2012 12th International Conference on Computational Science and Its Applications

A Survey: Soft computing in Intelligent Information Retrieval Systems

Mohd Wazih Ahmad
Research Scholor
Department of Computer Engineering
JMI New Delhi,India
Mail.java@yahoo.com

Dr. M A. Ansari
Deaprtment of Electrical Engineering,
School of Engineering,GBU Greater Noida,UP,India

***Abstract*— this paper provides an in-depth survey of challenges in the design of intelligent information retrieval systems, pointing out some similarities and differences in the core data mining and web based search operations. The procedures for evaluation of search engine performance and implicit feedback of the user with respect to a search result are studied with references to the different algorithms. We have proposed a novel neural satisfaction based feedback vector in contrast to the existing activity pattern based feedback as a future research direction in Intelligent IR. We addressed the select research work in the area of soft information retrieval using fuzzy sets, artificial neural networks, genetic algorithms and probabilistic information retrieval. As an instance of information retrieval, web mining is a set of operations to retrieve relevant documents from preprocessed, crawled and indexed web, and it can be categorized into more specialized tasks of web content mining, web structure mining and web usage mining. We have given a survey of the important reviews on topic of web mining and its associated tasks. On the basis of Identified challenges in information retrieval in general, and web mining in particular, we have concentrated on applicability of soft computing techniques and their hybrids in web mining, the performance related issues of select solutions, future of web mining and the next generation user's expectations from a search engine.**

***Index Terms*— Information retrieval survey, soft computing intelligent information retrieval.**

I. INTRODUCTION

The performance of an information retrieval system depends on a set of design time and run time factors and in literature performance of IR has been computed in terms of efficiency, effectiveness, quality of search and the user satisfaction with respect to a search result. Efficiency of the IR system is a design parameter, while effectiveness is a measure of usefulness of the system. One measure of effectiveness is precision and recall. In fact, efficiency and effectiveness of an IR system represents aggregate performance measure of individual operations involved in overall process. These operations are text preprocessing, document representation, indexing, searching and matching. But the quality of search result for a user query and the user satisfaction is determined by an exhaustive list of internal and external factors, which we have discussed in this paper. In addition to IR design issues we have also included the discussion on interaction based factors which affects the performance of a search engine. Performance of IR system is highly influenced by discriminative vocabulary size in corpus, domain of knowledge, user model and language model to be adopted, treatment of word sense ambiguities, synonymy and other linguistic morphology, ranking and rank aggregation techniques by which result lists of different search engines are aggregated into one final list. Some user specific issues are for example navigational skills of the users, definition of relevance in user context, way of presenting the information to the user by a search engine, user interface , whether the search engine interactively utilize the feedback from user to optimize future search, history of previous sessions between the current user and IR system etc. in a production environment there are more often, behind the scene factors which ultimately bring the user to a satisfaction or dissatisfaction level for a given search engine. Commercial search engines even consider more specific issues in addressing the need of their users. Expected design of a search engine is that which guess the user requirements, context and the topic of search from his or her behavior, if not in advance, than immediately after providing the result list. This can be implemented by implicitly observing the interaction pattern of user with a result list [6].implicit feedback mechanism is better than an explicit request for feedback because of the fact that in many situations users don't like to fill explicit feedback form designated for search evaluation. User satisfaction is a subjective measure of quality of search result. It can be computed through user activities and tracking the cognitive

978-0-7695-4710-7/12 $26.00 © 2012 IEEE
DOI 10.1109/ICCSA.2012.15

states of a user during information retrieval. Instead of merely focusing on the activity pattern based feedback vector, a breed of next generation search engines are aimed to catch the neuro-satisfaction of user by implementing more specific feedback mechanisms, like eyeball tracing, face expressions, mouse movements etc.

This Text focuses on the survey of Soft information retrieval, the design challenges, performance issues in web mining, and the directives for future research in web mining.

II. Web Based Information Retrieval

The web is a huge and fast growing collection of publicly available information over the internet. Searching the web for relevant information is becoming difficult as the new public and private networks, hardware devices and information formats being available at internet and the volume of data is growing every day. Deep web, privacy issues during crawling, identification and indexing of dynamic web pages are major concerns that the web search engines are considering as immediate challenges. Another dimension of the design challenges for a search engine is the interaction with novice users. Novice users are not trained for formal search engine query interface, unlike experts; they are unable to formulate exact information needs in the formal query language of search engine. This may happen because either they are not familiar with the formal query language itself or are unable to express the information needs in the query language. For novice users, a search engine provides a natural language query interface. A natural language query is capable to address inherent uncertainty in user queries. A reason of the uncertainty in a user query is the fact that human thinks and communicates in vague language. Natural languages queries regarding text, image and other information needs are infinitely capable to represent human thought process, uncertainty, incompleteness and imprecision, increasing the complexity of IR problem manifolds. The needs to handle the uncertainty in representation have resulted in approximated queries using weighted terms and fuzzy linguistic terms. Different types of document representation/indexing and explicit thesauruses are used to deal with problem of approximate search and fuzzy matching algorithms applied in the history of IR.

Recent developments in interdisciplinary research have proved the observation that the computational complexity of uncertain and vague processes is best handled by the integration of soft computing techniques and their hybrid schemes in finding the solution. In the past a variety of engineering applications including the information retrieval in specific domains, control systems, decision making in management sciences, medicine, industrial systems and products etc. have benefited with soft computing techniques like ANN, FL, GA and RS and significant improvement in the performance and capabilities of systems are reported, this has motivated our survey in the filed of applications of soft computing in web mining. In particular, fuzzy sets are useful in representing the neural satisfaction vector and the feedback of a user with a search engine, because these quantities are uncertain to determine at any point of time and the crisp representation of these quantities is inappropriate to deal with the underlying uncertainty.

Intelligent information retrieval deals with the uncertainty issues present in the underlying decision process. For example whether a document satisfies a fuzzy description of user requirement? If yes, than up to what degree? What is the answer of different search engines for a particular query? How to combine these opinions in an optimum way to give a final answer? We consider it an instance of generalized multi-objective optimization of list aggregation, constrained by the user requirements and neural satisfaction related feedback abstraction. Finding the best solution meant to find optimized resultant list of documents which includes all or a proportion of the given documents, and displayed list maximizes the gain in terms of user neural satisfaction as an objective function. Motivations to the research are subjective evaluation of result set on the basis of user feed back, Q o S parameters of search engines and next generation multimedia retrieval. Neural satisfaction driven optimization of information retrieval process is the continued future research direction rather than only focusing on the recall and precision curves in isolation. Soft computing techniques like Fuzzy logic, ANN, GA, Neuro -fuzzy systems, probabilistic fuzzy logic, fuzzy pattern matching are capable to address the issues of uncertainness of representation, specification and querying, uncertainty in the process of searching and matching at different levels of abstractions. Architectural design of information retrieval system is application specific unlike relational, object-oriented or any other generic database models which are applicable to a class of applications by providing a generalized and reusable framework for defining, updating, querying and managing the database in a DBMS. But actual success of an information retrieval system based on proposed model will depend the domain and complexity of the information, semantics, available knowledge about user and expertise of the user in posing queries to the system, dynamics with which user changes the queries, kind of documents (Text, Web documents, images), size of collection and the nature of search space in terms of update frequency, proportion of the non relevant documents to a query present in corpus, the dimensionality of availability of a documents in search space as well as the spam pages which are intentionally injected into the search results by following the best practices of black hat search engine optimization communities. Future of brain mapped information retrieval is highly dependent on how the above issues are addressed with respect to the neural satisfaction vector. In order to provide a self contained view of the current research work in the field of intelligent information retrieval, we have virtually divided this survey paper into multiple self contained parts, including introductory notes on current challenges in IR field and some Search Quality Measuring techniques in section I and II, we have described the important soft computing techniques currently published to be used in intelligent information retrieval systems. Some traditional and soft computing techniques are well practiced and can be assumed as standards for today’s research work and we have less concentration on referencing those. In second half of this paper we have described the significant work in the field of fuzzy, ANN, GA, probabilistic algorithms in web mining, summarizing some recent publication in this field. Rest of the paper is organized as follows: in section III a review of the challenges in IR process is described and one important survey is discussed, and section III (B) we have discussed SQM and major research in the field of quality measurement,

III. A Review Of Challenges In IR:

A. Major Research challenges in IR

Formally, information retrieval is defined as a problem of mapping an imprecise user query to the given set of documents in such a way that the result set contain the list of relevant documents in their sorted order of relevance with respect to the information requirement. Understanding the user expectations from a right information retrieval system in a given context is a prime challenge for researchers. What is going to make different an Intelligent IR system from a database search system? Is it only the query language which is different in two? Or the model used in searching the information is different in the two cases? Is it the difference of design of systems or the two are meant to address totally different problems of real world? Answer lies in the in-depth understanding of meaning of each of nearby problems and survey of context related solutions. This text includes the important work of researchers to identify the research challenges and classify these IR challenges for the ease of handling the complexity of design in the proposed solutions.

IR deals with organization, storage, matching and extraction of information from a large set of documents. The complexity of information retrieval task has derived the research to be performed in specific modules each addressing some aspects of the IR task. Subjective measures of user satisfaction are used for comparative performance evaluation in literature of IR to judge the supremacy of one model over other, given a common test data. In this paper we have summarized the IR challenges which need attention for insuring quality of search in information retrieval. The main Challenges in information retrieval (reported in the previous papers) are related to Query representation from vague user specification, document representation, classification, clustering and indexing, matching, ranking the result by relevance, language modeling, user feedback and performance measures in terms of precision and recall. Search engine must know importance of user feedback and device new techniques for implicit feedback rather than an explicit feedback form. User satisfaction, search quality measure and utility of result set are subjective in nature. We have introduced neural level user satisfaction in this paper as a future research challenge because modeling the effect of some information on human brain is more difficult to implement than an activity based feedback mechanism.

A description of various information retrieval challenges in given in [1] as well as in [2] .J. Allen et. al in [1] further extended the definition of information retrieval from the work of Salton[3] in 1960's that says:

"*Information retrieval is defined as a field concerned with structure, analysis, organization, storage searching and retrieval of information*".

According to [1], due to rapid growth of Internet and availability of different types of data on the web, the information retrieval is not only limited to the document retrieval but it has grown into much larger sphere and now research advances includes following hot keywords in their addresses:

"*Text retrieval, question answering, topic detection and tracking, summarization multimedia retrieval (i. g. image, video and music), software engineering and chemical and biological informatics, text structuring, text mining and genomics…."*

As the internationalization support in the software application is increasing, integration of the local information retrieval systems into meta-search engines to compute rank aggregation has also gained importance. The urgent need to handle the complexity of information access over internet, heterogeneity, variability in available information and changing information needs of different users and groups have lead the fast development of IR filed in 1960-80 era and maximum algorithms which used at that time are IR standards today. Internationalization has facilitated the research in the field of multilingual, distributed, heterogeneous, collaborative and context sensitive retrieval models of information. An important shift in computing paradigm has already taken place from increasing computational capability of machine architecture from hardware point of view to the increasing computational complexity that can be handled in the form of logical and intelligent procedures for next generation information retrieval on general purpose hardware clients. For example the chemical information retrieval systems which deals with the different chemical formulas, chemical representations, isotopes, rich information about the chemical changes, formation of intermediate compounds during a reaction, set of generalized chemical groups to a core compound, orientations of groups in space are real challenges, these graphical representations and chemical reactions can not be mimicked by specialized hardware and the complexity issues has to be still addressed by a logical framework at higher level of abstraction.

An excellent report on the challenges in information retrieval and language modeling by [1] has given a list of long term and near term challenges in information retrieval, language modeling and resource models etc. it has provided a list of next generation key research focuses in the different fields of IR. Global information retrieval and contextual information are identified as two long term challenges, along with the intelligent classification algorithms and automatic text summarization from semi structured data. In order to achieve the solution of these problems [1] has pointed out the need of collaborative solutions involving the possible fusion of the IR task with probability, Natural language processing, machine learning and other areas including database management systems, sensor based streams and distributed ,heterogeneous information systems. Near term issues are those which need to be handled within a time of five years and are generalization of specific IR models incorporating different forms of media from multiple sources, identification of more refined taxonomies and novel performance evaluation techniques for complex IR systems for comparative evaluation. other near term challenges under the hood of cross lingual information retrieval includes the modeling of the effective user functionality, new and more complex applications, handling languages with sparse data, learning semantic relationship from parallel and comparable corpora, merging the retrieval results, more tightly integrated retrieval models for the CLIR, accommodating more training data. Under the resource requirements there is a need of benchmarking, evaluation methods, TestBeds, and data sets like TREC, CLEF, and NII-NACSIS etc. in fact guided research is possible in IR only if the current flaws and shortcomings, findings and progress are kept in communication, in the form of standard data sets and conclusions are made available, communicated and discussed for each field in IR. Nicolas J. Belkin et.al. has proposed a new method for interactive information retrieval systems, in this work he has proposed a model for evaluation of performance of an IR system based on usefulness of result set. Unlike CRANFIELD/TREC performance evaluation on the basis of document relevance, this research work give more effective performance measure like implicit or explicit user feedback, utility and satisfaction (3), catching eye gaze focus of user while looking the result set, feedback from facial expressions, click through information [4]. However this work is written in context of Enterprise information retrieval systems, the information retrieval in enterprise search engines is quiet different from web search and can be recognized as a specific case of web mining with restricted scale, size and domain of information, limited number of categories of documents in the collection, more specific queries, absence of spam pages in database, more structured data and requirement of time-efficient system due to temporal constraints[3].

The major IR challenges and issues are issues discussed in [1] are identified in the categories of web search challenges, user modeling issues, filtering, TDT and classification, summarization and question answering issues, meta-search and distributed search issues, multimedia information retrieval and the information extraction specific issues. In this survey we have addressed these points in context of other important research work in this field.

A search engine assume web as a connected graph of nodes where nodes represents the web pages and arcs between nodes represents the hyper-links among web pages. Web search issues are related to web structure, crawling and indexing, searching, data collection and evaluation [1, 2].in his work in [1] users modeling issues and near term challenges are described as shared data-gathering environment, TestBed of IR interactions[1 4], evaluation successfully incorporating the user, privacy, extended user models, long term user models. The fields of Automatic disambiguation and the focused query expansion using the probabilistic models of user expectation is a potential solution for challenges in web search. User modeling in web search requires different kinds of approximations and adjustments in query, user profiles on the basis of past interaction of user with web and NLP issues. These approximations are discussed in (5); effect of these approximations however is to be analyzed in context of practical IR systems and their applications. Next few paragraphs address the valuable findings of [1] in contrast to [2] and others.

Classification is a task of assigning the predefined labels to the group of similar concepts. Advances in classification require the in depth understanding of procedures of finding the informal structured hierarchical taxonomies, leveraging unlabeled data, pseudo feedback, co training, active learning and transduction. Other issues in classification include handling semi-structured data and novelty detection in the given data [1]. Summarization issues are discussed in contrast with information extraction, lack of standards in summarization need to be addressed in short term separately for heading length summarization, topical summarization and summaries for news paper and news-wires. Models and summarization process based on the prior knowledge of user is also expected to drive the accuracy and of interactively generated summaries [1]. Result of good summarization can be used for automatic fuzzy thesauri generation.

With increasing dependency on the web for all kind of information, requirement of Natural language based question answering systems has increased and research challenges in the design of the QA systems includes: implementation of short or factual answers, passage and multi passage summaries, answers that involves complex reasoning [1] and can understand the user intention from a vague NLP statement [2]. Short term issues should be focused on providing a rich set of factoid as well as support to long answers, using dynamic resources with increased transparency.

Meta-search algorithms are based on distributed search, exploiting the link structure of web and the ranked lists of the documents from different search engines a meta-search engine provides aggregation of result lists (full lists or partial lists); rank aggregation is the process of merging the results of different search engines in a concise form and presenting the result set to the user [7]. Major challenges are description of resources over internet using a standard resource description framework, selection of resources on the basis of query, combining the individual ranks of resources from different search engines and than presenting user with aggregate list. Additionally, score normalization, ad hoc query evaluation, unsupervised learning, performance of distributed IR are important points to be addressed in near term before the cost of organization get impractical to achieve.

Nicolas J. Belkin in [2] has given emphasis on the fact that how a “user” views an information retrieval system and how he or she interacts with it, which unfortunately has been untouched in many publications of its time. This paper has listed a number of references and issues to depict importance of user in the process of information retrieval, scope and performance of interactive information retrieval, this paper revised a number of challenges for it. Discussion includes the introduction of more or less formal models for interactive information retrieval (language modeling), smoothing operation for refining the proposed models, implementation, and performance evaluation in terms of effectiveness.

B. Search Quality Measures: Effectiveness of IR System

The current web search challenges include the measurement and control of the web search quality, maximize the user satisfaction by presenting a list of highly relevant documents, minimizing the spam pages in the result lists, iteratively refined and guided result lists, driven by the user feedback for integrated information retrieval task with supports to audio, video, graphics and other elements of multimedia retrieval. According to the famous principle of control theory, anything which can be measured correctly can be controlled easily, therefore in order to control the quality and usefulness of a web search (result lists etc.) future implementations of web search techniques must be optimized for learning the past interaction of user with result lists of different queries aggregated in one optimum set of results.

In [7] there is an excellent application of soft computing techniques for computing the aggregated rank of a document while a list of document-ranks is given for each search engine. In [6] it is a very first successful attempt to address the quality of web search result after list aggregation in a quantitiave manner. In this paper the search quality of a result list is compared by the spearman's rank order correlation coefficient which is the numeric closeness of two rank lists. This research have modeled the user satisfaction in the form of a vector (V, T, P, S, B, E, C).this vector implicitly takes user feedback on the basis of the sequence in which a document is visited by user, time the user spend on a document, whether user takes a printout of the document, saves, bookmarks, email or copy some portion of a particular document from the given result set. Once this statistics is available, a numeric count of Search Quality measure is calculated which is represented as a seven dimensional normalized vector of above described parameters used to compute SQM, and update the weights of documents in the result list of current interaction so that in next display the document which were irrelevant are penalized in terms of their ranks and those which are relevant to user are improved in their ranking weights. The experiments have been performed and compared for different search engines like AltaVista, DirectHit, Excite,HotBot, Google, Lycos and Yahoo .totally fifteen queries are used and the Spearman's correlation coefficient between the different results is computed, from this computation it is inferred that Yahoo performed best followed by Google, Lycos, HotBot, AltaVista, Excite and DirectHit. Future work in this direction can be performed by extending this feedback vector to include more feedback sources like what area of result set the user was looking most of the time i.e. eye gaze and focus tracing, what were the facial expressions and what level of neuro satisfaction was present in the activities of user once he or she is exposed with the result list. The above SQM vector in fuzzy form has already been used by [7] to compute the penalties and improvements in the ranks of for those documents for which user has negative and positive feedback respectively. This approach can be further extended with probabilistic feedback computation and updating in ranks of documents. A neural network based approach of computing the user satisfaction with search engine is given in [18].

Information requirements of a user in an interactive retrieval system are represented by user queries, and somewhat this feedback vector in a closed loop system, but through limited number of query terms, user can model the information needs in a vague sense. Performing the trend analysis in querying the system, providing personalization features, effective indexing and fuzzy thesauri, this uncertainty can be handled at multiple levels. There are the soft computing techniques to transform computational uncertainty of information retrieval process (i.e. of defining information requirement, matching and ranking the documents) into deterministic, predictable but approximate solution. At the level of representation of query and documents, indexing and thesauri there has been significant of research performed in the past, that addressed representation of the documents in the form of a fuzzy sets, representation and processing of fuzzy queries, fuzzy thesauri computation, extended fuzzy queries etc. have given new ways to approximate user requirements, matching, computing result sets and incorporating fuzziness in the user feedback to form fuzzy closed loop interactive information retrieval systems. In order to judge the performance of a soft information retrieval system, researchers have decided the soft measures and matrices that fits above simple precision, recall, accuracy to the user expectations, subjective evaluation of quality of search results is computed and performance of the system is boosted by implicit relevance feedback .Today's competitive benchmarks states that a system which is highly accurate but unable to address the information requirements in real world sense, in time, space, dimensions, depth, level of summarization is not likely to survive unless it continuously optimizes its results as it learns from user interaction. Information retrieval process is under continuous change in its structure and implementation, with the advances in type of applications IR addresses, supported languages and localities, business interests and social integration over the internet, change in the definition of information relevance over the time. The need of new performance measures tools in increased to keep pace with this change at assessment coordinate system, for example two extreme cases of information retrieval are ad hoc retrieval and personalized searching. In each case the system have different amount of prior knowledge regarding searcher profile; therefore the performance of the system will be relatively different for same interaction type, one need to device distinct operators to measure performance in both cases. In this section we have addressed the advanced work on performance analysis of web based information retrieval systems and using that in the form of feedback to improve the next session result lists.

The feedback based IR systems concentrates on the increasing satisfaction of user. There are model based and adaptive systems in this field proposed by researchers and we have included references to little representative work in this text. Soft computing offers lots of opportunities in the design of fuzzy, learning systems which can optimize the result set of a search engine to fulfill user requirements. But the goal of these different approaches, models, hybrids and the methodologies discussed in this text is to decrease the dissatisfaction of user with a search result.

In [5] six models of implicit feedback are investigated for evaluation of interactive information retrieval systems by interaction with a searcher using implicit feedback mechanism. In fact the feedback collected by a search engine is highly dependent the user interface of the result set. in case of enterprise solutions the user interface is highly sensitive and records all the interactions of user with result set, while in case of a general purpose search engine user interface may not take into account the different activities of user like book-marking certain page, saving a page from result set, emailing it to a friend, further exploring the result page, time that user spent on the particular web page [M.M.S. Beg 98].we have summarized the reviewed models in [5] are given in fig.1

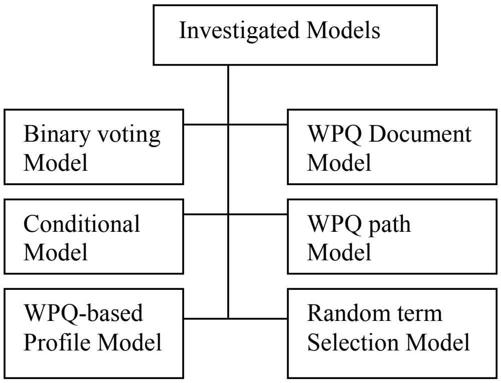

Figure 1: Models reviewed

In this paper binary voting model is a kind of competitive relevance feedback model in which each document either get votes in favor or it do not get votes. All the terms are participant for voting and if for a term a document is in viewed list, than it gets positively voted. This scheme is sensitive to document representation and to increase the resolution through voting the heuristic weights to representative terms are added. This modification provides greater control over how the document parts are important for a voting. Jefferies model revise the probability of relevance of a term based on implicit evidences captured during searcher interaction with result. This method approximates the relevance path between the document representations rather than documents. WPQ model [Robertson, 1990] based three variations are discussed, they are document WPQ which consider full length documents for RF,WPQ path model and WPQ ostensive profile model. Finally there are random term selection models which assign values between 0 and 1 to the terms in viewed documents. This research work has done simulation based evaluation using a matrix of assessment criteria for different models and successfully compared the performance of all the six models in average best and worst path performances, in this simulation Jeffery's conditioning model and WPQ document model outperformed other models. Results are compared for different scenarios; full details of experiment can be referred to original publication [5].

C. Soft computing in intelligent information retrieval:

A soft information retrieval system is that which can address the information need beyond the similarity measure and up to the level of relevance, satisfaction, repeatability, and the usefulness in the eyes of user. Main soft computing techniques are fuzzy logic, artificial neural network, genetic algorithms [11 16], some AI techniques [15], machine learning, pattern recognition and probabilistic information retrieval. Mining the web is different from the traditional and enterprise information retrieval [3], in terms of volume of data available on any topic on internet, different types of files, nature of users, heterogeneity and distribution of literature, dynamic nature and growth of the web, deletion of older and addition of new web pages etc. It is observed that many of the traditional data mining the algorithms for information retrieval can be extended to the web; usually they have been reported to perform well. A survey of soft computing techniques for traditional or core data mining task is given in [8], in this survey paper role of fuzzy sets, Artificial Neural networks, genetic algorithms and rough sets is described[8 11] and some of the key data mining challenges are discussed that lead to hybridization of different techniques. The need of intelligent data analysis techniques is explained due to inappropriateness of statistical data analysis due to volume of data and high dimensionality. In the light of challenges in the process of KDD [8] has identified the similar future questions as in case of web mining but the scale (data size and volume of queries), structure (considering metadata in search) and dynamics of web leads to the specialized research in the fields of web content mining, web usage mining, and web structure mining separately [9]. some of the key challenges that are common encountered in web mining as well as in core data mining research are reported in [5] are massive data sets and high dimensionality, user interaction and prior knowledge, overfitting and assessing the statistical significance, understandability of patterns, non-standard and incomplete data, mixed media data, managing the changing data and knowledge, integration. In case of web mining the problem of massive data is coupled with heterogeneity and geographical distribution, various categories of websites, hidden databases and totally incompatible systems in their data format and operating platforms, which may be interconnected or possibly unreachable through hyperlink structure like deep web. The problem of web search is not limited to knowledge discovery from web for indexing purpose, but identification of cluster of sources of data given a topic, classification of web documents into categories, and dealing with unstructured and heterogeneous data sets are immediate additional tasks. [8]

Has listed the important work in the field of the fuzzy set and its applications in associative rule mining, fuzzy taxonomies, clustering, generalized functional dependencies, fuzzy data summarization, fuzzy web user profile personalization, fuzzy set based categorizing web session by web mining, interactive image retrieval.

In case of data rich mining applications artificial neural networks are used for classification, clustering, learning and generalization of symbolic rules[8 17 18] from underlying data. In [8] the applications of ANN to data mining are categorized into rule extraction, rule evaluation, clustering and self organization and regression. Pointers to important hybrid neuro-fuzzy systems are given with their strengths and applications like fuzzy MLP, fuzzy Kohonen, networks neuro-fuzzy knowledge base, etc. . For the large data sets where user preferences can be modeled in the form of a fitness function, GA is best used soft computing technique. Rough set theory, which is another soft computing technique, is used as a mathematical tool to discover the redundancies and dependencies between the given features of a problem to be classified [8]. Soft computing techniques like rough-fuzzy, rough-neuro-fuzzy, rough-neuro-fuzzy-GA based frameworks for data mining are discussed. A fuzzy-GA based soft IR algorithm is revisited in [12] which use the fuzzy genes created from user profile of a web searcher to optimize the future search.

IV. Soft Computing in Web based search

Techniques for a web based search for relevant information is formally studied under the topic of web mining. In the next few headings we have described the types of web mining, challenges in different web mining tasks and applications of soft computing techniques in web mining.

A. Web mining: A Survey of Surveys

Web mining refers to the use of data mining techniques to automatically retrieve, extract, evaluate (generalize/analyze) information for knowledge discovery from the web documents and services [12].An excellent survey of web mining research is given in [9], in this paper author has suggested to decompose the web mining task into subtasks like resource finding, information selection and preprocessing, generalization and analysis. This work has surveyed the work related to interaction between research areas like web mining and Information Extraction, web mining vs. machine learning and web mining vs. agent paradigm. A more complete taxonomy of the web mining related activities is addressed in [10 12]. In [12], described taxonomy graphically mimics the activities of web mining. for instance, web structure mining deals with navigation, mining XML and HTML. Research work on the web content mining is concentrated at clustering, association rules, semantic web, web page content mining, search result mining, text mining and image mining. The web usage mining deals with the personalization, business intelligence, customer profile, use profile, system improvement, recommendation, e-commerce, intrusion detection and web agents. For more detailed study, [19] is a survey of the web content mining techniques, [20] can be consulted for web structure mining and [21] contains a survey of the web usage mining techniques and applications of the usage pattern collected from online users. One more extensive a survey of web mining in soft computing framework is [11]. It focused on the differences between the web mining and core data mining tasks, a survey of soft web mining techniques and the limitations of current web mining applications are given. Limitations of the web mining techniques described in [11] are in terms of inability of current mining algorithms to properly address the subjectivity, imprecision and uncertainty of web mining task, inability to deduct the correct answer for certain queries, inability of making soft decisions, page ranking problem, and dealing with outliers.

B. Soft Web Mining: Application of Probability Theory, FL, ANN and GA

In this section we have surveyed the research work in the field of soft web mining using fuzzy set theory, probabilistic information retrieval techniques to handle the uncertainties of IR, ANN, GA and hybrid approaches to solve the problems of IR in general and web mining to be specific. Soft information retrieval methods are more effective than their counterpart traditional model based solutions because they have advantage of learning, handling uncertainty and optimization of result sets.

1) Fuzzy Sets and their Applications in web mining:

A fuzzy set is defined as a generalization of crisp set, unlike crisp sets, which deals with the memberships of the set elements in binary fashion a fuzzy set is multi-valued set in which each member can have a membership value between zero and one .the membership value is determined by the membership function of the fuzzy set. The fuzzy sets are capable to represent the uncertainty and vagueness of search process; therefore researchers have used the concept of fuzzy sets to modify the Boolean search and extended Boolean and fuzzy search algorithms have been proposed. The Boolean search algorithm has a limitation that returns only those documents which exactly match the query terms, but if the terms present in the Boolean type of query are not directly appearing in the documents, a Boolean query processing algorithm will result no document marked as relevant to the user query. alternatively in a fuzzy information retrieval the similarity between query and a document is judged based on the partial relevance between them, even if the query terms are not present in a document collection directly, but there may be some of the documents which may be partially useful for the user, this partial similarity can be represented in the form of fuzzy membership in the set of similar documents. Crisp similarity measure is a specific case of fuzzy similarity measures which works on strict binary sense. Real world problems, such as fuzzy similarity measure is capable to represent more granular form of relevance as compared to Boolean similarity measures. The introduction and details of information retrieval using fuzzy sets is given in [22, 23, 24, 25] from different perspectives. In [23], Gloria Bordogna and Gabreilla Pasi have proposed a fuzzy rule based information retrieval system, which can be used in association with linguistic fuzzy terms in query. In this research work a method is proposed to compute a linguistic estimate of the relevance of a document to a user query. A novel fuzzy document representation is proposed, this representation is used to evaluate user queries against numerically assigned and linguistic weights to different terms. In [24] a method is proposed to represent the numerical weights of a query terms in the form of fuzzy label. Representation of information requirements in the form of fuzzy linguistic labels is advantageous in dealing with uncertainness and vagueness in user specification. This paper justifies the shift in paradigm in information retrieval from Boolean retrieval to fuzzy linguistic query based IR in the light of inability of Boolean retrieval systems to mimic the user requirements in real world. A short survey of applications of fuzzy set theory and ANN in IR systems is given in [25].

2) Probabilistic Information Retrieval:

A probabilistic indexing method considers that the query Q is a random variable and it finds the documents which are probability relevant to a term "i" because they were really relevant for all the searches for term 'I'. While the binary independent model consider the random variable as documents d. There are different applications of probabilistic inference based techniques in intelligent information retrieval starting from document surrogates, probabilistic indexing, probabilistic relevance matching, probabilistic sampling of the different hidden databases, probabilistic retrieval of Top-K documents etc.

3) ANN Based Applications for IRS:

Artificial Neural networks are used to learn from data. They are used in data rich applications for classification, generalization and other applications successfully. In Intelligent information retrieval systems ANN is used for classification, clustering, learning ontology's and concepts from user interaction.

4) Genetic Algorithms and IR:

Genetic Algorithms is based on the mechanism of reproduction process from biology and in computer science it is implemented as an instance of multi-objective optimization. In an IR system, Genetic algorithms are used in optimizing the ranks of documents in search results according to the user feedback.

V.Future Systems

Future work include the proposals for mind programmed information retrieval, Neuro-Sasisfaction based IR algorithms and the hybrid solutions for soft computing based Intelligent information retrieval.

VI. Conclusion

We have presented a review on some of the important and representative research works in the field of model based and soft information retrieval. We have discussed the topics of the performance of an IR system in terms of user feedback vector and proposed more granularities in feedback mechanism. Important work in the area of challenges in IR, evaluation, subjective quality measure and soft computing techniques for web mining are discussed. Some of the future directions are identified in each of the discussed topic.

However, the comparative study of performances and the effectiveness of the above methods is another research topic in itself, but scope of this survey is to concentrate on the challenges of IR and to identify few representative methodologies to handle these challenges. The selection of model based approaches from Belkin [2] is based on the fact that these are representative model based approaches, but still there is a huge space to survey Model based approaches separately. Also this paper highlighted the soft information retrieval starting from its meaning and up to some important research work these references are selected in the interest of those readers who want to understand the applicability of soft computing in the design of IR from scratch. Another Identified direction for research survey is the applicability of soft computing techniques in Intelligent IR, Used Feedback mechanisms to decrease the dissatisfaction. In fact in the next few survey papers I have a goal of being concentrated on the mathematics of body of researches in these dimensions.

The model based solution to SQM problems is discussed and the soft computing approach is introduced. Actually application of neuro-fuzzy and GA hybrid system is expected to yield more adaptive, optimized and fuzzy solutions to the problem of IR. In this context [6 7] are the representative methodologies which gives fair idea that how to use the soft computing in search quality measures for proposed algorithms for an IR. Readers can notice that this text has also proposed a future work of neuro satisfaction of user, in fact this survey is meant for identifying the problems which are not addressed in model based methodologies and addressing these problems with soft information retrieval with some leading example research work in the filed, out of numerous solutions.

Acknowledgment

I acknowledge the work of all those researchers who have contributed to the field of IR, which I have taken support of and could not mention due to limitation of space in this paper. I also acknowledge the work of all those authors which I have reviewed, my university and friends.

References

[1] James Allen et. al, "Challenges in Information Retrieval and Language Modeling" pub in SIGIR newsletter, ACM-SIGIR Forum, vol 37(1), Spring 2003

[2] Nicolas J Belkin ,"Some(what) Grand Challenges for Information Retrieval ".SIGIR Newsletter,ACM-SIGIR forum vol 42(1),2008

[3] Devid Hawking et.al."New Methods for Creating Testfiles:Tuning Enterprise Search with C-TEST ".SIGIR proc. of Future of IR evaluation July 2009,Boston.

[4] Nicolas J Belkin , et.el."A Model for Evaluation of Interactive Information Retrieval ".SIGIR proc. of Future of IR evaluation July 2009, Boston.

[5] Ryen W.White ,"Evaluating Implicit Feedback Models using Searcher Simulation", ACM Transaction on information systems vol.23(3),2005.

[6] MMS beg and N. ahmad ,"A Subjective Measure of Web Search Quality" IJIS, Elsevier Vol 169 no.3-4,2005 pp 365-381.

[7] MMS Beg and N.Ahmad," Soft Computing Techniques for Rank Aggregation on World Wide Web", World Wide web –an International Journal,Kluwer vol 6(1),march 2003 pp 5-22

[8] Sushmita Mitra et.al, "Data Mining in Soft Computing Framework: A Survey "IEEE Trans. Of Neural Networks, vol 13(1), July 2002.

[9] Raymond Kosala and Hendric Blockeel, "Web Mining Research:A Survey",SIGKDD Exploration,ACM,july 2000.

[10] Pranam Kolari and Anupam Joshi, "Web Mining Research and Practice", copublished by IEEE CS and AIP,July/Aug 2004.

[11] Sankar K..Pal et.al."Web Mining in Soft Computing Framework:Relevance,State of Art and Future directions",IEEE Trans of Neural Networks,vol 13(5),Sep 2002.

[12] Dragos Arotaritei et.al.,"Web Mining: A Survey in the Fuzzy Framework",Fuzzy Sets and Systems, Elsevier 148 (2004) pp5-19.

[13] Fabio Cristani and G.Pasi,"Soft Information Retrieval:Application of Fuzzy Set Theory and Neural Networks "

[14] O.Cordon et.al."A review on the application of evolutionary computing on information rterieval"International Journal of Approximate reasoning, Elsevier, vol. 34,(2003), pp241-264.

[15] KK Shukla, "Some AI Techniques for Information retrieval"DESIDOC Bulletin of Information Technology, vol. 16 (4), July 1996, pp13-18.

[16] Jin Zang et.al,"Impact of metadata implementation on webpage visibility in search engine result" Information Processing and Management Elsevier, vol. 41(2005).

[17] Qin He, "Neural Network and its Application in IR",UIUCLIS-1999/5+IRG.Spring,1999.

[18] Sunita Yadav et.al. "Neural Network based Approach for Predicting User Satisfaction with Search Engine" International Journal of Computer Applications (0975-8887), vol.18 (5), 2011.

[19] S.Chakrabarti, "Data Mining for Hypertext: A Tutorial Survey"ACM SIGKDD Explorations vol. 1(2),2000,pp1-11

[20] J. Furnkranz, "Web Structure Mining: Exploiting the Graph Structure of World Wide Web"OGAI vol.21(2),2002 pp17-26

[21] J.srivastava et.al."Web Usage Mining:Discovery and Applications of Usage Patterns from Web Data" ACM SIGKDD explorations,vol 1(2),2000, pp12-23.

[22] Limin Ren, "Research of Web Data Mining based on Fuzzy Logic and Neural Networks" proceedings of 6th IEEE Int.Conf. On Fuzzy Systems and Knowledge Discovery, 2009.

[23] Gloria Bordogna and Gabriella Pasi ,"Fuzzy Rule Based Information Retrieval",proc. of 18th conf on North American fuzzy information processing society,NAFIPS,1999 pp585-589

[24] Gloria Bordogna and Gabriella Pasi,"A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval: A model and its evaluation"

[25] Fabio cristani and G.Pasi" Soft information Retrieval: Applications of Fuzzy Set Theory and Neural Networks"N.kasabov and Robert Kozma -Editors,Physica-Verlag,springer-verlag group,278-313,1999

[26] VagelisHristidis et.al., "Relevance-based Retrieval on hidden-web Text Databse without Ranking Support", IEEE Transaction of Knowledge and Data Engineering, 2010.

[27] T.M.Nogueira et.al,"Fuzzy Rulesfor Document Classification to Improve Information retrieval System"International Journal of CIS and IMA,ISSN 2150-7988 vol(3) 2011,pp210-217.