Association Rule Extraction for Text Mining

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

Dpt. Computer Science and Artificial Intelligence, University of Granada
C/Periodista Daniel Saucedo s/n, 18071, Granada, Spain
{mdelgado, mbautis, daniel, jmserrano, vila@decsai.ugr.es}

Abstract. We present the definition of fuzzy association rules and fuzzy transactions in a text framework. The traditional mining techniques are applied to documents to extract rules. The fuzzy framework allows us to deal with a fuzzy extended Boolean model. Text mining with fuzzy association rules is applied to one of the classical problems in Information Retrieval: query refinement. The extracted rules help users to query the system by showing them a list of candidate terms to refine the query. Different procedures to apply these rules in an automatic and semi-automatic way are also presented.

1 Introduction

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 [10], [13], [17], [18].

Traditional mining techniques deal with Boolean association rules in the sense that they are generated from a set of Boolean transactions representing that an attribute is present in the transaction by 1 and non present by 0. However, most of the data implies the handling of uncertainty and quantitative attributes such as the age or the weight. The theory of fuzzy sets has been revealed as a good tool to deal with this kind of data in traditional data mining in relational databases [8], [9], [14]. In general, fuzzy association rules and fuzzy transactions can be defined to deal with structure data such as in relational databases and unstructured data, such as texts.

In this work, fuzzy association rules applying techniques from data mining will be discovered in a text framework as a process to select terms to be added to an original query. Some other approaches can be found in this direction. In [25] a vocabulary generated by association rules is used to improve the query. In [12] 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

T. Andreasen et al. (Eds.): FQAS 2002, LNAI 2522, pp. 154-162, 2002.

but in the Information Retrieval framework can be found in [20] where the extracted rules are employed for document classification.

This paper is organized as follows: from section 2 to section 4, general theory about association rules, fuzzy association rules and fuzzy transactions is presented. In section 5, an application to text framework is proposed and a definition of text transactions is given. In section 6, the extracted text association rules are applied to query reformulation in an Information Retrieval framework. Finally, conclusions and future trends are given in section 7.

2 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:

1. Find all the itemsets that have a support above threshold *minsupp*. These itemsets are called *frequent itemsets*.
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. However, 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. To solve this, one of the solutions is to make intervals considering all the possible values that can take the quantitative attributes [21]. The other solution to this problem is the use 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 [9], [14].

3 A Fuzzy Framework for 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 [5]. 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 [9].

3.1 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_o 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_o in a FT-set *T* based in *I* is represented by a fuzzy set $\tilde{\Gamma}_{I_0} \subseteq T$, defined as

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

3.2 Fuzzy Association Rules

A fuzzy association rule is a link of the form $A \Rightarrow B$ such that $A, B \subset I$ and $A \cap B = \varnothing$, 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 $\tilde{\Gamma}_A \subseteq \tilde{\Gamma}_B$. For more details about the definition of fuzzy association rules, see [9].

4 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 [26]. 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* $\tilde{\Gamma}_{I_0}$ 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* $\tilde{\Gamma}_A \cap \tilde{\Gamma}_B$ and its confidence is the evaluation of Q_M *of* $\tilde{\Gamma}_A$ *are* $\tilde{\Gamma}_B$.

We evaluate the sentences by means of method GD presented in [6]. 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

$$Q\left(\frac{|F \cap G|}{|F|}\right) \tag{5}$$

when F and G are crisp. The GD method verifies this property. For more details, see [6]. 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 [9].

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

$$Q_M \tau \in T, A \Rightarrow B \tag{6}$$

4.1 Measuring Accuracy of a Fuzzy Rule by Certainty

We propose the use of certainty factors to measure the accuracy of association rules. A previous study can be found in [7].

We define the 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 (7); otherwise, is given by expression (8), considering that if *supp(B)*=1, then $CF(A \Rightarrow B) = 1$ and if *supp(B)*=0, then $CF(A \Rightarrow B) = -1$

$$CF(A \Rightarrow B) = \frac{Conf(A \Rightarrow B) - supp(B)}{1 - supp(B)} \quad (7)$$

$$CF(A \Rightarrow B) = \frac{Conf(A \Rightarrow B) - supp(B)}{supp(B)} \quad (8)$$

We demonstrated in [2] that certainty factors verify the three properties by [21]. 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.

5 Text Mining

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 [10]. In our case, we use automatic indexing techniques coming from Information Retrieval [23]. 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 [24]. 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 the steps detailed below.

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 [22])
5. The representation of d_i obtained is a set of keywords $\{t_1, \ldots, t_m\} \in S$ with their associated weights $\{w_1, \ldots, w_m\}$

5.1 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, obtained from the algorithm included in the previous section. 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 [19]. 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.

6 Text Mining for Query Refinement

When a user try to express his/her 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 more 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 process can be understood as a feedback from the user part, but a term level, not at document level as traditionally is performed [4]. In this way, the relevance feedback is available during the query expansion, since the terms added to the query are directly selected by the user from a list generated by the system in a semi-automatic query refinement process, which is detailed in section 6.2. This process differs from the traditional relevance feedback methods, where the relevance is assigned to some retrieved documents by the user, and the terms to be added to the query are selected from the top-ranked documents retrieved once the feedback of the user has been incorporated to the original rank given by the system [16].

This problem of query optimization has been broadly treated in the field of Information Retrieval. We can find several references with solutions to this problem. A good review of the topic can be found in [11] and [15].

6.1 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.

6.2 Query Refinement Procedure

By means of fuzzy association rules, we can provide a system with a query reformulation ability in order to improve the retrieval process. We represent a query $Q = \{q_1, \ldots, q_k\}$ with associated weights $P = \{p_1, \ldots, p_k\}$. To obtain a relevance value for each document, the query representation is matched to each document representation, obtained as explained in section 5. 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 [3].

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. At this point, two different approaches can be considered: 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 are detailed in the following:

Automatic query refinement 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 and fuzzy association rules are generated.
4. The terms co-occurring in the rules with the query terms are added to the query, depending on a generalization or specialization process.
5. With the expanded query, the system is queried again.

Semi-automatic query refinement process

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

7 Concluding Remarks and Future Work

In this work, the application of traditional data mining techniques to text has been presented. The handling of uncertainty and quantitative attributes by means of fuzzy sets has been defined by the concepts of fuzzy transactions and fuzzy association rules. The application of these definitions in documents has been presented by the identification of with fuzzy transactions and fuzzy association rules with relations among terms presented in the document collection. The fuzzy framework in the mining environment allow us to work with a weighted extended Boolean model with fuzzy logic in the document and query space. Finally, an application of these elements to the problem of query optimization in Information Retrieval has been presented, identifying the generalization and specialization of the query with the inclusion of terms appearing in the antecedent and in the consequent of the rule, respectively.

In the present, we are experimenting with the presented approach and application to query refinement. A comparison to other classical approaches coming from Information Retrieval to solve the query optimization problem will be performed in the future.

References

1. Agrawal, R., Imielinski, T., Swami, A. Mining Association Rules between Set of Items in Large Databases. *Proc. of the 1993 ACM SIGMOD Conference*, pp. 207-216, 1993.
2. Berzal, F., Delgado, M., Sánchez, D., Vila, M.A. Measuring the accuracy and importance of association rules. Tech. Rep. CCIA-00-01-16, Department of Computer Science and Artificial Intelligence, University of Granada, 2000.
3. Buell, D.A., Kraft, D.H. Performance Measurement in a Fuzzy Retrieval Environment. In *Proceedings of the Fourth International Conference on Information Storage and Retrieval, ACM/SIGIR Forum*, 16(1), pp. 56-62, Oakland, CA, 1981.
4. Chang, C.H., Hsu, C.C. Enabling Concept-Based Relevance Feedback for Information Retrieval on the WWW. *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no.4, 1999.
5. Delgado, M., Sánchez, D., Vila, M.A. Acquisition of fuzzy association rules from medical data. In Barro, S. and Marín, R. (Eds.) *Fuzzy Logic in Medicine,* Physica-Verlag, 2000.
6. Delgado, M., Sánchez, D., Vila, M.A. Fuzzy cardinality based evaluation of quantified sentences. *International Journal of Approximate Reasoning,* vol. 23, pp. 23-66, 2000.

7. Delgado, M., Martín-Bautista, M.J., Sánchez, D., Vila, M.A. Mining strong approximate dependences from relational databases. *Proc. Of IPMU* 2000, Madrid.
8. Delgado, M., Martín-Bautista, M.J., Sánchez, D., Vila, M.A. Mining association rules with improved semantics in medical databases. *Artificial Intelligence in Medicine* vol. 21, pp. 241-245, 2001.
9. Delgado, M., Marín, N., Sánchez, D., Vila, M.A. Fuzzy Association Rules: General Model and Applications. *IEEE Transactions of Fuzzy Systems*, vol. 126, no.2, pp. 41-54, 2002.
10. Delgado, M., Martín-Bautista, M.J., Sánchez, D., Vila, M.A. Mining Text Data: Special Features and Patterns. *Proc. of EPS Exploratory Workshop on Pattern Detection and Discovery in Data Mining, Imperial College London, UK, September 2002.*
11. Efthimiadis, R. Query Expansion. *Annual Review of Information Systems and Technology,* vol. 31, pp. 121-187, 1996.
12. Feldman, R., Hirsh, H. Mining associations in text in the presence of Background Knowledge *Proc. of the Second International Conference on Knowledge Discovery from Databases,* 1996.
13. Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., Zamir, O. Text Mining at the Term Level. *Proc. of the 2nd European Symposium of Principles of Data Mining and Knowledge Discovery*, pp. 65-73, 1998.
14. Fu, A.W., Wong, M.H., Sze, S.C., Wong, W.C., Wong, W.L., Yu, W.K. 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), Hong Kong, pp.263-268, 1998.
15. Gauch, S., Smith, J.B. An Expert System for Automatic Query Reformulation. *Journal of the American Society for Information Science*, 44(3), pp. 124-136.
16. Harman, D.K. "Relevance Feedback and Other Query Modification Techniques". In W.B. Frakes and R. Baeza-Yates (Eds.) *Information Retrieval: Data Structures and Algorithms*, pp. 241-263, Prentice Hall, 1992.
17. Hearst, M. Untangling Text Data Mining. *Proc. of the 37th Annul Meeting of the Association for Computational Linguistics (ACL'99),* University of Maryland, June 1999.
18. Kodratoff, Y. Knowledge Discovery in Texts: A Definition, and Applications. In Z. W. Ras and A. Skowron (Eds.) *Foundation of Intelligent Systems*, Lectures Notes on Artificial Intelligence 1609, Springer Verlag, 1999.
19. Kraft, D.H., Petry, F.E., Buckles, B.P., Sadasivan, T. Genetic Algorithms for Query Optimization in Information Retrieval: Relevance Feedback. In E. Sanchez, T. Shibata and L. Zadeh, (Eds.), *Genetic Algorithms and Fuzzy Logic Systems*, in Advances in Fuzziness: Applications and Theory, vol. 7, pp. 157-173, World Scientific.
20. Lin, S.H., Shih, C.S., Chen, M.C., Ho, J.M., Ko, M.T., Huang, Y.M. Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach. *Proc. of ACM/SIGIR'98,* pp. 241-249, Melbourne, Australia, 1998.
21. Piatetsky-Shapiro, G. Discovery, Analysis, and Presentation of Strong Rules. In Piatetsky-Shapiro, G. and Frawley W.J. (Eds.) *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991.
22. Porter, M.F. An algorithm for suffix stripping. *Program,* 14(3): 130-137, 1980.
23. Salton, G., McGill, M.J. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
24. Salton, G., Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management,* vol. 24, no. 5, pp. 513-523, 1988.
25. Srinivasan, P., Ruiz, M.E., Kraft, D.H., Chen, J. Vocabulary mining for information retrieval: rough sets and fuzzy sets. *Information Processing and Management*, 37, pp. 15-38, 2001.
26. Zadeh, L.A. A computational approach to fuzzy quantifiers in natural languages. *Computing and Mathematics with Applications,* vol. 9, no. 1, pp. 149-184, 1983.