Web mining: a survey in the fuzzy framework☆

Dragos Arotaritei[a], Sushmita Mitra[b,*]

[a] *Department of Computer Science, Aalborg University Esbjerg, Niels Bohrs Vej 8, 6700 Esbjerg, Denmark*
[b] *Machine Intelligence Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India*

Abstract

This article provides a survey of the available literature on fuzzy Web mining. The different aspects of Web mining, like clustering, association rule mining, navigation, personalization, Semantic Web, information retrieval, text and image mining are considered under the existing taxonomy. The role of fuzzy sets in handling the different types of uncertainties/impreciseness is highlighted. A hybridization of fuzzy sets with genetic algorithms (another soft computing tool) is described for information retrieval. An extensive bibliography is also included.

Keywords: Fuzzy sets; Soft computing; Web mining; Information retrieval

1. Introduction

Web mining refers to the use of data mining techniques to automatically retrieve, extract and evaluate (generalize/analyze) information for knowledge discovery from Web documents and services. Web data is typically unlabelled, distributed, heterogeneous, semi-structured, time varying, and high dimensional. Hence any human interface needs to handle context sensitive and imprecise queries, and provide for summarization, deduction, personalization and learning. Almost 90% of the data is useless, and often does not represent any relevant information that the user is looking for. Taking into account the huge amount of data storage and manipulation needed for (say) a simple query, the processing essentially requires adequate tools suitable for extracting only the relevant, sometimes hidden, knowledge as the final result of the problem under consideration.

The use of soft computing tools, including fuzzy logic, in data mining has been adequately reported in literature [24]. However, the subtle differences between data mining and Web mining suggests the

use of new or modified tools and algorithms for appropriate handling of the Internet. Web mining typically addresses semistructured or unstructured data, like Web and log files with mixed knowledge involving multimedia, flow data, etc., often represented by imprecise or incomplete information. This implies that fuzzy set theoretic approaches are useful instruments in order to mine knowledge from such data.

The huge amounts of multivariate information offered by the Web has opened up new possibilities for many areas of research. Due to the involvement of human interaction in Web information (like text, image, sound, and linkages between them), new tools and methodologies need to be extended in order to deal with the incomplete or imprecise information. Web personalization, navigational Web, semistructured and structured data information are some of the major issues under Web mining. Semantic Web is also a target for data mining research using fuzzy logic, because of the inherent vagueness of human beings when expressing information in natural language. Considering the Web as a large distributed multimedia database, an extension of methodologies to deal with them and their mining algorithms can be considered under image mining. The huge volumes of compressed and uncompressed information stored in images, and their subjective evaluation by humans in interaction with the Web, is another focus of current data mining research requiring fuzzy treatment.

Web mining can be broadly categorized as

- Web Content Mining of multimedia documents, involving text, hypertext, images, audio and video information. This deals with the extraction of concept hierarchies/relations from the Web, and their automatic categorization.
- Web Structure Mining of inter-document links, provided as a graph of links in a site or between sites. For example, in Google a page is important if important pages point to it.
- Web Usage Mining of the data generated by the users' interactions with the Web, typically represented as Web server access logs, user profiles, user queries and mouse-clicks. This includes trend analysis (of the Web dynamics information), and Web access association/sequential pattern analysis.

Fig. 1 provides a taxonomy for Web mining. The different functions falling under the three major categories are highlighted. A relationship with information retrieval is also depicted.

The role of fuzzy sets in Web mining [5] holds promise mainly in (i) document and user clustering, (ii) deduction and summarization, (iii) handling of fuzzy queries involving natural language and/or linguistic quantifiers like *almost*, *about*, and (iv) information fusion in multimedia data. According to Zadeh [35], fuzzy logic may serve as the backbone of the Semantic Web, an extension of the current Web in which information is given well-defined meaning, thereby better enabling computers and people to work in cooperation.

Ordinary end-users often face difficulties in formulating a precise representation of their information needs in a Boolean query. This affects the efficiency of the information retrieval process. Hence Web search engines require the use of fuzzy aggregation operators. These are specially suitable in flexible query answering and information retrieval. User feedback is a process where the user expresses her/his opinion about the documents that the system has retrieved as an answer to a certain query. This user evaluation/opinion can be used both for training the classification system as well as reflected in the corresponding user profile. Fuzzy linguistic modeling is useful in handling such preference relations. This information can also be utilized by (say) marketing experts to analyze user interests.

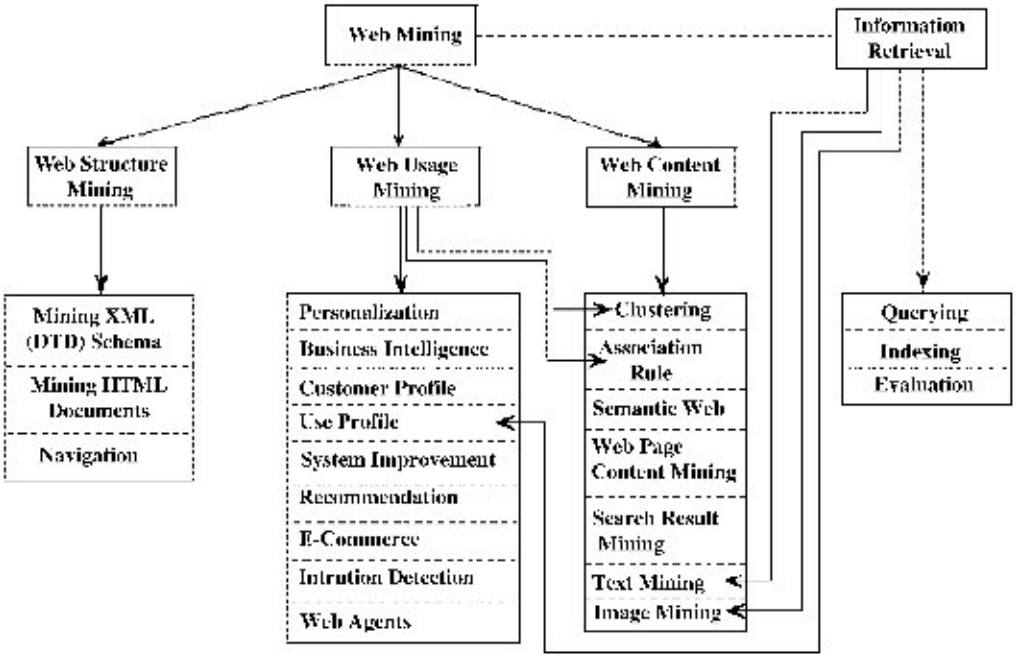

Fig. 1. A web mining taxonomy.

In this article we present a survey on the use of fuzzy sets in the different aspects of Web mining. In Sections 2–9 we dwell on varied issues like Web clustering, association rule mining, Web navigation, Web personalization, Semantic Web, information retrieval, text mining and image mining. The article is concluded in Section 10.

2. Web clustering

Clustering pertains to unsupervised learning, where no predefined classes are assigned. The key requirement is the need for a good measure of similarity between the instances/patterns. The problem is to group n patterns into c desired clusters, such that the data points within clusters are more similar than across clusters. Scalable clustering algorithms pertain to working with large volumes of high-dimensional data, that is inherent to data mining problems. Moreover, the presence of mixed media data over the Web call for the use of specialized techniques.

The use of fuzzy logic has been widely studied in clustering problems [1], and the results have also been extended to handle high dimensional data [17,28]. The importance of clustering to Web mining, specifically in the domains of Web Content and Web Usage mining, make Web clustering an interesting topic of research. This includes clustering of Web documents, snippets and access logs. Usually the Web involves overlapping clusters. So a crisp usage of metrics is better replaced by fuzzy sets which can reflect, in a more natural manner, the degree of belongingness/membership to a cluster.

In the fuzzy c-means algorithm, the objective is to minimize the function

$$J(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}diss(v_i,x_j) \quad \text{subject to} \quad \sum_{i=1}^{c}u_{ij}=1 \text{ for all } j, \tag{1}$$

where $diss(v_i,x_j)$ is the distance of pattern x_k from the ith cluster mean (typically expressed as squared Euclidean distance), u_{ik} is the membership function (in the interval [0,1]) of point x_k in the ith cluster such that $0<\sum_{j=1}^{n}u_{ij}<n\ \forall i$, U is the fuzzy c-partition of the data set, V is a set of c-prototypes and $m>1$ is the fuzzifier. However, even a few outliers or inherent noise in real data can affect the result of this algorithm. Fuzzy c-means has problems finding correct clusters in the presence of noise or outliers, because of its assumption that any point in a dataset must essentially belong to a cluster.

Using a noisy cluster concept in [6], all the noise points are collected in a separate cluster. These noise points are supposed to have a distance $diss(v_c,x_k)=\delta$ from their prototype v_c. Any pre-specification of δ is not easy, and the distance can be different for different kinds of problems. One way is to use a simplified statistical average to compute δ as

$$\delta=\lambda*\left[\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}diss(v_i,x_k)}{n(c-1)}\right], \tag{2}$$

where λ is a factor experimentally determined to lie between 0.1 and 100.

The probabilistic constraint that the membership u_{ij} of a data point x_j have a sum of one, is relaxed in the possibilistic approach to fuzzy c-means [18]. Here we have

$$\max_i\{u_{ij}\}>0 \quad \text{for all } j,$$

in Eq. (1). In order to avoid the trivial solution $u_{ik}=0$, a penalty function for low membership is appended to the objective of Eq. (1). The objective function now becomes [18]

$$J_P(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}diss(x_j,v_i)+\sum_{i=1}^{c}\alpha_i\sum_{j=1}^{n}(1-u_{ij})^m, \tag{3}$$

where the α_i's are a set of chosen numbers. The first part of this equation minimizes the distance of the feature vectors from the prototypes, while the second term maximizes the membership degree of this vector to the corresponding prototype. The membership functions u_{ij} are viewed as possibility distributions for clusters U_i over the domain of discourse of feature vectors x_j. The terms α_i are chosen based on the shape and size of cluster U_i, and the relative weight assigned to the second term in Eq. (3). A typical choice is to make α_i proportional to the average fuzzy intra-cluster distance of cluster U_i. Depending on the distance measure $diss(x_j,v_i)$ used, one derives the possibilistic c-means, possibilistic Gustafson–Kessel and possibilistic c-quadric shells algorithms.

A generalized model called alternating cluster estimation (ACE) [30] uses alternating iterations on model architectures, while the membership and prototype functions are selected directly by the user. (S)he can choose the form of updating of equations that calculate the new partition $U(t)$ and new prototypes $V(t)$ at each iteration t. The pair of updating equations can be changed in order to produce a model that matches better with the data to be clustered. Selection of membership function families (triangular, hyperconic, etc.) by approximating from available data, and the prototype function builder are the basics of an ACE interface.

A review of robust methods used for fuzzy Web clustering is presented in [13]. A robust algorithm is one whose performance is minimally affected in the presence of outliers. The authors present the fuzzy c-means (FCM), fuzzy trimmed c-prototype, fuzzy c-least medians and relational fuzzy c-means algorithms, along with some techniques used to make them robust.

The fuzzy c-medoids (FCMdd) algorithm can be summarized as follows. Let $X_c = \{x_i \mid i = 1, \ldots, n\}$ be a set of objects, $diss(x_i, x_j)$ be the matrix of dissimilarity between the objects x_i and x_j and $V = \{v_i \mid i = 1, \ldots, c\} \subseteq X_c$ be the set of representative points. The objective function E_m that must be minimized over all V in X_c is given as

$$E_m(V, X_c) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m diss(x_j, v_i), \tag{4}$$

where u_{ij}^m represents the fuzzy or possibilistic membership of x_j to cluster i with a heuristic determination of m (that controls the shape of this membership function). The steps of the algorithm are as follows:

(i) Set the number of clusters c
(ii) Randomly pick up initial set of medoids V from X_c
(iii) Set $iter = 0$
(iv) For $i \leftarrow 1$ to c do // Compute the membership
  For $j \leftarrow 1$ to n do
   Compute u_{ij}^m, using eqn. (6), for $x_j \neq v_p$
  endfor
 endfor
 $V^{old} = V$ // Store the current medoids
 // Compute the new medoids
 For $i \leftarrow 1$ to c do

$$q = arg \min_{1 \leqslant p \leqslant n} \sum_{j=1}^{n} u_{ij}^m * diss(x_p, x_j) \tag{5}$$

$$v_i = v_q$$

(v) If $iter = MAX\_ITERATIONS$ or $V^{old} = V$, then stop else go to step 4

The membership function is expressed as

$$u_{ij}^m = \frac{\left(\frac{1}{diss(x_j, v_i)}\right)^{1/(m-1)}}{\sum_{p=1}^{c} \left(\frac{1}{diss(x_j, v_p)}\right)^{1/(m-1)}}. \tag{6}$$

For $m = 2$ this boils down to the harmonic mean of the dissimilarities. The sum of square criterion is used in the dissimilarity matrix, typically expressed as the distance in the FCM and FCMdd algorithms. This is found to be not very robust.

In the robust FCM, one way of achieving robustness is by modifying Eq. (4) to

$$E_m^{\mathrm{R}}(V,X_c)=\sum_{j=1}^{n} g\left(\left(\sum_{i=1}^{c}(diss(v_i,x_j))^{1/(1-m)}\right)^{1-m}\right), \tag{7}$$

where g is a loss function that reduces the role of outliers and v_i refers to the ith class mean. Other approaches include possibilistic c-means [18], noise clustering [6] and alternating cluster estimation [30], as discussed earlier in this section.

The fuzzy c-medoids and fuzzy c-trimmed medoids are used to cluster relational data from Web documents and snippets in [16]. The algorithms are applied to a collection of 1042 abstracts from the Cambridge Scientific Abstract Web site, corresponding to 10 topics. A preprocessing stage is used to filter and remove irrelevant words, in order to generate the input feature vector that is computed using an *inverted document frequency* method. This 500-dimensional feature vector (keywords) is reduced using principal component analysis, resulting in a selection of 10 eigenvector values. The algorithms are also tested on a collection of snippets, corresponding to 200 Web documents collected by a search engine in response to a query "salsa". The results are extended in [17] using robust estimators, providing a computational complexity $O(n \log n)$ instead of $O(n^2)$. This is suitable for clustering noisy data that are characteristic of Web documents. An application of the robust clustering algorithm is made for mining Web access logs in [14], in order to automatically discover user session profiles.

The ACE model for numerical values [30] is extended to handle non-numerical patterns in relational data sets using relational ACE (RACE) [29]. The authors consider two types of patterns related to the Web, viz., (i) document contents such as text parts of Web pages (Web content mining), and (ii) sequences of Web pages visited by users, such as Web logs (Web usage mining). Both text and Web page sequences are handled using the Levenshtein (edit) distance, that determines the number of edits (delete, insert, change) steps needed to convert one word into another, and a graph-based distance measure. The prototypes found for Web page sequences are interpreted as typical click streams reflecting prototypical user interests. This provides an automatic data-driven user segmentation of the Web page sequences.

3. Association rule mining

Association rules, in the context of Web mining, refer to the determination of (say) those URLs that tend to be requested together. This can be categorized under Web Usage and Web Content mining. In this section we describe some methodologies that incorporate fuzzy set theoretic concepts.

Web mining of inference amplification is made in [21], using inference logic from fuzzy cognitive maps in three phases. The first stage mines association rules with the a priori algorithm, from a Web-log database. A corresponding fuzzy knowledge map involving causality (the cognitive map uses values between -1 and $+1$) is built in the second stage, incorporating both positive and negative rules. In the final stage, the system applies inference amplification in order to enrich the resulting Web mining association rules. Experiments are made using a Web-log file from a real online shop, specializing in computers and peripherals. The causal knowledge is represented as an adjacency

matrix, including the connectivity $W^{\mathrm{T}} = \{w_{ij} \mid w_{ij} \in \{-1, +1\}\}$. Here w_{ij} indicates the causal value (weight) of the arc from vertex/node C_i to C_j.

The definition "A causally increases B", implies that an increase/decrease of A causally increases/decreases B. The equivalent fuzzy relation (causal knowledge) used for inference amplification is expressed as

$$C_i \overset{w_{ij}}{\to} C_j;\ C_i \overset{\sim w_{ij}}{\to} \sim C_j;\ \sim C_i \overset{\sim w_{ij}}{\to} C_j;\ \text{and}\ \sim C_i \overset{w_{ij}}{\to} \sim C_j,$$

where $\sim$ indicates negation, and w_{ij} is the fuzzy membership function value $R(C_i, C_j)$ in the fuzzy relation between the fuzzy concept nodes C_i and C_j. After the chained rule is extracted, the inference amplification is made according to the above rules. The basic case $T \to T$, with $T = true$ and $F = false$, in the rule

$$A\ 0.2\ (T) \to 1.0\ B\ (T) \to 0.84\ C\ (T)$$

signifies that 20% of A browse B, and that 84% of them buy the product C. When inference amplification is applied, it produces the additional subtle information

$$A\ 0.8\ (F) \to 1.0\ B\ (T) \to 0.16\ C\ (F).$$

Here the case $F \to F$ is extended in a manner analogous to the case $T \to T$.

Fuzzy association rules are mined to discover access path prediction in [34]. The authors use a fuzzy index tree, involving fuzzy cases, in order to reduce the explosion in the number of rules generated in the corresponding crisp state. The experiments use the data from the logs made in www.microsoft.com.

A meta-search engine based on query keywords, that uses instant information retrieval given by Grey relational method, is described in [12]. A fuzzy inference model is used in the common case when the number of keywords is greater than or equal to two. The related sites are mined in order to verify the users' expectation. The term frequencies and document frequencies are fuzzified according to nine given rules. The meta-search engine finally locates the closest related site, based on keywords, from the queries.

4. Web navigation

Navigation is categorized as Web Structure mining. An optimization of the path for surfing the Web, given a target, is described in [4]. The connected Web sites are represented by a directed graph with source and destination nodes, and a links set along the path. The frequency of accessing various links (access rate) and the time taken to retrieve target pages (retrieval rate) are considered as the decisive factors. These are affected by criteria like the availability of channels, server capability and access time. The expected access rate and the required retrieval rate are expressed as fuzzy sets. Link weights are associated with appropriate linguistic values. A fuzzy opinion matrix expresses subjective estimation of users during their surfing of specified links on the path. The general evaluation of the link, with respect to interest rate of users, is provided as the intersection among the estimation made by the different users. The optimal path is computed as the minimum fuzzy distance estimated between a fuzzy *Hurwicz opinion set* (derived in terms of optimism–pessimism index) and the actual requirement set.

Web mining is applied to log files in order to discover user's behavior [11]. Numerical data, representing the time spent by an user in a page, is quantified using linguistic membership functions. The domain of discourse [0,130] is split into three trapezoidal membership functions, viz., short, middle and long. The preprocessing phase consists of sorting the log files by client and time access, and mapping the information to a flat database that is mined as multiple items transaction in the fuzzy framework. Maximally large sequences are discovered as a list of most frequent successively traversed Web pages, that satisfy the minimum support criterion for fuzzy association rules. However the fuzzy rule mining algorithm used to identify the sequential navigational patterns is complex, and costly in terms of time and memory requirements if the number of users and number of pages is high (say, thousands of pages). As a consequence, it is possible to have very long frequent sequences comprising of hundreds of items.

5. Web personalization

The increasing popularity of the Internet and the exponential increase in the number of its users has led to the creation of new paradigms of knowledge discovery, like Web personalization, mining bookmarks, mining e-mail correspondences, recommendation systems, and so on. These are grouped as Web Usage mining.

Mining typical user profiles and URL associations from the vast amount of access logs is an important component of Web personalization, that deals with tailoring a user's interaction with the Web information space based on information about her/him. Nasraoui et al. [26] have defined a *user session* as a temporally compact sequence of Web accesses by a user and used a dissimilarity measure between two Web sessions to capture the organization of a Web site. Automatic discovery of user session profile is made using fuzzy competitive agglomeration for relational data (CARD) algorithm. Complex, non-Euclidean distance/similarity measures can be handled in this framework. The objective function is now defined as

$$E^A(V,X_c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2.g\left(diss(x_j,v_i)^2-\alpha\sum_{i=1}^{c}\left(\sum_{j=1}^{n}w_{ij}.u_{ij}\right)^2\right) \tag{8}$$

subject to $\sum_{i=1}^{c}u_{ij}=1$ for all j, where α is a function of the iteration number, $g(.)$ is the loss function and $w_{ij}\in[0,1]$ are the robust weights. Log files are collected from a real site in order to mine the user profiles. The experimental results "capture" the pattern for users in profiles like outside visitor, prospective student, those that attend lessons by the same professor, an so on. To discover user profiles from real log files, various fuzzy clustering methods have been applied. The robust fuzzy c-medoids [14] is one such example. Another approach in this direction is the relational alternating cluster estimation [29].

A system that looks for Web documents using link-based search and a fuzzy concept network is described in [15]. The user's subjective interests are appropriately represented by a fuzzy concept network based on user profile. Missing information is inferred from a transitive closure of a matrix of knowledge in the network. The degree of relevance in the network is fuzzified as a value between 0 and 1. The importance of the document is computed by the fuzzy retrieval system that personalizes the results given by the search engine. It ranks the retrieved documents, and finds the authoritative

and hub sources. Five best authoritative sources for query are selected as the most representative documents corresponding to a user's query. Documents are ranked according to the user's interests stored in her/his profile (say, ten concepts as "Book", "Java", etc.) The experimental results present the ranked evaluation by three users, and the personalized result in response to their query for "Java".

A Fuzzy Cognitive Map (FCoM) is used in [23] for a causal modeling of the behavior of the user during her/his search on the Web. The learning of connections in the FCoM is made by neural learning of the weight matrix, using Hebbian law, when a change is detected. FCoM evolution proceeds along Markovian principles, by calculating the next concept based on the current concept and edge values. The convergence to a stable limit cycle or fixed points determines the "hidden patterns" in the causal Web represented by the FCoM. It reflects the users' behavior as their knowledge of the Web increases with time, and they either learn new patterns or reinforce old ones. Unlike Markov models, FCoM is able to detect hidden rules related to certain behavior of users like learning from experience (previous mistakes).

6. Semantic web

Despite the huge amount of effort made to tackle the information from the Web, the area of Semantic Web is yet to be fully exploited. Although fuzzy sets offer a possibility to work with linguistic terms, there does not exist much literature on the fuzzy Semantic Web. A fusion between conceptual graph and fuzzy formalism is made in [3] to handle natural language from hypermedia and hypertext semantics. A fuzzy conceptual graph (FCG) is used for knowledge representation of the Semantic Web, and is available both for human and machine processing. Concepts of simple FCG, fuzzy entity and fuzzy attribute are defined. FCG projection is used for matching graphs, and this helps evaluate the degree of matching of the concept and relation pair between two FCGs. Absolute and relative quantifiers from natural language are represented by a lambda expression. For example, the pages about "Books written by A" and "Books written about A" are indexed, and the result of the query is adequately generated. The results have also been extended for indexing images.

However, developments regarding fuzzy conceptual graphs are at an early stage now. Their extension to very large documents, structured and unstructured types of documents with possibly very long sentences, multimedia Web pages, etc., are subjects of ongoing investigation.

7. Information retrieval

Typically, four main constituents can be identified in the process of information retrieval from the Internet.

- *Indexing*: generation of document representation
- *Querying*: expression of user preferences through natural language or terms connected by logical operators
- *Evaluation*: performance of matching between user query and document representation
- *User profile construction*: storage of terms representing user preferences, specially to enhance the system retrieval in future accesses by the user

Due to the presence of *multimedia* information repositories consisting of mixed media data, the information retrieved can be text as well as image document or a mixture of both. An example for mining multimedia data using fuzzy methods is given in Ref. [27]. We deal with text and image mining, separately, in the following two sections in order to highlight their importance in Web mining.

Fuzzy genes are used as intelligent agents for information retrieval from the Web [22]. This incorporates a hybridization of fuzzy sets with genetic algorithms (GAs) in the soft computing framework. User profiles are built from the user preferences, represented by chromosomes made up of a vector of fuzzy genes. Each chromosome is associated with a fitness corresponding to the system's belief in the hypothesis that the chromosome, as a query, represents the user's information needs. Every gene represents, by a fuzzy set, the number of occurrences that characterizes the documents considered relevant by the user. The fitness of the chromosome is adjusted based on the comparison between the user's evaluation of the retrieved documents and the score computed by the system. GAs are used to track the user's preferences and adapt the profile by incorporating her/his relevance feedback, while fuzzy sets handle the imprecision in the user's preferences and evaluation of the retrieved documents.

The architecture of the system has a parser module that collects words and performs statistics, a learning module that uses GAs for modeling the behavior, and an evaluation module that scores the documents. Chromosomes with fuzzy genes are present in the learning module, and use three types of membership functions in the universe of discourse involving term occurrence values. The classical GA operations like selection, crossover and mutation are applied. The fitness function is given as

$$f_i^j = f_i^{j-1} + P_i^j - L_i^j, \tag{9}$$

with

$$P_i^j = [(1 - (C_i^j - U^j))^2]^{w_1} \cdot [(S^{j-1} - U^j)^2]^{w_2},$$

where f_i^j is the fitness of the ith chromosome in the jth generation, P_i^j corresponds to the payoff (used for accumulation of fitness), L_i^j is the lifetax (used as a restriction on the chromosome if the payoff is continually low), and w_1, w_2 are the weights (lying in the unit interval) assigned by experts. Here C_i^j is the score of the ith chromosome in generation j, U^j represents the user's evaluation of the document and S^j is the population score computed by an arithmetic average of the best chromosomes. $C_i^j = 1/\left(l \cdot \sum_{h=1}^{l} \mu_r^h\right)$, where l represents the length of the chromosome and μ_r^h is the membership value of the hth gene of function type r. The performance is evaluated in terms of a prediction precision of the feedback from the user, recapitulation ability, and the classification predictability (CP) expressed as

$$CP = 1 - \sqrt{\frac{1}{d-1} \sum_{i=1}^{d} (S^i - U^i)^2}$$

with d as the number of documents in the evaluated set. The authors use a set of 20 documents with a manual assignment of scores from 0.1 to 0.9 in the search history. In the test procedure the system receives 65 documents split in 13 sets of 5 documents each.

A system composed of WebAgents and MetaWebAgents is described in [2]. While the WebAgents retrieve certain information, the job of the MetaWebAgents is to select the appropriate WebAgents that can access the necessary information. Two characteristics, viz., *reply-time* (answer time from any WebAgent after a MetaWebAgent requests for information) and the number of solutions, are used for measuring the performance of the WebAgents. A fuzzy distance between the performance values of the WebAgents is computed to classify their behavior. A Fuzzy Relational Algorithm (FRA) in the *If* ... *And* ... *Then* ... form of Mamdani-implication is used to store the required knowledge. The defuzzification is made using the Center of Gravity method. This system, involving search and meta-search, is used in real problems related to querying for airlines flights.

An approach in Ref. [20] uses fuzzy association thesaurus and query expansion for information retrieval. Fuzzy composition operations like max–min, max–product and sum–product are used for constructing the thesaurus. Interactive query expansion shows the user, upon initial query, a ranked list of documents suggested by the system based on the fuzzy relation composition. A measure of similarity helps select the correlated terms using queries with/without weight. The experiments use a collection of daily news in Chinese, with 981 homogeneous and 700 heterogeneous text documents.

An HTML document can be viewed as a structured entity, in which document subparts are identified by tags and each such subpart consists of text delimited by a distinct tag. A fuzzy representation of HTML documents is described in Ref. [25]. The HTML document is represented by a sum of fuzzy set terms

$$R(doc) = \sum_{t \in T} \frac{\mu_{doc}(t)}{t}, \tag{10}$$

where the importance of each term t in document doc is given by the membership value $\mu_{doc}(t) = \sum_{i=1}^{n} \left(\mu_{tag_i}(doc, t) * w_i\right) * g(IDF_t)$, w_i is the normalized importance weight associated with tag_i, $g(.)$ is a normalization function and IDF_t is the inverse document frequency. The significance of an index term is computed by weighting the occurrence of the term with the importance of the tag associated with it.

Recommending alternate queries during information retrieval is an important feature in a Web-based search engine, because users often do not know the exact terms to locate the information relevant to their interests. Fuzzy ontology is used for query refinement in a domain search engine named Personalizing Abstract Search Service (PASS) [33]. Fuzzy *narrower-* and *broader-than* term relations are defined using a fuzzy conjunction operator. Pruning of redundant relations is made by analyzing the sets of relations involving more than two terms. PASS is used to provide the abstracts of papers from the IEEE Transactions on Neural Networks journal.

8. Text mining

Considering the Web as a huge repository of distributed hypertext, the results from text mining can be included as a topic under Web mining or information retrieval. However we deal with the subject as a separate section, along with image mining in the following section, to highlight their importance in the current context.

A fuzzy decision tree that uses a fuzzy inductive learning to acquire relations from examples is presented in [32]. The authors generate a concept relation dictionary and a classification tree from a random set of daily business reports database of text classes concerning retailing. The algorithm is outlined below.

(i) Allocate a training example set to a new node and stack up the node
(ii) While (pick-up-a-node) do
(iii) Evaluate the allocation of the class to node; if the degree of certainty is acceptable go to step 4
(iv) (a) Create fuzzy sets (fuzzy class items) for each attribute
(b) Calculate evaluation values of the attributes
(c) Select the attribute with the best evaluation value, and allocate the attribute to the node
(d) Decompose the learning set into new subsets, and allocate each subset to a new node
(e) Create new branches, that connect the original and new nodes, and allocate each item to the corresponding branches
(f) Stack up the new nodes and go to step 2

In their experiments the authors use 10,000 evaluation examples, with a decision tree of maximum 90 nodes (37 intermediate and 53 terminal).

A key issue in text mining is keyword extraction. This allows the automatic categorization and classification of text documents. Keyword extraction can be done using clustering methods. relational alternating cluster estimation was used to automatically extract the 20 most relevant keywords from Internet documents in Ref. [31]. Using these keywords, a classification rate of more than 80% could be achieved. The results demonstrate a high efficiency, considering the fact that no additional tools like ontologies were used.

9. Image mining

The literature related to various data mining problems applied to collections of images is large. Fuzzy techniques have been successfully used for information retrieval. A nice overview of activities in this field can be found in [19]. However most authors only refer to a possible application of their algorithm to the fuzzy mining of images from the Web. The heterogeneous structure of the Web is generally composed of multimedia information, viz., text, image (in *jpeg*, *gif* or *mpeg* formats) and sound (in a hypermedia environment). This implies specific variations in clustering or association rule mining, involving different technologies including fuzzy sets. Among the few papers that describe results in direct connection with fuzzy image mining applications from the Web, we mention [7,8].

Video as the format of computer related material is becoming more and more common these days, and many Web pages involve small pieces of movies, video-clips or animation. Fuzzy time related queries are used in [7] to retrieve information inside a video. The queries are handled using Zadeh's principle of *computing with words*, that allows a human-friendly interface. A fuzzy time vocabulary is designed using notions such as time positioning, time descriptors and time relationship. There is a fuzzy partitioning of the universe of discourse for each measure. The authors use the relative temporal relationship framework in order to define the time events. Considering R to be the membership function of a time relationship and T the membership of an event, a new moment/event

is localized in the video by

$$R \circ T(x) = \max_{y} (R(x - y) \wedge T(y)), \tag{11}$$

where $\wedge$ is the fuzzy *And*/*min* operator. The system is implemented on a Java Search Engine.

Querying for a target image and retrieving it from Web and image databases, based on image similarity, is presented in [8]. A fuzzy c-means algorithm is used to cluster intrinsic image characteristics extracted from subregions of the image. A measure of similarity between pairs of images is determined in terms of the rotation-invariant attributes like color, texture and shape. Color is defined in terms of the hue, saturation, value representation [9] of the average color of the pixel in the region. Texture is represented in terms of the co-occurrence matrices in four directions involving Haralick's parameters [10]. For each database image, the system calculates its attribute matrix and does a whitening transformation before fuzzy clustering to store the representative centroids. When a target image is supplied, the system adopts a similar procedure in order to retrieve the most similar image (in terms of the stored centroid) from the database in response to a query. Experiments are reported on 70 images extracted from the CorelDRAW image database.

10. Conclusions

There is a growing awareness that, in practice, it is easy to discover a huge amount of information from the Web, where most of these patterns are actually obvious, redundant, and useless or uninteresting to the user. To prevent the user from being overwhelmed by a large number of uninteresting patterns, techniques are needed to identify only the useful/interesting patterns and present them to the user.

Fuzzy sets, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster.

In this article we have presented a survey on Web mining involving fuzzy sets and their hybridization with other soft computing tools. The Web mining taxonomy has been described. The individual functions like Web clustering, association rule mining, Web navigation, Web personalization and Semantic Web have been discussed in the fuzzy framework. Finally information retrieval along with text and image mining have been highlighted.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[2] D. Camacho, C. Hernandez, J.M. Molina, Information classification using fuzzy knowledge based agents, in: Proc. 10th IEEE Internat. Conf. on Fuzzy Systems, 2001, pp. 4:2575–4:2580.
[3] T.H. Cao, Fuzzy conceptual graph for the semantic web, in: Proc. 2001 BISC Internat. Workshop on Fuzzy Logic and the Internet (FLINT'2001), Berkeley, USA, August 2001.
[4] C.W. Chong, V. Ramachandran, C. Eswaran, Path optimization using fuzzy distance approach, in: Proc. 1999 IEEE International Fuzzy Systems Conf. (FUZZ-IEEE'99), Seoul, Korea, August 1999, pp. III:1771–III:1774.
[5] F. Crestani, G. Pasi, (Eds.), Soft Computing in Information Retrieval: Techniques and Application, vol. 50. Physica-Verlag, Heidelberg, 2000.

[6] R.N. Davé, Characterization and detection of noise in clustering, Pattern Recogn. Lett. 12 (1991) 657–664.
[7] M. Detyniecki, C. Seyrat, R. Yager, Interacting with web video objects, in: Proc. 18th Internat. Conf. of the North American Fuzzy Information Processing Society (NAFIPS'99), 1999, pp. 914–917.
[8] A. Filho, G.L.A. Mota, M.M.B.R. Vellasco, M.A.C. Pacheco, Query by image similarity using a fuzzy logic approach, in: Proc. Fourth Internat. Conf. on Computational Intelligence and Multimedia Applications (ICCIMA 2001), 2001, pp. 389–394.
[9] R.C. Gonzalez, R.E. Woods, Digital Image Processing, Addison-Wesley, Reading, MA, 1992.
[10] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Systems Man Cybernet. 3 (1973) 610–621.
[11] T.-P. Hong, K.-Y. Lin, S.-L. Wang, Mining linguistic browsing patterns in the World Wide Web, Soft Comput. 6 (2002) 329–336.
[12] Y.-P. Huang, Y.-C. Lee, L. Lin, An intelligent approach to mining the related websites, in: Proc. 20th NAFIPS Conf. Vancouver, Canada, July 2001, pp. 435–440.
[13] A. Joshi, R. Krishnapuram, Robust fuzzy clustering methods to support web mining, in: Proc. ACM-SIGMOD Workshop on Data Mining and Knowledge Discovery, August 1998.
[14] A. Joshi, R. Krishnapuram, On mining web access logs, in: Proc. ACM-SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2000, pp. 63–69.
[15] K.-J. Kim, S.-B. Cho, A personalized web search engine using fuzzy concept network with link structure, in: Proc. Joint Nineth IFSA World Congress and 20th NAFIPS Internat. Conf. 2001, pp. 1:81–1:86.
[16] R. Krishnapuram, A. Joshi, L. Yi, A fuzzy relative of the *k*-medoids algorithm with application to document and snippet clustering, in: Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ IEEE'99), Korea, August 1999, pp. 3:1281–3:1286.
[17] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Low complexity fuzzy relational clustering algorithms for web mining, IEEE Trans. Fuzzy Systems 9 (2001) 595–607.
[18] R. Krishnapuram, J.M. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Systems 1 (1993) 98–110.
[19] R. Kruse, A. Klose, Information mining with fuzzy methods: Trends and current challenges, in: Methods and Models in Automation and Robotics, Szczecin, Poland, 2002, pp. 117–120.
[20] H.-M. Lee, S.-K. Lin, C.-W. Huang, Interactive query expansion based on fuzzy association thesaurus for web information retrieval, in: Proc. the 10th IEEE Internat. Conf. on Fuzzy Systems, 2001, pp. 2:724–2:727.
[21] K.C. Lee, J.S. Kim, N.H. Chung, S.J. Kwon, Fuzzy cognitive map approach to web mining inference amplification, Expert Systems Appl. 22 (2002) 197–211.
[22] M.J. Martin-Bautista, H.L. Larsen, M.A. Vila, A fuzzy genetic algorithm approach to an adaptive information retrieval agent, J. Amer. Soc. Inform. Sci. 50 (1999) 760–771.
[23] G. Meghabghab, Mining user's web searching skills through fuzzy cognitive state map, in: Proc. Joint 9th IFSA World Congress and 20th NAFIPS Internat. Conf. 2001, pp. 1:429–1:434.
[24] S. Mitra, S.K. Pal, P. Mitra, Data mining in soft computing framework: a survey, IEEE Trans. Neural Networks 13 (2002) 3–14.
[25] A. Molinari, G. Pasi, A fuzzy representation of HTML documents for information retrieval systems, in: Proc. Fifth IEEE Internat. Conf. on Fuzzy Systems, 1996, pp. 1:107–1:112.
[26] O. Nasraoui, H. Frigui, R. Krishnapuram, A. Joshi, Extracting web user profiles using relational competitive fuzzy clustering, Internat. J. Artificial Intelligence Tools 9 (2000) 509–526.
[27] A. Nürnberger, A. Klose, Improving clustering and visualization of multimedia data using interactive user feedback, in: Internat. Conf. on Information Processing and Management of Uncertainty in Knowledge Based Systems, Annecy, France, 2002.
[28] W. Pedrycz, Conditional fuzzy *c*-means, Pattern Recognition Lett. 17 (1996) 625–632.
[29] T.A. Runkler, J. Bezdek, Web mining with relational clustering, Internat. J. Approx. Reason. 32 (2003) 217–236.
[30] T.A. Runkler, J.C. Bezdek, Alternating cluster estimation: a new tool for clustering and function approximation, IEEE Trans. Fuzzy Systems 7 (1999) 377–393.
[31] T.A. Runkler, J.C. Bezdek, Relational clustering for the analysis of internet newsgroups, in: O. Opitz, M. Schwaiger, (Eds.), Exploratory Data Analysis in Empirical Research, Proceedings of the 25th Annual Conference of the German Classification Society, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, 2002, pp. 291–299.

[32] S. Sakurai, Y. Ichimura, A. Suyama, R. Orihara, Inductive learning of a knowledge dictionary for a text mining system, in: L. Monostori, J. Vancza, M. Ali (Eds.), Proc. 14th Internat. Conf. Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE 2001), Lecture Notes in Artificial Intelligence, vol. 2070, Springer, Berlin, 2001, pp. 247–252.
[33] D.H. Widyantoro, J. Yen, Using fuzzy ontology for query refinement in a personalized abstract search engine, in: Proc. Joint Nineth IFSA World Congress and 20th NAFIPS Internat. Conf. Vancouver, Canada, July 2001, pp. 1:610–1:615.
[34] C. Wong, S. Shiu, S.K. Pal, Mining fuzzy association rules for web access case adaptation, in: Proc. Fourth Internat. Conf. on Case-Based Reasoning, Vancouver, Canada, July 2001, pp. 213–220.
[35] L.A. Zadeh, A new direction in AI: towards a computational theory of perceptions, AI Magazine 22 (2001) 73–84.