Text Mining using Fuzzy Association Rules

M.J. Martín-Bautista, D. Sánchez, J.M. Serrano, and M.A. Vila

Dept. of Computer Science and Artificial Intelligence. University of Granada.
C/ Periodista Daniel Saucedo Aranda s/n, 18071, Granada, Spain.
mbautis@decsai.ugr.es

Summary. In this paper, fuzzy association rules are used in a text framework. Text transactions are defined based on the concept of fuzzy association rules considering each attribute as a term of a collection. The purpose of the use of text mining technologies presented in this paper is to assist users to find relevant information. The system helps the user to formulate queries by including related terms to the query using fuzzy association rules. The list of possible candidate terms extracted from the rules can be added automatically to the original query or can be shown to the user who selects the most relevant for her/his preferences in a semi-automatic process.

1 Introduction

The data in the Internet is not organized in a consistent way due to a lack of an authority that supervises the adding of data to the web. Even inside each web site, there is a lack of structure in the documents. Although the use of hypertext would help us to give some homogeneous structure to the documents in the web, and therefore, to use data mining techniques for structure data, as it happens in relational databases, the reality is that nobody follows a unique format to write documents for the web. This represents a disadvantage when techniques such as data mining are applied. This leads us to use techniques specifically for text, as if we were not dealing with web documents, but with text in general, since all of them have an unstructured form.

This lack of homogeneity in the web makes the search process of information in the web by querying not so successful as navigators expect. This fact is due to two basic reasons: first, because the user is not able to represent her/his needs in query terms and second, because the answer set of documents is so huge that the user feels overwhelmed. In this work, we address the first problem of query specification.

Data mining techniques has been broadly applied to text, generating what is called Text Mining. Sometimes, the data mining applications requires the

V. Loia et al. (eds.), *Fuzzy Logic and the Internet*

user to know how to manage the tool. In this paper, the rules extracted from texts are not shown to the user specifically. The generated rules are applied to help user to refine the query but the user only see, considering a process non automatic completely, a list of candidate terms to add to the query.

When a user try to express her/his needs in a query, the terms that finally appear in the query are usually not very specific due to the lack of background knowledge of the user about the topic or just because in the moment of the query, the terms do not come to the user's mind. To help the user with the query construction, terms related to the words of a first query may be added to the query. From a first set of documents retrieved, data mining techniques are applied in order to find association rules among the terms in the set. The most accurate rules that include the original query words in the antecedent / consequent of the rule, are used to modify the query by automatically adding these terms to the query or, by showing to the user the related terms in those rules, so the modification of the query depends on the user's decision. A generalization or specification of the query will occur when the terms used to reformulate the query appear in the consequent / antecedent of the rule, respectively. This suggestion of terms helps the user to reduce the set of documents, leading the search through the desired direction.

This paper is organized as follows: in section 2, a summary of literature with the same purpose of this work is included. From section 3 to section 6, general theory about data mining and new proposals in the fuzzy framework are presented. Concretely, in section 3 and 4, the concepts of association rules, fuzzy association rules and fuzzy transactions are presented. In section 5, new measures for importance and accuracy of association rules are proposed. An algorithm to generate fuzzy association rules is presented in section 6. An application of this theory to text framework is proposed in section 7 and 8. The definition of text transactions is given in section 7, while the extracted text association rules are applied to query reformulation in an Information Retrieval framework in section 8. Finally, concluding remarks and future trends are given in section 9.

2 Related Work

One of the possible applications of Text Mining is the problem of query refinement, which has been treated from several frameworks. On the one hand, in the field of Information Retrieval, the problem has been defined as query expansion, and we can find several references with solutions to this problem. A good review in the topic can be found in [20]. On the other hand, techniques such as Data Mining, that have been applied successfully in the last decade in the field of Databases, have been also applied to solve some classical Information Retrieval problems such as document classification [33] and query optimization [46]. In this section, prior work in both frameworks, Information

Retrieval and Data Mining is presented, although the number of approaches presented in the first one, is much more extended than in the second one.

2.1 Previous Research in the Data Mining and Knowledge Discovery Framework

In general terms, the application of Data Mining and Knowledge Discovery techniques to text has been called Text Mining and Knowledge Discovery in Texts, respectively. The main difference to apply these techniques in a text framework is the special characteristics of text as unstructured data, totally different from databases, where mining techniques are usually applied and structured data is managed. Some general approaches about Text Mining and Knowledge Discovery in Texts can be found in [17], [21], [28],[31]

In this work, association rules applying techniques form data mining will be discovered as a process to select the terms to be added to the original query. Some other approaches can be found in this direction. In [46] a vocabulary generated by the association rules is used to improve the query. In [22] a system for Finding Associations in Collections of Text (FACT) is presented. The system takes background knowledge to show the user a simple graphical interface providing a query language with well-defined semantics for the discovery actions based on term taxonomy at different granularity levels. A different application of association rules but in the Information Retrieval framework can be found in [33] where the extracted rules are employed for document classification.

2.2 Previous Research in the Information Retrieval Framework

Several classifications can be made in this field according to the documents considered to expand the query, the selection of the terms to include in the query, and the way to include them. In [50] the authors make a study of expansion techniques based on the set of documents considered to analyze for the query expansion. If these documents are the corpus as a whole, from which all the queries are realized, then the technique is called *global analysis.* However, if the expansion of the query is performed based on the documents retrieved from the first query, the technique is denominated *local analysis,* and the set of documents is called *local set.* This local technique can also be classified into two types. On the one hand, *local feedback* adds common words from the top-ranked documents of the local set. These words are identified sometimes by clustering the document collection [3]. In this group we can include the relevance feedback process, since the user have to evaluate the top ranked documents from which the terms to be added to the query are selected. On the other hand, *local context analysis* [50], which combines global analysis and context local feedback to add words based on relationships of the top-ranked documents. The co-occurrences of terms are calculated based on passages (text windows of fixed size), as in global analysis, instead of complete

documents. The authors show that, in general, local analysis performs better than global one.

In our approach, both a global and a local technique are considered. On the one hand, association rules will be extracted from the corpus and applied to expand the query, and on the other hand, only the top ranked documents will be considered to carry out the same process.

Regarding the selection of the terms, some approaches use several techniques to identify terms that should be added to the original query. The first group is based on their association relation by co-occurrence to query terms [47]. Instead of simply terms, in [50] find co-occurrences of concepts given by noun groups with the query terms. Some other approaches based on concept space are [12]. The statistical information can be extracted from a clustering process and ranking of documents from the local set, as it is shown in [13] or by similarity of the top-ranked documents [36]. All these approaches where a co-occurrence calculus is performed has been said to be suitable for construct specific knowledge base domains, since the terms are related, but it can not be distinguished how [8]. The second group searches terms based on their similarity to the query terms, constructing a similarity term thesaurus [41]. Other approaches in this same group, use techniques to find out the most discriminatory terms, which are the candidates to be added to the query. These two characteristics can be combined by first calculating the nearest neighbors and second by measuring the discriminatory abilities of the terms [38]. The last group is formed by approaches based on lexical variants of query terms extracted from a lexical knowledge base such as Wordnet [35]. Some approaches in this group are [49], and [8] where a semantic network with term hierarchies is constructed. The authors reveal the adequacy of this approach for general knowledge base, which can be identified in general terms with global analysis, since the set of documents from which the hierarchies are constructed is the corpus, and not the local set of a first query. Previous approaches with the idea of hierarchical thesaurus can be also found in the literature, where an expert system of rules interprets the user's queries and controls the search process [25].

In our approach, since we are performing a local analysis, fuzzy association rules are used as a technique to find relations among the terms. The aim of the use of this technique is detail and give more information by means of inclusion relations about the connection of the terms, avoiding the inherent statistical nature of systems using co-occurrences as relationships among terms, which performance is only good where the terms selected to expand the query comes from relevant documents of the local set [27]. Previous good results of the use of fuzzy association rules in comparison with crisp association rules and pure statistical methods have been presented in the relational database framework [4], [16], [18].

As for the way to include the terms in the query, we can distinguish between automatic and semi-automatic query expansion [41]. In the first group, the selected terms can substitute or be added to the original query without

the intervention of the user [10], [25], [47]. In the second group, a list of candidate terms is shown to the user, which makes the selection [48]. Generally, automatic query expansion is used in local analysis and semi-automatic query expansion is more adequate for global analysis, since the user has to decide from a broad set of terms from the corpus which are more related to her/his needs.

3 Association Rules

The obtaining and mining of association rules is one of the main research problems in data mining framework [1]. Given a database of transactions, where each transaction is an itemset, the obtaining of association rules is a process guided by the constrains of *support* and *confidence* specified by the user. Support is the percentage of transactions containing an itemset, calculated in a statistical manner, while confidence measures the strength of the rule. Formally, let T be a set of transactions containing items of a set of items I. Let us consider two itemsets $I_1, I_2 \subseteq I$, where $I_1 \cap I_2 = \emptyset$. A rule $I_1 \Rightarrow I_2$ is an implication rule meaning that the apparition of itemset I_1 implies the apparition of itemset I_2 in the set of transactions T. I_1 and I_2 are called antecedent and consequent of the rule, respectively. Given a support of an itemset noted by $supp(I_k)$, and the rule $I_1 \Rightarrow I_2$, the support and the confidence of the rule noted by $Supp\,(I_1 \Rightarrow I_2)$ and $Conf\,(I_1 \Rightarrow I_2)$, respectively, are calculated as follows:

$$Supp\,(I_1 \Rightarrow I_2) = supp\,(I_1 \cup I_2) \tag{1}$$

$$Conf\,(I_1 \Rightarrow I_2) = \frac{supp\,(I_1 \cup I_2)}{supp\,(I_1)} \tag{2}$$

The constrains of minimum support and minimum confidence are established by the user with two threshold values: *minsupp* for the support and *minconf* for the confidence. A *strong rule* is an association rule whose support and confidence are greater that thresholds minsupp and minconf, respectively. Once the user has determined these values, the process of obtaining association rules can be decomposed in two different steps:

Step 1.- Find all the itemsets that have a support above threshold *minsupp*. These itemsets are called *frequent itemsets.*

Step 2.- Generate the rules, discarding those rules below threshold *minconf.*

The rules obtained with this process are called boolean association rules in the sense that they are generated from a set of boolean transactions where the values of the tuples are 1 or 0 meaning that the attribute is present in the transaction or not, respectively.

The application of these processes is becoming quite valuable to extract knowledge in business world. This is the reason why the examples given in the literature to explain generation and mining processes of association rules are based, generally, on sale examples of customers shopping. One of the most famous examples of this kind is the market basket example introduced in [1], where the basket of customers is analyzed with the purpose of know the relation among the products that everybody buy usually. For instance, a rule with the form *bread⇒milk* means that everybody that buy bread also buy milk, that is, the products bread an milk usually appears together in the market basket of customers. We have to take into account, however, that this rule obtaining has an inherent statistical nature, and is the role of an expert the interpretation of such rules in order to extract the knowledge that reflects human behavior. This fact implies the generation of easy rules understandable for an expert of the field described by the rules, but probably with no background knowledge of the data mining concepts and techniques.

The consideration of rules coming from real world implies, most of the times, the handling of uncertainty and quantitative association rules, that is, rules with quantitative attributes such as, for example, the age or the weight of a person. Since the origin of these rules is still considered as a set of boolean transactions, a partition into intervals of the quantitative attributes is needed in order to transform the quantitative problem in a boolean one. The discover of suitable intervals with enough support is one of the problems to solve in the field proposed and addressed in several works [14], [23], [39]. In the first work, an algorithm to deal with non binary attributes, considering all the possible values that can take the quantitative attributes to find the rules. In the last two works, however, the authors strengthen the suitability of the theory of fuzzy sets to model quantitative data and, therefore, deal with the problem of quantitative rules. The rules generated using this theory are called fuzzy association rules, and their principal bases as well as the concept of fuzzy transactions are presented in next section.

4 Fuzzy Transactions and Fuzzy Association Rules

Fuzzy association rules are defined as those rules that associate items of the form *(Attribute, Label)*, where the label has an internal representation as fuzzy set over the domain of the attribute [18]. The obtaining of these rules comes from the consideration of fuzzy transactions. In the following, we present the main and features related to fuzzy transactions and fuzzy association rules. The complete model and applications of these concepts can be found in [14].

4.1 Fuzzy Transactions

Given a finite set of items I, we define a fuzzy transaction as any nonempty fuzzy subset $\tilde{\tau} \subseteq I$. For every $i \in I$, the membership degree of i in a fuzzy

transaction $\tilde{\tau}$ is noted by $\tilde{\tau}(i)$. Therefore, given an itemset $I_o \subseteq I$, we note $\tilde{\tau}(I_0)$ the membership degree of I_0 to a fuzzy transaction $\tilde{\tau}$. We can deduce from this definition that boolean transactions are a special case of fuzzy transactions. We call FT-set the set of fuzzy transactions, remarking that it is a crisp set.

A set of fuzzy transactions FT-set is represented as a table where columns and rows are labeled with identifiers of items and transactions, respectively. Each cell of a pair *(transaction, itemset)* of the form $(I_0, \tilde{\tau}_j)$ contains the membership degree of I_0 in $\tilde{\tau}_j$, noted $\tilde{\tau}_j(I_0)$ and defined as

$$\tilde{\tau}(I_0) = \min_{i \in I_0} \tilde{\tau}(i) \tag{3}$$

The representation of an item I_0 in a FT-set T based in I is represented by a fuzzy set $\tilde{\Gamma}_{I_0} \subseteq T$, defined as

$$\tilde{\Gamma}_{I_0} = \sum_{\tilde{\tau} \in T} \tilde{\tau}(I_0)/\tilde{\tau} \tag{4}$$

4.2 Fuzzy Association Rules

A fuzzy association rule is a link of the form $A \Rightarrow B$ such that $A, B \subset I$ and $A \cap B = \emptyset$, where A is the antecedent and B is the consequent of the rule, being both of them fuzzy itemsets. An ordinary association rule is a fuzzy association rule. The meaning of a fuzzy association rule is, therefore, analogous to the one of an ordinary association rule, but the set of transactions where the rule holds, which is a FT-set. If we call $\tilde{\Gamma}_A$ and $\tilde{\Gamma}_B$ the degrees of attributes A and B in every transaction $\tilde{\tau} \in T$, we can assert that the rule $A \Rightarrow B$ holds with totally accuracy in T when $\tilde{\Gamma}_A \subseteq \tilde{\Gamma}_B$.

5 Importance and Accuracy Measures for Fuzzy Association Rules

The imprecision latent in fuzzy transactions makes us consider a generalization of classical measures of support and confidence by using approximate reasoning tools. One of these tools is the evaluation of quantified sentences presented in [51]. A quantified sentence is and expression of the form "Q of F are G", where F and G are two fuzzy subsets on a finite set X, and Q is a relative fuzzy quantifier. We focus on quantifiers representing fuzzy percentages with fuzzy values in the interval [0,1] such as "most", "almost all" or "many". These quantifiers are called *relative quantifiers.*

Let us consider Q_M a quantifier defined as $Q_M(x) = x$, $\forall x \in [0, 1]$. We define the **support of an itemset** I_0 in an FT-set T as the evaluation of the quantified sentence,

$$Q_M \; of \; T \; are \; \tilde{\Gamma}_{I_0} \tag{5}$$

while the **support of a rule** $A \Rightarrow B$ in T is given by the evaluation of

$$Q_M \; of \; T \; are \; \tilde{\Gamma}_{A \cup B} = Q_M \; of \; T \; are \; \tilde{\Gamma}_A \cap \tilde{\Gamma}_B \tag{6}$$

and its confidence is the evaluation of

$$Q_M \; of \; \tilde{\Gamma}_A \; are \; \tilde{\Gamma}_B \tag{7}$$

We evaluate the sentences by means of method GD presented in [19]. To evaluate the sentence "Q of F are G", a compatibility degree between the relative cardinality of G with respect to F and the quantifier is represented by $GD_Q\,(G/F)$ and defined as

$$GD_Q\,(G/F) = \sum_{\alpha_i \in \Delta(G/F)} (\alpha_i - \alpha_{i+1}) \cdot Q\left(\frac{|(G \cap F)_{\alpha_i}|}{|F_{\alpha_i}|}\right) \tag{8}$$

where $\Delta\,(G/F) = \Lambda\,(G \cap F) \cup \Lambda\,(F)$, $\Lambda\,(F)$ being the level set of F, and $\Delta\,(G/F) = \{\alpha_1, \ldots, \alpha_p\}$ with $\alpha_i > \alpha_{i+1}$ for every $i \in \{1, \ldots p\}$. The set F is assumed to be normalized. If not, F is normalized and the normalization factor is applied to $G \cap F$.

We must point out, moreover, that when we are dealing with crisp data in a T-set T , the evaluation of sentences are the ordinary measures of support and confidence of crisp association rules. Therefore, the evaluation of sentence "Q of F are G" is

$$Q\left(\frac{|F \cap G|}{|F|}\right) \tag{9}$$

when F and G are crisp. The GD method verifies this property. For more details, see [19]. We can interpret the ordinary measures of confidence and support as the degree to which the confidence and support of an association rule is Q_M. Other properties of this quantifier can be seen in [14].

This generalization of the ordinary measures allow us, using Q_M , provide an accomplishment degree, basically. Hence, for fuzzy association rules we can assert

$$Q_M \tau \in T, \; A \Rightarrow B \tag{10}$$

5.1 Certainty as a New Measure for Rule Accuracy

We propose the use of certainty factors to measure the accuracy of association rules. A previous study can be found in [15]. Certainty factors were developed as a model for the representation of uncertainty and reasoning in rule-based systems [45], although they have been used in knowledge discovery too [24].

We define certainty factor (CF) of a fuzzy association rule $A \Rightarrow B$ based on the value of the confidence of the rule. If $Conf\,(A \Rightarrow B) > supp\,(B)$ the value of the factor is given by expression (11); otherwise, is given by expression (12), considering that if $supp(B)$=1, then $CF\,(A \Rightarrow B) = 1$ and if $supp(B)$=0, then $CF\,(A \Rightarrow B) = -1$

$$CF\,(A \Rightarrow B) = \frac{Conf\,(A \Rightarrow B) - supp\,(B)}{1 - supp\,(B)} \tag{11}$$

$$CF\,(A \Rightarrow B) = \frac{Conf\,(A \Rightarrow B) - supp\,(B)}{supp\,(B)} \tag{12}$$

We demonstrated in [7] that certainty factors verify the three properties by [29]. From now on, we shall use certainty factors to measure the accuracy of a fuzzy association rule. We consider a fuzzy association rule as strong when its support and certainty factor are greater than thresholds *minsupp* and *minCF*, respectively.

6 Generation of Fuzzy Association Rules

Several approaches can be found in the literature where efficient algorithms for association rule generation like Apriori and AprioriTid [2], OCD [34], SETM [30], DHP [37], DIC [9], FP-Growth [26] and TBAR [6], have been presented. Most of them include and describe the process of generating fuzzy association rules with two basic steps, as we mentioned in Sect. 3: the generation of frequent itemsets and the obtaining of the rules, with their associated grades of support and confidence. As we are considering fuzzy association rules, in Algorithm 1, we show a process to find the frequent itemsets. For this purpose, the transactions are analyzed one by one and the itemsets whose support is greater than threshold minsupp are selected. The items are processed ordered by size. First 1-itemsets, next 2-itemsets and so on. The variable l stores the actual size. The set L_l stores the l-itemsets that are being analyzed and, at the end, it stores the frequent l-itemsets.

In order to deal with fuzzy transactions, we need to store the difference between the cardinality of every α-cut of $\tilde{\Gamma}_{I_0}$ and the cardinality of the corresponding strong α-cut, $\alpha \in [0, 1]$, for all considered itemsets I_0. Specifically

$$\left|\left(\tilde{\Gamma}_{I_0}\right)_{\alpha}\right| - \left|\left(\tilde{\Gamma}_{I_0}\right)_{\alpha+}\right|$$

where $\left(\tilde{\Gamma}_{I_0}\right)_{\alpha} = \left\{\tilde{\tau} \in T \;\middle|\; \tilde{\Gamma}_{I_0}(\tilde{\tau}) \geq \alpha\right\}$ and $\left(\tilde{\Gamma}_{I_0}\right)_{\alpha+} = \left\{\tilde{\tau} \in T \;\middle|\; \tilde{\Gamma}_{I_0}(\tilde{\tau}) > \alpha\right\}$

We use a used a fixed number of k equidistant α-cuts, (specifically k=100, although we a lesser value would be sufficient). By this information, we obtain the fuzzy cardinality of the representation of the items, which is stored in an

array V_{I_0}. This array can be easily obtained from an FT-set by adding 1 to $V_{I_0}\left(\tilde{I}_{I_0}(\tilde{\tau})\right)$ for every itemset I_0 each time a transaction $\tilde{\tau}$ is considered. The function *R(x,k)* maps the real value x to the nearest value in the set of k equidistant levels we are using.

The procedure *CreateLevel(i, L)* generates a set of i-itemsets such that every proper subset with i-1 items is frequent (i.e. is in L_{i-1}). Since every proper subset of a frequent itemset is also a frequent itemset, with this procedure we avoid analyzing itemsets that do not verify this property, saving space and time.

Algorithm 1 Basic algorithm to find frequent itemsets in a FT-set T

Input: a set I of items and an a FT-set T based on I.
Output: a set of frequent itemsets F.

1. {Initialization}
   a) Create an array $V_{\{i\}}$of size $k+1$ for every $i \in I$
   b) $L_1 \leftarrow \{\{i\} \mid i \in I\}$
   c) F=0
   d) $l \leftarrow 1$
2. Repeat until $l > |I|$ or $L_l = 0$
   a) For every $\tilde{\tau} \in T$
      i. For every $I_* \in L_l$
         A. $V_{I_*}\left(R\left(\tilde{I}_{I_*}(\tilde{\tau}), k\right)\right) \leftarrow V_{I_*}\left(R\left(\tilde{I}_{I_*}(\tilde{\tau}), k\right)\right) + 1$
   b) For every $I_* \in L_l$
      i. Calculate $GD_Q\left(\tilde{I}_{I_*}/T\right)$
      ii. If $GD_Q\left(\tilde{I}_{I_*}/T\right) < minsupp \times |T|$
         A. $L_l \leftarrow L_l \backslash \{I_*\}$
         B. Free the memory used by V_{I_*}
   c) {Variables updating}
      i. $F = F \cup L_l$
      ii. $L_{l+1} \leftarrow CreateLevel(l+1, L_l)$
      iii. $l \leftarrow l + 1$
3. Return(F)

The complexity of this algorithm is an exponential function of the number of items. The hidden constant is increased in a factor that depends on k as this value affects the size of the arrays V. For more details of the algorithm, see [14]

Once we have obtained the frequent itemsets with the former algorithm, we obtain the confidence by calculating $GD_Q(B/A)$ from V_A and $V_{A \cup B}$. From confidence and support of the consequent, both available, we obtain the certainty factor of the rules. Finally, we can identifier the strong rules by analyzing the values of support and certainty for the rules.

7 Text Mining for Information Access

The main problem when the general techniques of data mining are applied to text, is to deal with unstructured data, in comparison to structured data coming from relational databases. Therefore, with the purpose to perform a knowledge discovery process, we need to obtain some kind of structure in the texts. Different representations of text for association rules extraction have been considered: bag of words, indexing keywords, term taxonomy and multi-term text phrases [17]. In our case, we use automatic indexing techniques coming from Information Retrieval [44]. We represent each document by a set of terms with a weight meaning the presence of the term in the document. Some weighting schemes for this purpose can be found in [43]. One of the more successful and more used representation schemes is the *tf-idf* scheme, which takes into account the term frequency and the inverse document frequency, that is, if a term occurs frequently in a document but infrequently in the collection, a high weight will be assigned to that term in the document. This is the scheme we consider in this work. The algorithm to get the representation by terms and weights of a document d_i can be detailed by the known following steps in Algorithm 2.

Algorithm 2 Basic algorithm to obtain the representation of documents in a collection

Input: a set of documents $D = \{d_1, \ldots d_n\}$.
Output: a representation for all documents in D.

1. Let $D = \{d_1, \ldots d_n\}$ be a collection of documents
2. Extract an initial set of terms S from each document $d_i \in D$
3. Remove stop words
4. Apply stemming (*via* Porter's algorithm [40])
5. The representation of d_i obtained is a set of keywords $\{t_1, \ldots, t_m\} \in S$ with their associated weights $\{w_1, \ldots, w_m\}$

We must point out that, as it has been commented and shown in [21], [42], standard Text Mining usually deal with categorized documents, in the sense of documents which representation is a set of *keywords*, that is, terms that really describe the content of the document. This means that usually a full text is not considered and its description is not formed by all the words in the document, even without stop words, but also by keywords. The authors justify the use of keywords because of the appearing of useless rules. Some additional commentaries about this problem regarding the poor discriminatory power of frequent terms can be found in [38], where the authors comment the fact that the expanded query may result worst performance than the original one due to the poor discriminatory ability of the added terms.

However, in document collections where the categorization is not always available, full text is necessary to be considered as starting point. Additionally, special pre-processing tasks of term extraction and selection can be applied to get keywords in these collections. We are not referring here to statistical counts of term occurrences and assigning of weighting schemes such as the *tf-idf* one, but to more elaborated methods that imply additional time process, such as term taxonomy construction, thesauri or controlled vocabulary.

Nevertheless, in dynamic environments or systems where the response-time is important, the application of this pre-processing stage may not be suitable. This is the case of the problem we deal with in this work, the query refinement in Internet, where an automatic process would be necessary. Two time constraints have to be into account: first, the fact that not all web documents have identified keywords when is retrieved, or if they have, we do not have the guarantee that the keywords are appropriate in all the cases. Second, in the case of query refinement, information rule must be shown to the user online, that is, while she/he is query the system. Therefore, instead of improve document representation in this situation, we can improve the rule obtaining process. The use of alternative measures of importance and accuracy such as the ones presented in Sect. 5 is considered in this work in order to avoid the problem of non appropriate rule generation.

7.1 Text Transactions

From a collection of documents $D = \{d_1, \ldots, d_n\}$ we can obtain a set of terms $I = \{t_1, \ldots, t_m\}$ which is the union of the keywords for all the documents in the collection. The weights associated to these terms are represented by $W = \{w_1, \ldots, w_m\}$. Therefore, for each document d_i, we consider an extended representation where a weight of 0 will be assigned to every term appearing in some of the documents of the collection but not in d_i.

Considering these elements, we can define a *text transaction* $\tau_i \in T$ as the extended representation of document d_i. Without loosing generalization, we can write $T = \{d_1, \ldots, d_n\}$. However, as we are dealing with fuzzy association rules, we will consider a fuzzy representation of the presence of the terms in documents, by using the normalized tf-idf scheme [32]. Analogously to the former case, we can define a set of *fuzzy text transactions* $FT = \{d_1, \ldots, d_n\}$, where each document d_i corresponds to a fuzzy transaction $\tilde{\tau}_i$, and where the weights $W = \{w_1, \ldots, w_m\}$ of the keyword set $I = \{t_1, \ldots, t_m\}$ are fuzzy values.

8 Query Reformulation Procedure

The purpose of this work is to provide a system with a query reformulation ability in order to improve the retrieval process. We represent the query a $Q = \{q_1, \ldots, q_m\}$ with associated weights $P = \{p_1, \ldots, p_m\}$. To obtain a

relevance value for each document, the query representation is matched to each document representation, obtained as explained in Algorithm 2. If a document term does not appear in the query, its value will be assumed as 0. The considered operators and measures are the one from the generalized Boolean model with fuzzy logic [11].

The user's initial query generates a set of ranked documents. If the top-ranked documents do not satisfy user's needs, the query improvement process starts. From the retrieved set of documents, association relations are found. As we explain in Sect.2, two different approaches can be considered at this point: an automatic expansion of the query or a semi-automatic expansion, based on the intervention of the user in the selection process of the terms to be added to the query. The complete process in both cases is detailed in the following:

Case 1: Automatic query reformulation process

1. The user queries the system
2. A first set of documents is retrieved
3. From this set, the representation of documents is extracted following Algorithm 2 and fuzzy association rules are generated following Algorithm 1 and the extraction rule procedure.
4. The terms co-occurring in the rules with the query terms are added to the query.
5. With the expanded query, the system is queried again.

Case 2: Semi-automatic query reformulation process

1. The user queries the system
2. A first set of documents is retrieved
3. From this set, the representation of documents is extracted following Algorithm 2 and fuzzy association rules are generated following Algorithm 1 and the extraction rule procedure
4. The terms co-occurring in the rules with the query terms are shown to the user
5. The user selects those terms more related to her/his needs
6. The selected terms are added to the query, which is used to again to query the system

We must point out that, in both cases, the obtained association rules conform a knowledge base specific for the domain of the first query. Where several queries are performed, a broader knowledge base may be constructed, so original queries will be enriched with more terms as the time passes. However, the obtaining of a huge knowledge-based from iterated query expansions even in different domains probably can not be used for any query in a successful way, since additional semantic relation information should be also take into account

in order to get a general knowledge-base. As a future proposal, we can think about combine both domain-specific knowledge base and general knowledge base, looking at the terms appearing in association rules together with query terms appear, and searching in a general knowledge-base additional terms, WordNet [35], for instance, with a semantic relation with all the terms in the rule. Some further discussion about this point can be found in [8]

8.1 Generalization and Specialization of a Query

Once the first query is constructed, and the association rules are extracted, we make a selection of rules where the terms of the original query appear. However, the terms of the query can appear in the antecedent or in the consequent of the rule. If a query term appears in the antecedent of a rule, and we consider the terms appearing in the consequent of the rule to expand the query, a generalization of the query will be carried out. Therefore, a generalization of a query gives us a query on the same topic as the original one, but looking for more general information. However, if query term appears in the consequent of the rule, and we reformulate the query by adding the terms appearing in the antecedent of the rule, then a specialization of the query will be performed, and the precision of the system should increase. The specialization of a query looks for more specific information than the original query but in the same topic. In order to obtain as much documents as possible, terms appearing in both sides of the rules can also be considered.

9 Conclusion and Future Work

In this work, an application of traditional data mining techniques in a text framework is proposed. Classical transactions in data mining are first extended to the fuzzy transactions, proposing new measures to measure the accuracy of a rule. Text transactions are defined based on fuzzy transactions, considering that each transaction correspond to a document representation. The set of transactions represents, therefore, a document collection from which the fuzzy association rules are extracted. One of the applications of this process is to solve the problem of refinement of a query, very well known in the field of Information Retrieval. A list of terms extracted from the fuzzy association rules related to the terms in the query can be automatically added to the original query to optimize the search. This process can also be done with the user intervention, selecting the terms more related to her/his preferences.

As future work, we will implement the application of the model to this query reformulation procedure and compare the results with other approaches to query refinement coming from Information Retrieval.

References

1. Agrawal R, Imielinski T, Swami A (1993) Mining Association Rules between Set of Items in Large Databases. Proc. of the 1993 ACM SIGMOD Conference, pp 207-216
2. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. Proc. Of the 20^{th} VLDB Conference, pp 478-499
3. Attar R, Fraenkel AS (1977) Local Feedback in Full-Text Retrieval Systems. Journal of the Association for Computing Machinery 24(3):397-417
4. Au WH, Chan KCC (1998) An effective algorithm for discovering fuzzy rules in relational databases. Proc. Of IEEE International Conference on Fuzzy Systems, vol II, pp 1314-1319
5. Baeza-Yates R, Ribeiro-Nieto B (1999) Modern Information Retrieval, Addison-Wesley, USA
6. Berzal F, Cubero JC, Marín N, Serrano JM (2001) TBAR: An efficient method for association rule mining in relational databases. Data and Knowledge Engineering 37(1):47-84
7. Berzal F, Blanco I, Sánchez, Vila MA (2002) Measuring the Accuracy and Importance of Association Rules: A New Framework. Intelligent Data Analysis 6:221-235
8. Bodner RC, Song F (1996) Knowledge-Based Approaches to Query Expansion in Information Retrieval. In: McGalla G (ed) Advances in Artificial Intelligence pp 146-158. Springer, New York
9. Brin S, Motwani JD, Ullman JD, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. SIGMOD Record 26(2):255-264
10. Buckley C, Salton G, Allan J, Singhal A (1993) Automatic Query Expansion using SMART: TREC 3". *Proc. of the* 3^{rd} Text Retrieval Conference. NIST Special Publication 500-225, pp 69-80
11. Buell DA, Kraft DH (1981) Performance Measurement in a Fuzzy Retrieval Environment. Proceedings of the Fourth International Conference on Information Storage and Retrieval, ACM/SIGIR Forum 16(1): 56-62, Oakland, CA
12. Chen H, Ng T, Martinez J, Schatz BR (1997) A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. Journal of the American Society for Information Science 48(1):17-31
13. Croft WB, Thompson RH (1987) I^3R: A New Approach to the Design of Document Retrieval Systems. Journal of the American Society for Information Science 38(6):389-404
14. Delgado M, Marín N, Sánchez D, Vila MA (2001). Fuzzy Association Rules: General Model and Applications. IEEE Transactions of Fuzzy Systems (accepted)
15. Delgado M, Martín-Bautista MJ, Sánchez D, Vila MA (2000). Mining strong approximate dependences from relational databases. Proc. Of IPMU 2000 2:1123-1130. Madrid, Spain
16. Delgado M, Martín-Bautista MJ, Sánchez D, Vila MA (2001) Mining association rules with improved semantics in medical databases. Artificial Intelligence in Medicine 21:241-245
17. Delgado M, Martín-Bautista MJ, Sánchez D, Vila MA (2002) Mining Text Data: Special Features and Patterns. Proc. of EPS Exploratory Workshop on

Pattern Detection and Discovery in Data Mining, pp 140-153. Imperial College Londres, UK
18. Delgado M, Sánchez D, Vila MA (2000) Acquisition of fuzzy association rules from medical data. In Barro S, Marín R (eds) Fuzzy Logic in Medicine. Physica-Verlag
19. Delgado M, Sánchez D, Vila MA (2000) Fuzzy cardinality based evaluation of quantified sentences. International Journal of Approximate Reasoning 23:23-66
20. Efthimiadis E (1996) Query Expansion. Annual Review of Information Systems and Technology 31:121-187
21. Feldman R, Fresko M, Kinar Y, Lindell Y, Liphstat O, Rajman M, Schler Y, Zamir O (1998) Text Mining at the Term Level. Proc. of the 2^{nd} European Symposium of Principles of Data Mining and Knowledge Discovery, pp 65-73
22. Feldman R, Hirsh H (1996) Mining associations in text in the presence of Background Knowledge. Proc. of the Second International Conference on Knowledge Discovery from Databases
23. Fu AW, Wong MH, Sze SC, Wong WC, Wong WL, Yu WK (1998) Finding Fuzzy Sets for the Mining of Fuzzy Association Rules for Numerical Attributes. Proc. of Int. Symp. on Intelligent Data Engineering and Learning (IDEAL'98), pp 263-268, Hong Kong
24. Fu LM, Shortliffe EH (2000) The application of certainty factors to neural computing for rule discovery. IEEE Transactions on Neural Networks 11(3):647-657
25. Gauch S, Smith JB (1993) An Expert System for Automatic Query Reformulation. Journal of the American Society for Information Science 44(3):124-136
26. Han J, Pei J, Yin Y (2000)Mining frequent patterns without candidate generation. Proc. ACM SIGMOD Int. Conf. On Management of Data, pp 1-12. Dallas, TX, USA
27. Harman D (1988) Towards interactive query expansion. Proc. of the Eleventh Annual International ACMSIGIR Conference on Research and Development in Information Retrieval pp 321-331. ACM Press
28. Hearst M (1999) Untangling Text Data Mining. Proc. of the 37^{th} Annual Meeting of the Association for Computational Linguistics (ACL'99). University of Maryland
29. Hearst M (2000) Next Generation Web Search: Setting our Sites. IEEE Data Engineering Bulletin, Special issue on Next Generation Web Search, Gravano L (ed)
30. Houtsma M, Swami A (1995) Set-oriented mining for association rules in relational databases. Proc. Of the 11^{th} International Conference on Data Engineering pp 25-33.
31. Kodratoff Y (1999) Knowledge Discovery in Texts: A Definition, and Applications. In: Ras ZW, Skowron A (eds) Foundation of Intelligent Systems, Lectures Notes on Artificial Intelligence 1609. Springer Verlag
32. Kraft D, Petry FE, Buckles BP, Sadasivan T (1997) Genetic Algorithms for Query Optimization in Information Retrieval: Relevance Feedback. In: Sanchez E, Shibata T, Zadeh LA, (eds) Genetic Algorithms and Fuzzy Logic Systems, Advances in Fuzziness: Applications and Theory 7:157-173, World Scientific
33. Lin SH, Shih CS, Chen MC, Ho JM, Ko MT, Huang YM (1998) Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach. Proc. of ACM/SIGIR'98 pp 241-249. Melbourne, Australia

34. Mannila H, Toivonen H, Verkamo I (1994) Efficient algorithms for discovering association rules. Proc. Of AAAI Workshop on Knowledge Discovery in Databases pp 181-192
35. Miller G (1990) WordNet: An on-line lexical database. International Journal of Lexicography 3(4)
36. Mitra M, Singhal A, Buckley C (1998) Improving Automatic Query Expansion. Proc. Of ACM SIGIR pp 206-214. Melbourne, Australia
37. Park JS, Chen MS, Yu PS (1995) An effective hash based algorithm for mining association rules. SIGMOD Record 24(2):175-186
38. Peat HJ, Willet P (1991) The limitations of term co-occurrence Data for Query Expansion in Document Retrieval Systems. Journal of the American Society for Information Science 42(5):378-383
39. Piatetsky-Shapiro G (1991) Discovery, Analysis, and Presentation of Strong Rules. In: Piatetsky-Shapiro G, Frawley WJ (eds) Knowledge Discovery in Databases, AAAI/MIT Press
40. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130-137
41. Qui Y, Frei HP (1993) Concept Based Query Expansion. Proc. Of the Sixteenth Annual International ACM-SIGIR'93 Conference on Research and Development in Information Retrieval pp 160-169
42. Rajman M, Besançon R (1997) Text Mining: Natural Language Techniques and Text Mining Applications. Proc. of the 3^{rd} International Conference on Database Semantics (DS-7)Chapam & Hall IFIP Proceedings serie
43. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing and Management 24(5):513-523
44. Salton G, McGill MJ (1983) Introduction to Modern Information Retrieval. McGraw-Hill
45. Shortliffe E, Buchanan B (1975) A model of inexact reasoning in medicine. Mathematical Biosciences 23:351-379
46. Srinivasan P, Ruiz ME, Kraft DH, Chen J (2001) Vocabulary mining for information retrieval: rough sets and fuzzy sets. Information Processing and Management 37:15-38
47. Van Rijsbergen CJ, Harper DJ, Porter MF (1981) The selection of good search terms. Information Processing and Management 17:77-91
48. Vélez B, Weiss R, Sheldon MA, Gifford DK (1997) Fast and Effective Query Refinement. Proc. Of the 20^{th} ACM Conference on Research and Development in Information Retrieval (SIGIR'97). Philadelphia, Pennsylvania
49. Voorhees EM (1994)Query expansion using Lexical-Semantic Relations. ACM SIGIR pp 61-70
50. Xu J, Croft WB (1996) Query Expansion Using Local and Global Document Analysis. Proc. of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval pp 4-11
51. Zadeh LA (1983) A computational approach to fuzzy quantifiers in natural languages. Computing and Mathematics with Applications 9(1):149-184