *Proc. 2nd Int. Conf. KDD Data Mining*, 1998, pp. 329–333.

[40] 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.

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

[42] 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.

[43] 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 (WEBKDD'00). [Online]http://robotiocs.stanford.edu/~ronnyk/WEBDD2000/papers/alvarez.pdf

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

[45] 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.

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

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

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

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

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

[51] 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.

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

[53] R. Yager, "A framework for linguistic and hierarchical queries for document retrieval," in *Soft Computing in Information Retrieval: Techniques and Applications*, F. Crestani and G. Pasi, Eds, Heidelberg: Physica-Verlag, 2000, vol. 50, pp. 3–20.

[54] T. Gedeon and L. Koczy, "A model of intelligent information retrieval using fuzzy tolerance relations based on hierarchical co-occurrence of words," in *Soft Computing in Information Retrieval: Techniques and Applications*, F. Crestani and G. Pasi, Eds. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50, pp. 48–74.

[55] G. Pasi and G. Bordonga, "Application of fuzzy set theory to extend boolean 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. 21–47.

[56] R. Krishnapuram, A. Joshi, and L. Yi, "A fuzzy relative of the $k$-medoids algorithm with application to document and snippet clustering," *Proc. IEEE Int. Conf. Fuzzy Syst.*, 1999.

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

[58] B. Mobasher, V. Kumar, and E. H. Han, *Clustering in a High Dimensional Space Using Hypergraph Models*. Minneapolis: Univ. Minnesota, 1997, Tech. Rep. TR-97-063.

[59] A. Gyenesei, "A Fuzzy Approach for Mining Quantitative Association Rules," Univ. Turku, Dept. Comput. Sci., Lemminkisenkatu 14, Finland, TUCS Tech. Rep. 336, Mar. 2000.

[60] D. Nauck, "Using symbolic data in neuro-fuzy classification," in *Proc. NAFIPS'99*, New York, June 1999, pp. 536–540.

[61] 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.

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

[63] 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.

[64] M. Boughanem, T. Dkaki, J. Mothe, and C. Soule-Dupuy, "Mercure at trec7," presented at the Proc. 7th Int. Conf. Text Retrieval, TREC7, Gaithersburg, MD, 1998, pp. —355–360.

[65] J. H. Lim, "Visual keywords: From text retrieval to multimedia retrieval," in *Soft Computing in Information Retrieval: Techniques and Applications*, F. Crestani and G. Pasi, Eds. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50, pp. 77–101.

[66] D. Merkl and A. Rauber, "Document classification with unsupervised artificial neural networks," 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.

[67] H. Fukuda, E. Passos, A. M. Pacheco, L. B. Neto, J. Valerio, V. J. D. Roberto, E. R. Antonio, and L. Chigener, "Web text mining using a hybrid system," in *Proc. 6th Brazilian Symp. Neural Networks*, 2000, pp. 131–136.

[68] 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.

[69] D. Bikel, R. Schwartz, and R. Weischedel, "An algorithm that learns what's in a name," *Machine Learning (Special Issue on Natural Language Learning)*, vol. 34, no. 1/3, pp. 211–231, 1999.

[70] D. Freitag and N. Kushmerick, "Boosted wrapper induction," in *Proc. AAAI*, 2000, pp. 577–583.

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

[72] ——, *Self-Organizing Maps*, 2nd ed, Berlin, Germany: Springer-Verlag, 1997.

[73] M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill and webert: Identifying interesting web sites," in *Proc. 13th Nat. Conf. AI*, 1996, pp. 54–61.

[74] C. Drummond, D. Ionescu, and R. Holte, "A Learning Agent That Assists the Browsing of Software Libraries," Univ. Ottawa, Ottawa, ON, Canada, Tech. Rep. TR-95-12, 1995.

[75] S. Mitra and S. K. Pal, "Fuzzy multi-layer perceptron, inferencing and rule generation," *IEEE Trans. Neural Networks*, vol. 6, pp. 51–63, Jan. 1995.

[76] F. Crestani and G. Pasi, Eds., *Soft Computing in Information Retrieval: Techniques and Application*. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50.

[77] 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, Aug. 2000, pp. 13–23.

[78] M. Martin-Bautista and M. A. Vila, "A survey of genetic feature selection in mining issues," in *Proc. Congr. Evol. Comput. (CEC99)*, 1999, pp. 13–23.

[79] 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.

[80] 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.

[81] 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.

[82] M. D. Gordon, "Probabilistic and genetic algorithms for document retrieval," *Commun. ACM*, vol. 31, no. 10, pp. 208–218, 1988.

[83] 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.

[84] H. Kargupta, "The gene expression messy genetic algorithm," *Proc. IEEE Int. Conf. Evol. Comput.*, pp. 631–636, 1996.

[85] H. Kargupta, B. H. Park, D. Hershberger, and E. Johnson, "Collective data mining: A new perspective toward distributed data mining," in *Advances in Distributed and Parallel Knowledge Discovery*. Cambridge, MA: MIT/AAAI Press, 1999.

[86] O. Etzioni and M. Perkowitz, "Adaptive web sites: An AI challenge," presented at the Proc. 15th Int. Joint Conf. Artificial Intell. (IJCAI'97), Nagoya, Japan, July 1997, pp. 16–23.

[87] A. Skowron and L. Polkowski, Eds., *Rough Sets in Knowledge Discovery*. Heidelberg, Germany: Physica-Verlag, 1998.

[88] S. K. Pal, S. Mitra, and P. Mitra, "Rough fuzzy MLP: Modular evolution, rule generation and evaluation," IEEE Trans. Knowledge Data Eng., 2002, to be published.

[89] M. Banerjee, S. Mitra, and S. K. Pal, "Rough fuzzy MLP: Knowledge encoding and classification," *IEEE Trans. Neural Networks*, vol. 9, pp. 1203–1216, 1998.

[90] S. K. Wong, Y. Y. Yao, and C. J. Butz, "Granular 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. 317–331.

[91] S. K. Pal and A. Skowron, *Rough Fuzzy Hybridization: A New Trend in Decision Making*, Singapore: Springer-Verlag, 1999.

[92] U. Straccia, "A framework for the retrieval of multimedia objects based on four-valued fuzzy description logics," in *Soft Computing in Information Retrieval: Techniques and Applications*, F. Crestani and G. Pasi, Eds. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50, pp. 332–357.

[93] S. Kawasaki, N. B. Nguyen, and T. B. Ho. Hierarchical document clustering based on tolerance rough set model. presented at Proc. 6th Int. Conf. Knowledge Discovery Data Mining (KDD-2000) Workshop Text Mining. [Online]http://www-2.cs.cmu.edu/~dunja/KDDpapers/Bao-TM.ps

[94] 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, . .

[95] L. Polkowski and A. Skowron, "Rough mereology: A new paradigm for approximate reasoning," *Int. J. Approximate Reasoning*, vol. 15, no. 4, pp. 333–365, 1996.

[96] C. Wan, M. Liu, and L. Wang, "Content-based sound retrieval for web application," in *Web Intelligence: Research and Development*, N. Zhong, Y. Yao, J. Liu, and S. Oshuga, Eds, Singapore: Springer-Verlag, 2001, vol. LNCS 2198, pp. 389–393.

[97] K. Yanai, M. Shindo, and K. Noshita, "A fast image-gathering system on the world wide web using a PC cluster," in *Web Intelligence: Research and Development*, N. Zhong, Y. Yao, J. Liu, and S. Oshuga, Eds, Singapore: Springer-Verlag, 2001, vol. LNCS 2198, pp. 324–334.

[98] C.-H. Lee and H.-C. Yang, "Developing an adaptive search engine for e-commerce using a web mining approach," in *Proc. Int. Conf. Inform. Technol.: Coding and Computing*, 2001, pp. 604–608.

[99] S. K. Pal, T. S. Dillon, and D. S. Yeung, Eds., *Soft Computing in Case Based Reasoning*, Singapore: Springer-Verlag, 2001.

**Sankar K. Pal** (M'81–SM'84–F'93) received the M.Tech. and Ph.D. degrees in radio physics and electronics in 1974 and 1979, respectively, from the University of Calcutta, Calcutta, India. He received the Ph.D. degree in electrical engineering and the DIC degree from Imperial College, University of London, London, U.K., in 1982.

He is a Professor and Distinguished Scientist at the Indian Statistical Institute, Calcutta. He is also the Founding Head of the Machine Intelligence Unit there. He was with the University of California, Berkeley, and the University of Maryland, College Park, from 1986 to 1987 as a Fulbright Postdoctoral Visiting Fellow; at the NASA Johnson Space Center, Houston, TX, from 1990 to 1992 and 1994 as a Guest Investigator under the NRC-NASA Senior Research Associateship Program; and at the Hong Kong Polytechnic University, Hong Kong, in 1999 and 2000 as a Visiting Professor. He served as a Distinguished Visitor of IEEE Computer Society (USA) for the Asia-Pacific Region during 1997 to 1999. His research interests include pattern recognition, image processing, data mining, soft computing, neural nets, genetic algorithms, and fuzzy systems. He is a co-author/co-editor of ten books including *Fuzzy Mathematical Approach to Pattern Recognition* (New York: Wiley, 1986), *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing* (New York: Wiley, 1999) and has about 300 research publications.

Dr. Pal is a Fellow of the Third World Academy of Sciences, Italy, the International Association of Pattern Recognition, and all the four National Academies for Science/Engineering in India. He has received the 1990 S. S. Bhatnagar Prize, the 1993 Jawaharlal Nehru Fellowship, the 1993 Vikram Sarabhai Research Award, 1993 NASA Tech Brief Award, the 1994 IEEE Transactions on Neural Networks Outstanding Paper Award, the 1995 NASA Patent Application Award, the 1997 IETE—Ram Lal Wadhwa Gold Medal, the 1998 Om Bhasin Foundation Award, the 1999 G. D. Birla Award for Scientific Research, the 2000 Khwarizmi International Award (first winner) from the Islamic Republic of Iran, the 2001 Syed Husain Zaheer Medal from Indian National Science Academy, and the 2001 FICCI Award for Engineering and Technology from the Federation of Indian Chamber of Commerce and Industries. He was an Associate Editor, the IEEE Transactions on Neural Networks from 1994 to 1998, and is currently Editor of the IEEE IEEE Transactions on Pattern Analysis and Machine Intelligence, *Pattern Recognition Letters*, the *International Journal of Pattern Recognition and Artificial Intelligence*, *Neurocomputing*, *Applied Intelligence*, *Information Sciences*, *Fuzzy Sets and Systems*, and *Fundamenta Informaticae*; a Member of the Executive Advisory Editorial Board for the IEEE Transactions on Fuzzy Systems, the *International Journal on Image and Graphics*, and the *International Journal of Approximate Reasoning*; and a Guest Editor of many publications, including the IEEE Computer.

**Varun Talwar** (S'01) received the Bachelor's degree in computer engineering at Netaji Subhas Institute of Technology, University of Delhi, Delhi, India, in 2002. He is currently pursuing the M.S. degree in computer science with the Singapore-Massachusetts Institute of Technology (MIT) Alliance at the National University of Singapore, Singapore.

This paper was written while he was working on the project on soft computing in web mining as a Research Trainee at Machine Intelligence Unit, Indian Statistical Institute, Calcutta, India. His primary research interests are related to data and web mining.

**Pabitra Mitra** (S'99) received the B.Tech. degree in electrical engineering from Indian Insitute of Technology, Kharagpur, India, in 1996.

He was a Scientist with the Center for Artificial Intelligence and Robotics, Bangalore, India. Currently, he is a Senior Research Fellow of Indian Statistical Institute, Calcuuta. His research interests are in the area of data mining and knowledge discovery, pattern recognition, learning theory, and soft computing.