Mining texts by association rules discovery in a technical corpus

Mathieu Roche, Jérôme Azé, Oriane Matte-Tailliez, and Yves Kodratoff

LRI, Bât 490, Université Paris-Sud, 91405 Orsay Cedex, France

Abstract. The text mining tools proposed in this paper extract association rules from a set of specialized and homogeneous texts (corpus). This tool is built in several steps and, at each of them, the expert plays a fundamental role. The first step extracts the terms from the corpus, and clusters them in classes by semantic similarity, associating each class to a concept meaningful to a field expert. Using the knowledge thus obtained, the corpus generates a table of concept frequencies in the texts. Next, we discretize the values of this table, and finally we are able to extract association rules among the concept occurrences.

1 Introduction

Our approach extracts knowledge directly from the initial corpus. It is made up of three main phases (1) collecting a homogeneous corpus on a given topic, (2) building a categorization specific to the corpus, (3) extracting and validating the extracted knowledge. Figure 1 shows the various phases of the text mining process.

The first phase will not be described in this paper, since we had to perform it for all of our corpora, though not in a completely automated way. It is important however not to forget that this phase is the basis upon which the whole text mining process is built, since its success depends heavily on the quality and the homogeneity of the collected corpora. The various treatments described in this paper have been carried out on several corpora, two in English, on the topics of Molecular Biology (9424 Ko) and Data Mining (369 Ko), two in French on the topics of Human Resources (3784 Ko) and Curriculum Vitae (2470 Ko). We shall mostly illustrate this paper by the Data Mining corpus (hereafter called DMC), made of the introductions of papers relative to the field of Data Mining, and the Human Resources corpus (hereafter called HRC), a set of texts commenting in natural language the results of ability tests.

Phase #2 of the text mining process is made of two steps: identify terms and syntactic relationships in the texts, and terms and syntactic relationships categorization. The search for terms finds the significant word groupings for the specialty field, for example *decision-tree* or *genetic-algorithm*. The search for syntactic relationships is performed by a version of Xerox'

Shallow Parser. Syntactic relationships are, for example, of the form *(ability:Subject,predict:Verb)* describing a noun-verb relation.

The terms and syntactic relationships thus obtained are associated to concepts, i.e., those showing the same semantics are clustered. Each cluster represents a concept, meaningful to the field expert, and these basic concepts can be, and usually are, the first level of an ontology of concepts. The existence of more general concepts is asserted by the expert, when needed. For instance, our expert recognized from the DMC that the set of terms *frequent-pattern*, *interesting-rule*, *decision-table* were the linguistic observables indicating the existence of concept *NatofOutput* in the texts. Similarly, *decision-tree*, *genetic-algorithm*, indicate the presence of concept *KnownMethods* in the texts. These indications of presence of a concept are defined like "traces" of this concept. It was also recognized that the concept *KnownMethods* and another concept called *Model* (the author evokes the use of a classical model for representing knowledge) were subsumed by the more general concept *Methods*, evoked by all the terms evoking its two instances. In the same way, the syntactic relationships *(find:Verb, cluster:Object)* and *(bidirectional:Adjective,rule:Noun)* point at the presence of the concept *NatofOutput*. We thus obtain a mixed ontology, the leaves of which are terms and syntactic relationships, the nodes of which are concepts.

Phase #3 is also divided into two steps, knowledge extraction, and validation. The ontology makes it possible to rewrite the corpus in a more compact way by replacing all different occurrences of a concept by the name of this concept. A system of knowledge extraction, as described in section 4, is applied to the transformed corpus. The knowledge extracted is presented in the form of association rules between concepts. The knowledge obtained is finally validated by a field expert (D, Figure 1).

2 Extracting terms as linguistic instances of a concept (*A*, Figure 1)

Spotting terms in texts is essential to detect instances of concepts. The quality of our terminology is evaluated according to precision. The definition of this measure is indicated below (1), where $\mathcal{L}$ is the whole of the terms evoking a concept. This list $\mathcal{L}$ was manually made up by the field expert.

$$Precision = \frac{\text{number of terms extracted present in } \mathcal{L}}{\text{number of terms extracted}} \tag{1}$$

Since we don't know what is the total number of pertinent terms, recall and ROC curves [8] are not computable. In the following of this paper, we will also use the *lift chart*, particularly well adapted for testing the results in

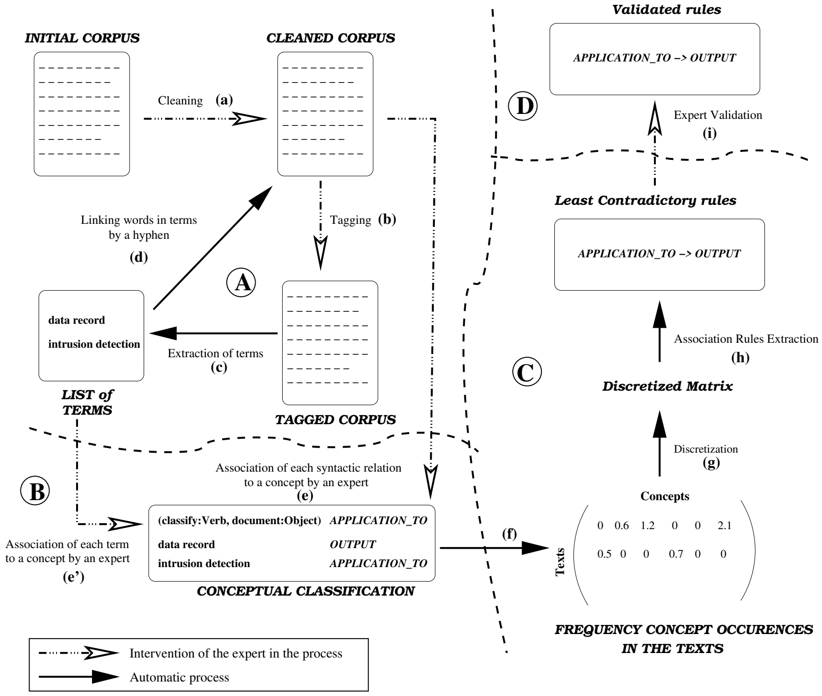

Fig. 1. The total process of text-mining to extract association rules.

unsupervised learning. The *lift chart* measures the variation of the precision as a function of the proportion of terms found by the system.

Finding a good terminology requests two preliminary steps, both asking for intervention of the field expert, namely Cleaning and Labeling.

2.1 Cleaning (*a*, Figure 1) and Labeling (*b*, Figure 1)

All the corpora we worked on with had in common to be largely unsuitable to further linguistic treatment due to the variety of unexpected forms they contain. For instance in the case of DMC, we eliminated authors names, affiliations, etc. The main point to signal is that the writing of cleaning procedures cannot be done without the help of field expert who provides sets of rules specific to the domain. For instance, in Molecular Biology, the complex combinations of upper and lower case letters convey a meaning the expert only can deal with.

Tagging labels each word of the cleaned texts by a grammatical label. At this step, we used Brill's tagger [3]. After tagging, we are able to extract the doublets or triplets of single words, called collocations [10], showing a specific grammatical label, Noun-Noun, Noun-Preposition-Noun, Adjective-Noun, Noun-Adjective[1]. The whole set of collocations will be noted $\mathcal{C}$. However, in DMC for example, more than 28% of the words of our corpus are unknown to the standard lexicon of Brill's tagger, and this implies many tagging errors. The field expert can help improve the quality of the labeling, especially in the context of Brill's tagger since it is made of rules. For instance, for DMC, the percentage of unknown words by adding 14 lexical rules to the existing ones becomes 8%.

2.2 Extracting terms (*c*, Figure 1)

This step automatically extracts the terms from the texts. In the next subsection, we shall describe our method to extract the most relevant terms for the field. Before choosing the most suitable measure according to our task of finding the relevant terms, we perform some preliminary cleaning, driven by the expert, of the collocations list $\mathcal{C}$. We exclude the terms containing words that the expert classes as non significant for term formation. For example, the collocation containing the adjectives "other", "same", "such". Note that some rules are more subtle and ask for domain expertise. For instance, some terms are not significant until they contain enough many words.

Choice of the basic measure. In order to choose the most adapted measurement, we carried out a certain amount of experiments. We present here two basic measurements made on our corpus. Mutual information [5] is based on the dependence of the two words x and y composing the collocations. Let $P(x,y)$ represent the probability of observing x and y together in this order. $P(x)$ (resp. $P(y)$) represents the probability of observing x (resp. y). The mutual information $MI(x,y)$ is defined by (2):

$$MI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{2}$$

The second measure discussed in this paper is the loglikelihood [7]. Before defining it, we must give several definitions. Consider a contingency table associated with each couple containing the words (x_i, x_j).

	x_j	$x_{j'}$ with $j' \neq j$
x_i	a	b
$x_{i'}$ with $i' \neq i$	c	d

The values a, b, c and d define occurrences of a couple where $a+b+c+d=N$ is the number of occurences of all couples.

[1] French corpus only

The loglikelihood $L(x_i, x_j)$ is defined by (3) :

$$L(x_i, x_j) = a\log(a) + b\log(b) + c\log(c) + d\log(d) - (a+b)\log(a+b) \\ -(a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) \\ +N\log(N) \quad (3)$$

These two measurements provide a likelihood score to the couples (x, y). They differ mainly because mutual information takes only into account the independence of x and y while loglikelihood takes also into account the number of collocations without x nor y.

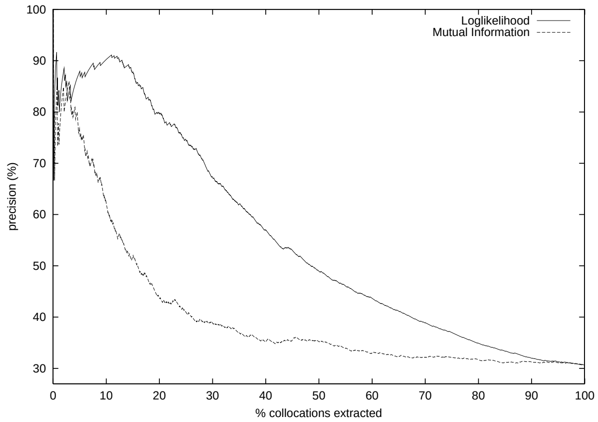

Fig. 2. Lift chart with Mutual Information and Loglikelihood on the relation "noun-adjective" in HRC. Here, we select the terms appearing more than three times.

We noticed that for all proportions of terms found, loglikelihood always gives the higher proportion of correct terms (Figure 2). This confirms Daille's assertion that loglikelihood is the best measure for the detection of terms [6].

Note, ordering collocations by their frequency, we obtain a good performance too. But to order equal frequencies of occurrences, it is necessary to use a statistic measure.

Iterative Method. In our approach, words of binary terms extracted are linked by a hyphen in texts (see d, Figure 1). For example, when finding the term "data mining," words "data" and "mining" are linked by a hyphen to form "data-mining." The originality of our approach is the iterative procedure used to extract terms. To extract terminology at the n-th iteration, we can use terms found at the $(n-1)$th one to build more specific terms. For instance in DMC, during the first iteration, we extracted the binary term "data mining." Then, during the second iteration, we extracted the term "data-mining application."

Parameters added. In order to improve the precision, we modified the scores of the terms by parameters [13]. For example, we added a parameter to favor the collocations found in several different texts rather than in a single one. An other parameter used is a simple pruning, fixing a minimum number of occurrences of a collocation, whatever its loglikelihood, in order to accept it as a term.

3 Categorization of the terms and collocations that are traces of a concept (*B*, Figure 1)

This step is still entirely carried out manually by a field expert using our software named “Rowan”[9], which decides what are the significant concepts. Rowan is a visualization tool enabling an expert to decide which concept is illustrated by a given term or syntactic relationship. Table 1 below summarizes the amount of information we collected and shows that our main effort until now has been concept recognition, rather than building ontologies.

Since we wanted to establish at first some kind of benchmark, we did not apply at this stage any inductive method to complete the categorization. We have already written an induction program based on an idea similar to LSA’s [11] i.e., terms or nouns are likely to belong to the same concept when they are in the same context, and in our case this means sharing the same term or syntactic environment.

Corpus	**Data Mining**	**Human res.**	**CV**	**Mol. Bio.**
#terms+syntax relationships	3415	9092	1958	5675
#concepts of level 1	19	18	29	204
#concepts of level 2	4	3	2	9
#concepts of level 3	1	1	0	1

Table 1. Results with different corpora.

4 Knowledge discovery (*C*, Figure 1)

This section presents the type of knowledge we look for. We try to generate association rules from the texts. The discovered knowledge will have thus the form $concept_1 \rightarrow concept_2$. Some other approaches use a taxonomy, for example to generalize the extracted association rules [15]. In our work, we directly search association rules between concepts. As usually, each rule has a support and a confidence measure [1]. To these two basic measurements, we find useful to add a third one, as we want to obtain the “least-contradicted” rules, i.e., that accepting rule $concept_1 \rightarrow concept_2$ means also checking that we almost never meet $concept_1 \rightarrow \overline{concept_2}$. Let us consider the rule : $concept_{NatofOutput} \rightarrow concept_{KnownMethods}$ with $support = 0.16$ and $confidence = 0.64$. This rule is interpreted as : "in DMC, when the nature of the outputs is described then, in 64% of the cases, also some well-known methodology is spoken of." This rule is followed by 16% of the corpus.

To obtain this type of knowledge from our corpora, we used the categorization defined in section 3. This categorization enables us to rewrite the corpus in the form of a matrix $\mathcal{M}$ of size $n*p$ where n is the number of different texts in the corpus and p is the number of concepts. The element $\mathcal{M}(i,c)$ represents the frequency of the concept c in text i. To obtain association rules such as that presented above, we must discretize the matrix $\mathcal{M}$.

4.1 Discretization (*g*, Figure 1)

For a given concept, the simplest discretization consists in dividing the texts into two subsets: texts containing no occurrences of the concept and those containing at least one. The concept under study is then divided into two subconcepts: *concept_0* and *concept_1* representing the absence of the concept and the presence of the concept in a text. A richer representation is obtained by increasing the number of possible subconcepts. For instance, a discretization in three parts generates three subconcepts: absence of the concept, weak presence of the concept and strong presence of the concept in the studied text. This new representation of the corpus in the form of a Boolean matrix is used to discover knowledge. When *concept_1*, representing texts containing at least one occurence of the concept, is divided into other sub concepts, we use a naive discretization based on equi-repartition of the frequence into each sub concept [4].

4.2 Association rules discovery (*h*, Figure 1)

Many algorithms can be used to discover the association rules from a Boolean matrix. Many approaches use the anti-monotony property of the support to carry out efficiently the discovery of all the association rules [1]. Let us recall shortly the definition of anti-monotony. A measure $\mathcal{S}$ is anti-monotonous if (*if R and R' are two rules such that $R \subset R'$ then $S(R) > S(R')$*). The use of a minimal threshold for the support makes it possible to avoid searching the entire space of the possible rules, some being pruned during the search.

The major drawback of this approach lies precisely in the need to define a minimum support. When we discover knowledge in the data (be it obtained from texts or not), we seldom are able to decide of a minimal threshold below which we are sure to stop finding interesting knowledge. Moreover, it is obvious that high support means finding rules already well-known in the field, or even trivial ones. If we withdraw the constraint related to the use of the support, we cannot use any more the traditional association rule detection algorithms. Inversely, if we accept to ask the expert to fix this threshold, we often find ourselves in situations where the selected threshold is very small and where the discovery time of the rules, and the number of obtained rules are overwhelming. In order to address the two problems, we propose a new approach based on the discovery of a subset of rules. We used the "least contradicted" measure of interest detailed in [2]. "Least contradicted" is defined by $mc(A \rightarrow B) = \frac{n_{AB} - n_{A\overline{B}}}{n_B}$, whre n_{AB} represents the number of examples of the rule $A \rightarrow B$ and $n_{A\overline{B}}$ the number of counter examples (Figure 3). Moreover, in order to discover the "best" rules from a set of candidate rules

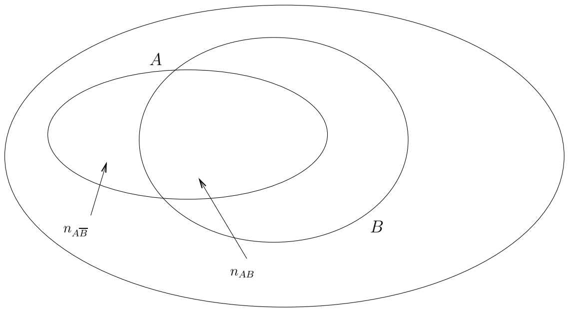

Fig. 3. Illustration of n_{AB} and $n_{A\overline{B}}$ for the rule $A \rightarrow B$.

we used a method, traditional in the field of Data Analysis [12], based on the standardization of the measure of interest, as compared to the average and the standard deviation of measures associated with the candidate rules.

Let K_{max} be the maximum number of concepts (defined by an expert) in the premise of the rules and $k \leq K_{max}$ be the number of concepts in the premise of the rule currently under study. We then study a set $\mathcal{E}^k$ of candidate rules of the type $A \rightarrow B$, with $size(A) = k$ and $size(B) = 1$. Let m be the quality measure, μ_m^k the average of m on $\mathcal{E}^k$ and σ_m^k the standard deviation of m observed on $\mathcal{E}^k$. A rule $R = A \rightarrow B$ is looked upon as interesting iff $\frac{m(R) - \mu_m^k}{\sigma_m^k} > 1$.

This approach makes it possible to keep the most interesting rules, taking into account the whole set of the candidate rules. Keeping the most general interesting rule means : if $A \rightarrow B$ is considered as interesting then $\forall C, A \wedge C \rightarrow B$ will be excluded from the set of candidate rules. This approach is similar to a method proposed by [14] who, when pruning a rule $A \rightarrow B$, prunes also all its specializations.

5 Validation of discovered knowledge (*D*, Figure 1)

In this last step, the expert must validate the association rules obtained by our algorithm. The parameters of discretization are chosen by the expert with an interactive interface. This freedom enables him/her to control the number of subconcepts and how much the rules can be given an interpretation. In general, the subconcepts representing the absence of a concept in the texts are ignored during the discovery phase because these subconcepts completely spoil understandability.

When applied to DMC, and given the discretization realised by the expert, the discovery algorithm produces 11 rules containing at most 2 subconcepts in the premise and one subconcept in the conclusion. As a means of comparison, when using the algorithm Apriori [1], we obtain 372 rules with support and confidence higher or equal to the minimal values observed among the 11 rules obtained by our approach. These 11 rules were considered to be correct by our expert and two of them were considered to be very interesting because they brought new knowledge.

Applied to HRC, our association detection algorithm generates 27 rules, while Apriori generates 3027 rules. The set of our rules have been judged interesting and non trivial by the expert. For example, the association rule *"strong presence of concept: company activity"* → *"mean presence of concept: individual relationships"* (Confidence = 0.67, Support = 0.32). This rule shows that, in the psychological tests commented by HRC, it is implicitly assumed that to be active in the company and contacts with the colleagues are linked.

6 Conclusions

We obtained very promising results in our work for dealing with the complete chain of treatment, from the raw texts to discovering association rules valid in these texts. This whole process includes several steps gathered from various fields, mainly primarily Natural Language Processing and Data Analysis. The essential contribution of our work consists in the adaptation and transformation of the various techniques in a way suitable for text mining in order to obtain a set of rules from a specialized corpus. This knowledge, represented in the form of association rules, becomes a source of new information for various field experts.

At each step of the process, the quality of the results obtained relies on one of the results obtained at the preceding steps. Thus, an error appearing at the beginning of the process can completely spoil the results of the following steps. This is why real-world applications demand a complete chain of treatment, instead of concentrating on improving the final step of knowledge discovery without considering the preceding ones.

References

1. Agrawal R., Imielinski T., Swami A-N. (1993) Mining Association Rules between Sets of Items in Large Databases. Proceedings of ACM SIGMOD International Conference on Management of Data. 207–216
2. Azé J., Kodratoff Y. (2002) A study of the Effect of Noisy Data in Rule Extraction Systems. Proceedings of EMCSR. **2**, 781–786
3. Brill E. (1994) Some Advances in Transformation-Based Part of Speech Tagging. AAAI. **1**, 722–727
4. Chickering D.M., Meek C., Rounthwaite R. (2001) Efficient Determination of Dynamic Split Points in a Decision Tree. Proceedings of the 2001 IEEE International Conference on Data Mining. IEEE Computer Society. 91-98
5. Church K-W., Hanks P. (1990) Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics. **16**, 22–29
6. Daille B., Gaussier E., Langé J-M. (1998) An Evaluation of Statistical Scores for Word Association. The Tbilisi Symposium on Logic, Language and Computation: Selected Papers, CSLI Publications. 177–188
7. Dunning T-E. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics. **19**(1), 61–74
8. Ferri C., Flach P., Hernández-Orallo J. (2002) Learning Decision Trees Using the Area Under the ROC Curve. 9th International Conference on Machine Learning, ICML'02. 139-146
9. Fontaine L., Kodratoff Y. (2003) Comparaison du rôle de la progression thématique et de la texture conceptuelle chez les scientifiques anglophones et francophones s'exprimant en Anglais. Revue du GERAS. *Accepted Asp publications*. An english version is available at http://www.lri.fr/∼yk,
10. Halliday M.A.K. (1976) System and Function in Language. Oxford University Press.
11. Landauer T., Dumais S. (1997) A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review. **104**(2), 211-240
12. Lerman I.C. (1984) Justification et validité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. Publications de l'Institut de Statistique des Universités de Paris. **29**, 27–57
13. Roche M. (2003) Extraction paramétrée de la terminologie du domaine. Revue RIA-ECA. **17**, 295–306
14. Sahar S. (1999) Interestingness via What is Not Interesting. Knowledge Discovery and Data Mining. 332–336
15. Srikant R., Agrawal R. (1997) Mining generalized association rules. Future Generation Computer Systems. **13**(2-3), 161–180