

Association rules and fuzzy association rules to find new query terms

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

Department of Computer Science and Artificial Intelligence

University of Granada

C/ Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
mdelgado@ugr.es, {mbautis, daniel, jmserrano, vila}@decsai.ugr.es

Abstract

We present an application of association rules and fuzzy association rules to find new terms that help the user to search in the web. A first query is made to the web and an initial set of documents is retrieved. Considering the terms of the collection as items, text transactions are constructed and association rules are extracted. Fuzzy association rules are also considered when the presence of an item in a transaction is a weight between zero and one. A selection of the best rules is carried out and the terms in those rules are offered to the user to refine the query.

Keywords: information retrieval, association rules, fuzzy logic, query refinement

1 Introduction

The lack of homogeneity in web documents both in the structure and in their indexing by the search robots makes difficult to find relevant information in the web. Most of the retrieved set of documents in a web search meet the search criteria but do not satisfy the user needs, and the amount of documents is so huge that the user feels overwhelmed. This is due generally to a lack of specificity in the formulation of the queries. Most of the times, the user does not know the vocabulary of the topic or query terms do not come to user's mind at the query moment. If the retrieved documents do not satisfy user's needs, the query improvement process starts.

In the field of Information Retrieval, this problem has been treated as query expansion or query refinement (a good review of the topic can be found in [7]). The main solutions given to solve it are based mainly on two approaches: the *automatic query expansion*, [10], where the augmentation of query terms to improve the retrieval process is carry out without the intervention of the user, and the second one is the suggestion of new terms to the user to be added to the original query in order to guide the search towards a more specific document space, called *semi-automatic query expansion* [12].

In our system, we use mining techniques to solve this problem. Crisp association rules and fuzzy association rules are used as a technique to find presence dependence relations among the terms from an initial set of retrieved documents. A group of selected terms from the extracted rules generate a vocabulary related to the search topic that helps the user to narrow the query with the aim of reduce the number of documents retrieved with the first query. Given a query and a set of retrieved documents, the query representation is matched to each document representation in order to obtain a relevance value for each document. If a document term does not appear in the query, its value will be assumed to be 0. In the case of the crisp case, the considered model is the Boolean one [14], while in the fuzzy case the considered model is the generalized Boolean model with fuzzy logic [4]. The user's initial query generates a set of ranked documents. If the top-ranked documents do not satisfy user's needs, the query refinement process starts.

In the next section, several mining concepts such as association rules and transactions are defined. In section 3, association rules are applied to query reformulation in an Information Retrieval framework. An ex-

*Corresponding author.

perimental example is shown in section 4 and conclusions are given in section 5.

2 Association Rules and Fuzzy Association Rules

Given a database of transactions, where each transaction is an itemset, we can extract association rules. 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 in a transaction implies the apparition of itemset I_2 in the same transaction. The reciprocal does not have to happen necessarily [11]. I_1 and I_2 are called antecedent and consequent of the rule, respectively. The rules obtained with this process are called boolean association rules or in general association rules since they are generated from a set of boolean or crisp transactions.

Fuzzy association rules are defined as those rules extracted from a set of fuzzy transactions FT where the presence of an item in a transaction is given by a fuzzy value of membership.

The complete model of transactions and association rules in the fuzzy framework can be found in [5].

2.1 Measures for Association Rules

The extraction of association rules is based on the values of *support* and *confidence*. We shall note $supp(I_k)$ the support of the itemset I_k . The support and the confidence of the rule $I_1 \Rightarrow I_2$ noted by $Supp(I_1 \Rightarrow I_2)$ and $Conf(I_1 \Rightarrow I_2)$, respectively. Support is the percentage of transactions containing an itemset, calculated by its probability, while confidence measures the strength of the rule calculated by the conditioned probability of the consequent with respect to the antecedent of the rule. Only itemsets with a support greater than a threshold $minsupp$ are considered, and from the resulting association rules, those ones with a confidence less than a threshold $minconf$ are discarded. Both thresholds must be fixed by the user before starting the process.

To deal with the imprecision of fuzzy transactions, we need to obtain the support and the confidence values with alternative methods which can be found mainly in the framework of approximate reasoning. We have selected the the evaluation of quantified sentences pre-

sented in [16], evaluated by means of method GD presented in [6].

Moreover, as an alternative to confidence, we propose the use of *certainty factors* to measure the accuracy of association rules, since they have been revealed as a good measure in knowledge discovery too [9]. A detailed study of the application of this measure can be found in [1].

The certainty factor (CF) of an association rule is defined as $I_1 \Rightarrow I_2$ based on the value of the confidence of the rule. If $Conf(I_1 \Rightarrow I_2) > supp(I_2)$ the value of the factor is given by expression (1); otherwise, is given by expression (2), considering that if $supp(I_2)=1$, then $CF(I_1 \Rightarrow I_2) = 1$ and if $supp(I_2)=0$, then $CF(I_1 \Rightarrow I_2) = -1$

$$CF(I_1 \Rightarrow I_2) = \frac{Conf(I_1 \Rightarrow I_2) - supp(I_2)}{1 - supp(I_2)} \quad (1)$$

$$CF(I_1 \Rightarrow I_2) = \frac{Conf(I_1 \Rightarrow I_2) - supp(I_2)}{supp(I_2)} \quad (2)$$

3 Association Rules and Fuzzy Association Rules for Query Refinement

Once the user queries the system, a first set of documents is retrieved. From this set, the representation of documents is extracted. We consider each document as a transaction called *text transaction*. Let us consider $T_D = \{d_1, \dots, d_n\}$ as the set of *text transactions* from the collection of documents D , and $I = \{t_1, \dots, t_m\}$ as the text items obtained as representation of each $d_i \in D$ with their membership to the transaction expressed by $W = \{w_{i1}, \dots, w_{im}\}$. On this set of transactions we extract the association rules.

We must note that we do not distinguish in this process the crisp and the fuzzy case, but we give general steps to extract association rules from text transactions. The specific cases will be given by the item weighting scheme that we consider in each case. When the weights associated to the transactions take the values $\{0, 1\}$, it means that the attribute is not present in the transaction or it is, respectively. These transactions can be called boolean or crisp transactions. In the fuzzy case, we can consider a weighted representation of the presence of the terms in the documents by a nor-

malized weighting scheme in the unit interval. Concretely, we consider two fuzzy weighting schemes, namely the frequency weighting scheme [3] and the TDIDF weighting scheme [2], both normalized. After the rules are extracted, a selection of them is carried out and the terms appearing in the selected rules are shown to the user. Then, the user chooses the most appropriate terms to guide the search towards her/his needs. Those terms are added to the query and the system is queried again.

As several authors assert, the selection of good terms in the query is crucial for the goodness of rules [8], [13] and for query expansion [12], since the poor discriminatory power of frequent terms can generate a query with a worst performance than the original one due to the poor discriminatory ability of the added terms. Therefore, the problem of selecting good terms to be added to the query have two faces. One the one hand, if the terms are not good discriminators, the expansion of the query may not improve the result. But, on the other hand, in dynamic environments or systems where the response-time is important, the application of a pre-processing stage to select good discriminatory terms may not be suitable. In our case, since we are dealing with a problem of query refinement in Internet, information must be shown on-line to the user, so a time constraint is present.

Solutions for both problems can be given. In the first case, discriminatory schemes almost automatic can be used alternatively to a preprocessing stage for selecting the most discriminatory terms. This is the case of the *TFIDF* weighting scheme. In the second case, when we work in a dynamic environment, we have to remind that to calculate the term weights following the *TFIDF* scheme, we need to know the presence of a term in the whole collection, which limits in some way its use in dynamic collections, as usually occurs in Internet. Therefore, instead of improving 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 section 2.1 is considered in this work in order to avoid the problem of non appropriate rule generation.

3.1 The Selection of Rules for Query Refinement

The extraction of rules is usually guided by several parameters such as the minimum support (*minsupp*),

the minimum value of certainty factor (*mincf*), and the number of terms in the antecedent and consequent of the rule. Rules with support and certainty factor over the respective thresholds are called *strong rules*.

Once the strong association rules are extracted, the selection of useful rules for query refinement depends on the appearance in antecedent and/or consequent of query terms. Let us suppose that *qterm* is a term that appears in the query and let $term \in S$, $S_0 \subseteq S$. Some possibilities are the following:

- Rules of the form $term \Rightarrow qterm$. We could suggest the term *term* to the user as a way to restrict the set of results.
- Rules of the form $S_0 \Rightarrow qterm$ with $S_0 \subseteq S$. We could suggest the set of terms S_0 to the user as a whole, i.e., to add S_0 to the query.
- Rules of the form $qterm \Rightarrow term$ with $term \in S$. We could suggest the user to replace *qterm* with *term* in order to obtain a set of documents that include the actual set (this is interesting if we're going to perform the query again in the web, since perhaps *qterm* is more specific that the user intended).

The previous examples allow us to provide the user with some alternatives in order to discard some retrieved documents or obtain others by performing the query again in the web. However, it is necessary to take into account the reciprocal of the rules. For example, if both $term \Rightarrow qterm$ and $qterm \Rightarrow term$ are strong, then that means that having *term* and *qterm* in a query is equivalent to some extent (depending on the *mincf* threshold employed). Then, the rules are uninteresting in order to specialice/generalize the results. But these rules can be interesting if we are going to perform the query again in Internet, since new documents not previously retrieved and interesting for the user can be obtained by replacing *term* with *qterm*.

4 An Experimental Example

To carry out the experimental stage, we have made an initial query to the search engine *Alltheweb* (<http://www.alltheweb.com>) with the search and results in Spanish. For the query terms, we have taken a short query (only one term), and a term with more

than one meaning. The term query is '*fresas*' which means '*strawberries*' in Spanish but also '*milling cutter*'. The purpose of this kind of query is to find additional terms that can broaden the query but narrow the set of retrieved documents. Therefore, if the user retrieves a set of one hundred documents with the term '*fresas*' with the intention of looking for the industrial tool and she/he does not know more vocabulary related to that concept, the resulting rules can suggest her/him some terms to add to the query, which can discard the documents related to other meanings (always that the additional terms are not in the vocabulary of the other meanings).

From the more than 61.000 retrieved documents, we analyze the 100 top-ranked documents. This technique of considering also the top-ranked documents from the document set resulted from the first query is called *local analysis* [15]. After the extraction of document representation, we obtain 832 terms. If we obtain the text transactions, we have 100 transactions with 832 items. Considering the possible weighting schemes we have proposed in section 3, we can distinguish broadly between the crisp and the fuzzy case. This last case can have a frequency weighting scheme or a TFIDF weighting scheme. Based on the selected case, we extract the rules without establishing a threshold for the confidence or the certainty measure. The number of components of the rule (antecedent and consequent) can not be more than 5. In the case of the crisp case, the number of rules extracted is 87954 (with a *minsupp* of 5%); in the case of the fuzzy frequency scheme, we obtained 68 rules (also with a *minsupp* of 5%); and, in the case of the fuzzy TFIDF scheme we obtained 3686 rules with a *minsupp* of 2%. We decide to decrease the support threshold in this last case because the number of obtained rules with a support of 5% was only 4.

These results reveal one of the main advantages of the fuzzy approach: the selection of good rules. In the crisp case, the weight of an item in a transaction can be only 0 or 1. This means that, if a term appears only one time in a document but another term appears 10 times in the same document, both of them will have a weight of 1. This generates a huge number of rules that, on one hand does not reflect the real presence relation among terms in the documents, and on the other hand, overwhelms the user, who is not able to handle and understand so many rules. The fact that in

the TFIDF scheme with a *minsupp* of 5% only 4 rules have been obtained shows that, really, this scheme discards in some way those terms with a poor discriminatory power, so the terms appearing very frequently in the whole collection have a low weight. The principle of this scheme accords to the selection of rules in the sense that those rules where a high frequent term appears do not give new information. For instance, those rules where the term *fresas* appear do not provide information about the relation of presence with the other terms in the rule, since the term *fresas* appears in all the documents (otherwise they would not have been retrieved). When the TFIDF scheme is used, the term *fresas* has assigned a weight of 0, since it appears in all the documents of the collection. This means that any rule with the term *fresas* will appear in the set of extracted rules when the TFIDF weighting scheme is applied.

However, the terms appear with *fresas* in the same rule can decrease the number of documents retrieved. For instance, in the case of the frequency weighting scheme, the rule *frontales*¹ \rightarrow *fresas* appears with a certainty factor of 1. Although from the point of view of new information the interpretation of this rule does not provide anything new, from the point of view of reducing the number of documents, the term *frontales* can suggest to the user a new term related to the meaning of the industrial tool, which she/he did not know before due to a lack of vocabulary in the topic. Other rules that provide new vocabulary terms about the industrial tool with the same weighting scheme are *herramientas*² \rightarrow *fresas* with a confidence of 70% and a certainty of 0.8.

As for the accuracy measures, some results are counterintuitive when we compare the values of confidence and certainty, which reveals that when the rules relate two items very frequent, the confidence is quite high but the certainty is not. For instance, in the crisp case, the rule *fresas* \rightarrow *productos*³ has a confidence of 13.26% while the certainty value is of 0.001.

5 Conclusion and Future Work

We have presented an application of association rules and fuzzy association rules to solve the information

¹ '*frontales*' in Spanish means '*profiles*'

² '*herramientas*' in Spanish means '*tools*'

³ '*productos*' in Spanish means '*products*'

retrieval problem of query refinement in the web. Those rules are extracted from a set of transactions which represents the document set retrieved by an initial query to the web. A list of terms extracted from the rules related to the terms in the query are shown to the user in order to refine the query.

The experimental example shows the extraction of good rules. From this rules a selection process is carried out on the basis on the form of the rule and the purpose of the process. The user can decrease the number of documents retrieved by the first query or can query the web again with a different vocabulary to retrieve a new set of documents.

As future work, we will extend this approach to the automatic case and will compare the results with other approaches to query refinement found in the literature.

References

- [1] Berzal, F., Blanco, I., Sánchez, D. & Vila, M.A. "Measuring the Accuracy and Importance of Association Rules: A New Framework". *Intelligent Data Analysis* 6:221-235, 2002.
- [2] Bordogna, G., Carrara, P. & Pasi, G. "Fuzzy Approaches to Extend Boolean Information Retrieval". In Bosc., Kacprzyk, J. *Fuzziness in Database Management Systems*, 231-274. Germany: Physica Verlag, 1995.
- [3] Bordogna, G. & Pasi, G. "A Fuzzy Linguistic Approach Generalizing Boolean Information Retrieval: A Model and Its Evaluation". *Journal of the American Society for Information Science* 44(2):70-82, 1993.
- [4] Buell, D.A. & Kraft, D.H. "Performance Measurement in a Fuzzy Retrieval Environment". In *Proc. of the Fourth International Conference on Information Storage and Retrieval, ACM/SIGIR Forum* 16(1): 56-62. Oakland, CA, USA, 1981.
- [5] Delgado, M., Marín, N., Sánchez, D. & Vila, M.A. "Fuzzy Association Rules: General Model and Applications". *IEEE Transactions on Fuzzy Systems* (accepted), 2001a.
- [6] Delgado, M., Sánchez, D. & Vila, M.A. "Fuzzy cardinality based evaluation of quantified sentences". *International Journal of Approximate Reasoning* 23:23-66, 2000c.
- [7] Efthimiadis, E. "Query Expansion". *Annual Review of Information Systems and Technology* 31:121-187, 1996.
- [8] Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y. & Zamir, O. "Text Mining at the Term Level". In *Proc. of the 2nd European Symposium of Principles of Data Mining and Knowledge Discovery*, 65-73, 1998.
- [9] Fu, L.M. & Shortliffe, E.H. "The application of certainty factors to neural computing for rule discovery". *IEEE Transactions on Neural Networks* 11(3):647-657, 2000.
- [10] Gauch, S. & Smith, J.B. "An Expert System for Automatic Query Reformulation". *Journal of the American Society for Information Science* 44(3):124-136, 1993.
- [11] Kraft, D.H., Martín-Bautista, M.J., Chen, J. & Vila, M.A., "Rules and fuzzy rules in text: concept, extraction and usage". *International Journal of Approximate Reasoning* (to appear).
- [12] Peat, H.P. & Willet, P. "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, 1991.
- [13] Rajman, M. & Besançon, R. "Text Mining: Natural Language Techniques and Text Mining Applications". In *Proc. of the 3rd International Conference on Database Semantics (DS-7)*. Chapam & Hall IFIP Proc. serie, 1997.
- [14] Salton, G. & McGill, M.J. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [15] Xu, J. & Croft, W.B. "Query Expansion Using Local and Global Document Analysis". In *Proc. of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 4-11, 1996.
- [16] Zadeh, L.A. "A computational approach to fuzzy quantifiers in natural languages". *Computing and Mathematics with Applications* 9(1):149-184, 1983.