Towards a Text Mining Methodology Using Frequent Itemsets and Association Rule Extraction

Hacène Cherfi, Amedeo Napoli, Yannick Toussaint

Towards a Text Mining Methodology Using Frequent Itemsets and Association Rule Extraction

Hacène Cherfi * Amedeo Napoli * Yannick Toussaint *

Abstract: *This paper proposes a methodology for text mining relying on the classical knowledge discovery loop, with a number of adaptations. First, texts are indexed and prepared to be processed by frequent itemset levelwise search. Association rules are then extracted and interpreted, with respect to a set of quality measures and domain knowledge, under the control of an analyst. The article includes an experimentation on a real-world text corpus holding on molecular biology.*
Keywords: Association rules, text mining, quality measures, molecular biology..

1 Introduction

The access to a large amount of textual documents becomes more and more effective, considering the growth of the Web, digital libraries, technical documentation, medical data, ... These textual data constitute resources that it is worth exploiting. In this way, knowledge discovery from textual databases, or for short, text mining (TM), is an important and difficult challenge, because of the richness and ambiguity of natural language (used in most of the available textual documents).

Following some previous works [9], we present in this paper the application of association rule extraction for TM. Association rules highlight correlations between elements in the texts, *e.g.* keywords. Moreover, association rules are easy to understand and to interpret for an analyst, *i.e.* the person in charge of the mining process. However, it should be mentioned that the association rule extraction is of exponential growth (based on lattice classification) and a very large number of rules can be produced. It is mandatory to provide means for managing the set of produced rules.

In this paper, we propose a text mining process based on the general knowledge discovery process introduced in [8]. We start with a set of textual documents that are first prepared, *i.e.* selected, preprocessed, and indexed. Then we apply a text mining method, *i.e.* association rule extraction. In a final step, association rules are classified according to numerical measures, and validated by an analyst, actually an expert of the data domain. Once validated, the association rules are considered as new knowledge units and are used in turn to enrich the text indexing and annotation. It must be noticed that the whole text mining process depends on domain knowledge and on the analyst knowledge: we claim that the text mining process cannot be successfully carried out without a model of the data domain.

The outline of the paper is as follows: Section 2 describes the global process of TM as a loop, starting from the collection of texts until the extraction and validation of knowledge. Section 3 proposes a definition of text mining. Section 4 describes the use of natural language processing (NLP) tools for modelling the contents of the texts. Section 5 presents the process of association rule extraction. A set of quality measures is then associated to each extracted rule for classifying and ranking the rules (section 6). A discussion on the rule quality ends the paper.

2 Text Mining: a KDD Paradigm

Knowledge Discovery in Databases (KDD) consists in analysing raw data, or structured data in a database, in order to extract exploitable knowledge [8]. An analyst, usually an expert of the data domain, is in charge of the KDD process. The analyst selects a dataset and runs datamining tools to build one or more models that explain the dataset. Then, he chooses the more appropriate model according to his needs, completes it with his background knowledge in order to enrich the knowledge of the domain.

*LORIA (CNRS - INRIA - Nancy universities) — Campus scientifique — BP 239 Vandœuvre-lès-Nancy — F-54506 (France). Mail: {cherfi, napoli, yannick}@loria.fr

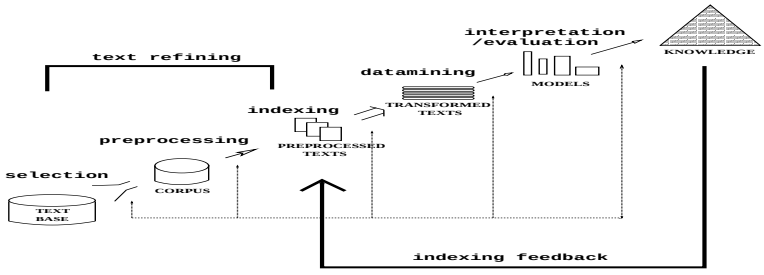

Figure 1: Text mining process.

In our approach, a KDD system relies on four main components:

- databases and the associated database management systems;
- a knowledge base system (KBS) for managing the background and the domain knowledge and for coding the knowledge extracted;
- a set of datamining tools based on symbolic or numerical techniques such as lattice classification, decision trees, induction, data analysis, statistics, ...;
- a graphical user interface dedicated to the visualisation of the results and to the navigation through the data and the knowledge units.

A KDD system usually works on large and evolving databases. The KDD process can be considered as a semi-automatic knowledge acquisition process, feeding a knowledge base with discovered knowledge units after validating the units by the analyst.

In the following, we consider a special knowledge discovery (KD) process holding on text databases.

3 Text Mining as a KD Process

We consider TM as an interactive, iterative and incremental process, as displayed in figure 1. The TM loop involves three main steps: modelling and preparing the information contained in the texts, applying a datamining method, and interpreting the extracted units. The TM is expected to provide a synthetic view of a collection of texts, *i.e.* a classification according to a set of keywords contained in the texts, or a set of association rules reflecting correlations between keywords.

The units provided by the TM process are then presented to the analyst that validates some of these units (and they can become new knowledge units), or rejects them, running the TM process with new inputs or new objectives. The TM process can be parametered using a number of numerical thresholds that can be adjusted to return only the units that are expected to be interesting (just as in the iceberg techniques presented in [15]). The TM process detailed in the following is based on the search of closed frequent itemsets (in a boolean array describing a Cartesian product $\texttt{texts} \times \texttt{keyterms}$), and association rule extraction.

Itemsets and the subsequent association rules can be interpreted in terms of term cooccurrences, and thus, may reflect semantic links between terms [1].

Itemsets and association rule extraction produce an exponential number of units. Thus, measures are introduces in [13] for reducing the set of extracted rules, and ranking the rules for allowing the analyst interpretation.

4 From Text to its Modelling

Textual documents are particular data and they need as much a particular preprocessing before the KD process. We assume that there are five different levels: the logical structure (*i.e.* parts) of the text (introduction, hypothesis, development, conclusion, etc.); the discourse structure that takes into account the link between sentences or between paragraphs; the semantics of the sentence; its syntax; and finally the lexicon.

There is also no unified semantic modelling [4] taking into account these five levels. Thus, we have to prepare the contents of the texts to ensure the quality of the TM process. This preparation relies on two main tasks: (1) the selection and annotation of textual raw data; and (2) for allowing the application of NLP tools (returning the keyterms associated to texts).

Document: **#391**
Title: Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained In vitro.
Author(s): Dessus-Babus-S ; Bebear-CM ; Charron-A ; Bebear-C ; de-Barbeyrac-B
Abstract: The L2 reference strain of Chlamydia trachomatis was exposed to subinhibitory concentrations of ofbxacin (0.5 microg/ml) and sparfbxacin (0.015 microg/ml) to select fluoroquinolone-resistant mutants. In this study, two resistant strains were isolated after four rounds of selection **[...]** A point mutation was found in the gyrA quinolone-resistance-determining region (QRDR) of both resistant strains, leading to a Ser83–>Ile substitution (Escherichia coli numbering) in the corresponding protein. The gyrB, parC, and parE QRDRs of the resistant strains were identical to those of the reference strain. These results suggest that in C. trachomatis, DNA gyrase is the primary target of ofbxacin and sparfbxacin.

Figure 2: An excerpt from the bibliographical record #391 (shorten abstract).

4.1 Text Refining

4.1.1 Extracting Textual Fields in the Sources

In our experimentations, we have worked with a collection of texts in molecular biology.[1] Actually, the texts are bibliographical notes characterised by contextual data and metadata encoded in XML tags, *e.g.* title, author(s), date, status (published or not), keywords, ... (see figure 2 for an example of such a note).A first processing of this collection of notes is used to extract two textual fields, the title and the abstract, thanks to the functionalities of the DILIB library (specialised in textual document processing [5]).

4.1.2 Part-of-Speech Tagging

Part-of-Speech (POS) tagging associates to each word of the texts a linguistic tag corresponding to its morpho-syntactic category (noun, adjective, verb, etc.). Several taggers exist for English and show high performances ($\approx$ 99,5% of correctness). They basically use a statistical model to learn how to predict the category of a word with respect to the preceding word categorisation. For example, sentence (1) extracted from figure 2 gives the tagged sentence (2):

1. Two resistant strains were isolated after four rounds of selection
2. Two/CD resistant/JJ strains/NNS:pl were/VBD isolated/VBN after/IN four/CD rounds/NNS:pl of/IN selection/NN

Even if POS taggers are robust, they are sensitive to the vocabulary and to the syntax of the sentences. Thus, if the vocabulary is very different from the one used during the learning phase, performance will decrease. The same problem occurs if the syntax is very different: it is not a problem of complexity of a sentence but a problem due to unusual syntactic structures. For example, a thesaurus usually contains nouns but no verbs neither determiners (such as numbers, articles, all, with, ...). The correctness may fall down to 95%, but still remains acceptable.

In our experimentations, we have used the Brill tagger [3] that generates a lexicon, lexical and contextual rules. The rules can be manually adapted and the tagger can be configurated for specific manipulation, in our case, the manipulation of nominal sequences.

4.2 Modelling the Contents of the Texts: the Terminological Indexing

In our experimentations, the texts have been processed according to the given representation level; a text is represented by a set of keyterms following an indexing process as exposed hereafter. A concept belonging to a specific knowledge base is associated to a noun phrase, ensuring the transition from the linguistic to the knowledge level. Such a transition is well adapted for processing text abstracts, that usually contain a high density of keyterms.

4.2.1 Term Identification and its Variants

FASTR [11] is the system that we have used for identifying the terms issued from a given vocabulary. In our case, this vocabulary results from a merge of several thesauri of the domain. In order to reduce the silence (missing to recognise a term in a text), FASTR allows to recognise a term in several variant forms. For example, the term "transfer of capsular biosynthesis genes" is considered as a variant form of the term "gene transfer" belonging to the vocabulary. However, all the variants are not acceptable, and FASTR uses linguistic rules to keep only the variants preserving the initial sense of the term.

[1] The Pascal-BioMed documentary database from the French institute for scientifi c and technical information: INIST.

Each term of the vocabulary is tagged by its syntactic structure. A syntactic variant is obtained by application of a linguistic transformation, coded by a meta-rule in the FASTR system. For example, the expression "transfer of genes" is recognised as an *inversion* of the expression "gene transfer" (belonging to the thesaurus). Moreover, expressions like "transfer of capsular biosynthesis genes" can be recognised by an *insertion* in the same way.

4.2.2 Data Description

Our corpus is composed of a set of $1,361$ documents of about $240,000$ words (1.6 Mb). A document is composed of an identifier (*i.e.* a number), a title, authors, an abstract (text in natural language), and a list of keyterms (see figure 2, augmented with the list of keyterms). The texts hold on gene mutations that cause resistance of bacteria to antibiotics. A first tagging has been done with the FASTR system, yielding $22,885$ terms corresponding to $3,337$ different terms. Among these terms, $1,762$ (*i.e.* 52.8%) appear in a single text (*i.e.* hapax). A subsequent tagging has been done on the results of the first tagging under the supervision of the analyst. This time, the corpus has been indexed by $14,374$ terms, with 632 different terms (*i.e.* 18.94% of the $3,337$ different terms of the first tagging).

5 Text Mining Process

After the preparation of the texts, the TM process is based on:

(1) the application of a datamining method, *i.e.* closed and frequent itemsets search and association rule extraction, using the so-called "*Close*" algorithm [14],

(2) the classification of the extracted association rules according to a set of *quality measures*,

(3) an interactive access to the rules and to the texts for validation of the extracted rules. An interface allows the analyst to browse the association rules extracted and classified according to a set of quality measures. Moreover, the analyst has access to the text if necessary when he wants to validate some of the rules.

5.1 Frequent Itemsets for TM

The mining algorithm that we have used for TM is based on closed and frequent itemsets search.

An itemset is a set of items that characterise objects, *e.g.* an item can be a property that is or not part of an object. Given a population of objects denoted by $\mathcal{D}$ and a set of items denoted by $\mathcal{T}$, an itemset X, *i.e.* an element of $2^{\mathcal{T}}$, is frequent if X appears in a number of objects greater than a fixed threshold denoted by $minsup$ (actually, the number of objects containing X normalised by the size of $\mathcal{D}$, also called the *support* of the itemset X).

In the context of TM, the objects in $\mathcal{D}$ are texts and the items in $\mathcal{T}$ are terms (or keyterms) of the texts. A relation R can be build on the product $\mathcal{D} \times \mathcal{T}$: $R(d, t) = 1$ if the text d contains the term t and 0 otherwise. This relation can be extended to set of texts and keyterms. The boolean table corresponding to the relation R on $\mathcal{D} \times \mathcal{T}$ constitutes the input of the text mining process. The search for frequent itemsets in this table is based on levelwise search, starting from the shorter itemsets and pruning non frequent itemsets (see [14, 15] for details). Actually, the search is based on the closed itemsets (that are maximal itemsets) and the building of the associated *Galois lattice* [6] (as done in formal concept analysis [10]).

5.2 Association Rules for TM

Association rules have already been used in TM [9, 12]. Below, we define and describe association rules in the context of TM and from our point of view.

Given a set of terms $\{t_1, ..., t_n\}$, an association rule is of the form:

$$AR : t_1, \ldots, t_k \Longrightarrow t_{k+1}, \ldots, t_n$$

meaning that a text containing the terms t_1 and $t_2 \ldots$ and t_k also contains the terms t_{k+1} and $t_{k+2} \ldots$ and t_n.

A number of measures can be associated with association rules, as explained below. Let us denote a rule by $B \Longrightarrow H$ (where B stands for *body* or *antecedent*, and H for *head* or *consequent*).

The **support** of the rule ($AR : B \Longrightarrow H$) is defined as the support of the itemset $B \bigcup H$. Usually in terms of itemsets, the itemset $B \bigcup H$ stands for the whole set $\{t_1, ..., t_n\}$, meaning that all terms t_i, $i = \{1, n\}$, have to appear simultaneously. By contrast, such an union is denoted by an intersection in the context of probability theory. Thus, as we have to use elements of probability further, we will denote the set $\{t_1, ..., t_n\}$ by $B \bigcap H$, meaning that the terms in B and H have to appear simultaneously.

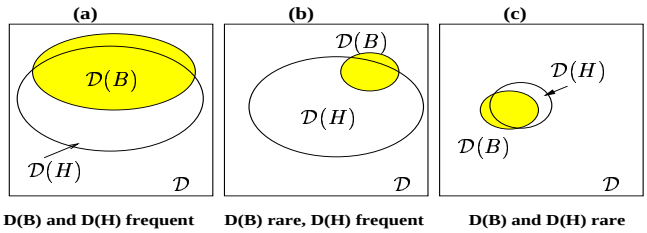

Figure 3: Three major cases illustrating the variations of $\mathcal{D}(\mathtt{B})$ and $\mathcal{D}(\mathtt{H})$.

The **confidence** of the rule AR is defined as the quotient $\frac{\mathtt{support(B \bigcap H)}}{\mathtt{support(B)}}$. Actually, the confidence can be likened to the conditional probability $\mathtt{P(H|B)}$. It measures the proportion of objects verifying the rule (that can be considered as examples of the rule, counterexample meaning in this context that there exist texts with B and without all the terms of H). When confidence of the rule is equal to 1, the rule is called *exact*, otherwise *partial* or *approximative*. For example, $\mathtt{quinupristin} \Longrightarrow \mathtt{dalfopristin}$ with confidence 0.6 means that the texts including the term quinupristin include the term dalfopristin in 60% of the cases.

An association rule AR is said to be *valid* if the support of AR is greater than a threshold minsup and the confidence of AR is greater than another threshold denoted by minconf. It can be noticed that a valid rule must be built from a frequent itemset.

These two measures, support and confidence, are the usual measures used to prune the set of rules, to obtaining a more reasonable size of this set (that grows exponentially).

In our approach, we try in fact to classify the extracted association rules according to some viewpoints that depend on the quality measures that we used. The viewpoints relying on the support and the confidence are the most well known and widely used. However, as explained in the following, there is a number of other quality measures that can be used to classify rules for interpretation by the analyst.

6 Quality Measures for Association Rules

Let B be the set of terms $\{\mathtt{t_1}, ..., \mathtt{t_k}\}$, H the set of terms $\{\mathtt{t_{k+1}}, ..., \mathtt{t_n}\}$, and $\mathtt{B} \bigcap \mathtt{H}$ the set of terms $\{\mathtt{t_1}, ..., \mathtt{t_n}\}$. Let $\mathcal{D}(\mathtt{B})$, $\mathcal{D}(\mathtt{H})$, and $\mathcal{D}(\mathtt{B} \bigcap \mathtt{H})$ be the set of texts including respectively the set of terms B, H, and $\mathtt{B} \bigcap \mathtt{H}$. Three probabilities can be defined, $\mathtt{P(B)}$, $\mathtt{P(H)}$, and $\mathtt{P(B} \bigcap \mathtt{H)}$ with the generic formula: $\left(\mathtt{P(X)} = \frac{|\mathcal{D}(\mathtt{X})|}{|\mathcal{D}|}\right)$.

As already mentioned, $\mathtt{P(B} \bigcap \mathtt{H)} = \frac{|\mathcal{D}(\mathtt{B} \bigcap \mathtt{H})|}{|\mathcal{D}|}$ is the support of the rule, and the conditional probability $\mathtt{P(H|B)} = \frac{\mathtt{P(B} \bigcap \mathtt{H)}}{\mathtt{P(B)}}$ is the confidence of the rule.

In a general way, the larger $\mathcal{D}(\mathtt{X})$ is, the higher $\mathtt{P(X)}$ is (*i.e.* close to 1). This is the case when B and H are frequent itemsets with a high frequency. However, the knowledge associated with such itemsets will not be considered as interesting, because the sets B and H cannot be used to discriminate the set of texts. More generally, we explain below a set of term distributions that are, in our view, of main interest for text mining (see figure 3).

- In case (a), the probability distributions $\mathtt{P(B)}$ and $\mathtt{P(H)}$ are both high, meaning that the terms in B and H are widespread in the whole set of texts $\mathcal{D}$. This kind of rule is not informative: a set of terms included in a large number of texts usually denote generic *concepts* in the domain of the texts. For example, in our context, a rule such as "mutation" $\Longrightarrow$ "resistance" would not be of great interest since these two terms are "key" terms and all texts are about mutation and resistance;
- In case (b), the probability distribution $\mathtt{P(B)}$ is low and $\mathtt{P(H)}$ is high. This kind of rule can be more interesting and can be interpreted as the following: texts including the set of terms B tends to include the set of terms H;
- in case (c), the probability distributions $\mathtt{P(B)}$ and $\mathtt{P(H)}$ are both low, *i.e.* terms of B and H are rare in the texts, and they often occur together (symbolised by the overlapping of sets $\mathcal{D}(\mathtt{B})$ and $\mathcal{D}(\mathtt{H})$). This means that terms in B and H are related in the present context.

In the following, we show that some quality measures reflect the three cases displayed in figure 3.

6.1 Support and Confidence Measures

Support and confidence do not allow to differentiate the three cases in figure 3. The support focuses on $\mathcal{D}(\mathtt{B}) \bigcap \mathcal{D}(\mathtt{H})$ and can be used to distinguish the case (a) from the two other cases. Confidence may be interpreted as a measure

of the inclusion degree of $\mathcal{D}(B)$ in $\mathcal{D}(H)$.

Moreover, confidence is not a discriminating measure and can stay rather constant in the three cases.

In order to separate the three cases above, and mainly the third case (the most significant one), we introduce in the following a number of other quality measures that have different and useful discrimination properties. These measures are introduced and described in [13] for example.

6.2 Related Quality Measures for Association Rules

6.2.1 Interest Measure

The **interest**, or *lift*, measures the independence degree of B and H, and is defined as: $\texttt{int}[B \Longrightarrow H] = \frac{P(B \bigcap H)}{P(B) \times P(H)}$.

The sets B and H are independent (from a probabilistic viewpoint) when $\texttt{int}[B \Longrightarrow H]$ is equal to 1. The more B and H are incompatible, *i.e.* they cannot occur simultaneously, the more $P(B \bigcap H)$ and thus the interest tend to 0. The more B and H are dependent, *i.e.* they occur simultaneously, the more the interest is greater than 1.

Provided that $(\mathcal{D}(B) \bigcap \mathcal{D}(H)) \subseteq \mathcal{D}(B)$, and $(\mathcal{D}(B) \bigcap \mathcal{D}(H)) \subseteq \mathcal{D}(H)$, and that $\mathcal{D}(B)$ and $\mathcal{D}(H)$ are small and close to $\mathcal{D}(B) \bigcap \mathcal{D}(H)$, then the value of the interest $\texttt{int}[B \Longrightarrow H]$ is high. When $P(B \bigcap H) \simeq P(B)$, or $P(B \bigcap H) \simeq P(H)$, then $\texttt{int}\ [B \Longrightarrow H] \simeq \frac{P(B)}{P(B) \times P(H)} = \frac{1}{P(H)}$, or $\texttt{int}[B \Longrightarrow H] \simeq \frac{P(H)}{P(B) \times P(H)} = \frac{1}{P(B)}$ (the interest is symmetrical). If $P(B)$ or $P(H)$ are close to 0, then the interest value is high.

Thus, the interest is a good value for discriminating case (c) in figure 3: B and H are small and have a great overlapping degree.

6.2.2 Conviction Measure

The **conviction** is defined as: $\texttt{conv}[B \Longrightarrow H] = \frac{P(B) \times P(\neg H)}{P(B \bigcap \neg H)}$. The conviction measures the *deviation* of the rule $B \Longrightarrow H$ by taking into account the rule $B \Longrightarrow \neg H$. The $\neg H$ means that at least one term of H is not present. Moreover, $|\mathcal{D}(\neg H)| = |\mathcal{D}| - |\mathcal{D}(H)|$, and $P(\neg H) = 1 - P(H)$.

The conviction equals to $\frac{1}{\texttt{int}[B \Longrightarrow \neg H]}$ and measures the degree of implication of the rule $B \Longrightarrow \neg H$. The value of conviction is high when $P(\neg H)$ is high (and thus $P(H)$ is small), when $P(B)$ is high, or when $P(B \bigcap \neg H)$ is small, meaning that $P(B \bigcap H) \simeq P(B)$ because $P(B) = P(B \bigcap H) + P(B \bigcap \neg H)$. Once again, the conviction is a measure that ranks in a first positions the rules corresponding to case (c) of figure 3. Moreover, the conviction is not symmetrical and cannot be computed for exact rules since the denominator $P(B \bigcap \neg H) = 0$ because there are no counterexamples.

6.2.3 Dependency Measure

The **dependency** is defined as follows: $\texttt{dep}\ [B \Longrightarrow H] = |P(H|B) - P(H)| = \left| \frac{P(B \bigcap H) - P(B) \times P(H)}{P(B)} \right|$. It measures the independence degree of B and H. The more close to 0 the dependency is, the more B and H are independent, *i.e.* they do not occur simultaneously.

The dependency has a similar behaviour for the cases (a) and (b) of figure 3. It can be noticed that $\texttt{dep}\ [B \Longrightarrow H] = (1 - P(H))$ and does not depend on B for exact rules.

Two other measures are introduced to study this particular last case.

6.2.4 Novelty and Satisfaction Measures

The **novelty** is defined as: $\texttt{nov}\ [B \Longrightarrow H] = P(B \bigcap H) - P(B) \times P(H)$.

We have: $|\texttt{nov}\ [B \Longrightarrow H]| = \texttt{dep}\ [B \Longrightarrow H] \times P(B)$.

The lower $P(B)$ is, the lower the novelty is. in this way, the cases (b) and (c) in figure 3 have a low novelty (and are thus ranked at the end).

The novelty varies in $]-1, 1[$ and is negative when $P(B \bigcap H) \approx 0$. The novelty is symmetrical, whereas there can exist more counterexamples for say the rule $B \Longrightarrow H$ than the rule $H \Longrightarrow B$.

This is why the **satisfaction** measure is defined as: $\texttt{sat}\ [B \Longrightarrow H] = \frac{(P(\neg H) - P(\neg H|B))}{P(\neg H)}$.

We also have $|\texttt{sat}\ [B \Longrightarrow H]| = \frac{P(H|B) - P(H)}{1 - P(H)} = \frac{\texttt{dep}\ [B \Longrightarrow H]}{P(\neg H)}$ because $P(\neg H) - P(\neg H|B) = (1 - P(H)) - (1 - P(H|B))$ $= P(H|B) - P(H)$, and $P(H|B) + P(H|\neg B) = 1$.

The satisfaction $\left(\frac{P(H|B) - P(H)}{1 - P(H)}\right)$ is not useful for classifying exact rules since $P(H|B) = 1$. For approximative rules, $P(H)$ appears in the numerator and the denominator, therefore the variation depends on $P(B)$. The lower $P(B)$ is, the higher its value is. The rules of the case (a) are bottom classified and are distinguished from the case (b). We are searching for high values of this measure (*i.e.* closer to 1).

The novelty and satisfaction must jointly examined to separate the cases (a) and (b) in figure 3. The lower the novelty is and the higher the satisfaction is, the more the rule is meaningful.

Below, we give a table synthesising the main characteristics of the quality measures that have been introduced so far.

Table 1: Quality measure characteristics

Measure	Formula	Range	Independence event	Special values	Symmetry
int $[B \Longrightarrow H]$	$\frac{P(B \bigcap H)}{P(B) \times P(H)}$	$[0, +\infty[$	1	incompatible $= 0$	$\times$
conv $[B \Longrightarrow H]$	$\frac{P(B) \times P(\neg H)}{P(B \bigcap \neg H)}$	$[0, +\infty[$	1	> 1 dep., $[0, 1[$ not dep.	
dep $[B \Longrightarrow H]$	$\lvert P(H\lvert B) - P(H) \rvert$	$[0, 1[$	0	$\simeq 1$ dep.	
nov $[B \Longrightarrow H]$	$P(B \bigcap H) - P(B) \times P(H)$	$]-1, 1[$	0	$\simeq -1$, low support	$\times$
sat $[B \Longrightarrow H]$	$\frac{(P(\neg H) - P(\neg H\lvert B))}{P(\neg H)}$	$[0, 1]$	0	$= 1$ exact rule	

7 Experiments, Interpretation and Discussion

In this section, we describe the interpretation of the extracted association rules, the last step of the TM process. As we will see, the values of the different measures vary in subsets of the set of all possible values.

7.1 The Description of the Extraction Results

Two experiments have been conducted on the corpus of texts holding on molecular biology. The TM process has been carried on while the two indexing methods, presented in paragraph 4.2.2, have been used.

When `minsup` is set to 0.7% because $49\ \%$ of the terms appear between 5 and 15 times in the texts, and `minconf` is set to 100% (*i.e.* only exact rules), $1,202$ rules are generated. 713 have a support in $[0.07, 0.11]$ corresponding to a range of $[10, 15]$ texts. However, the rules extracted are so numerous that the analyst cannot perform a precise analysis.

In the second experiment on filtered terms, `minsup` is kept unchanged and `minconf` is set to 80% (*i.e.* almost exact), 347 rules, including 128 of exact rules, are obtained.

Among these 347 rules, more than 60% of the rules are of the type of the case (c) in figure 3, *i.e.* the most interesting case from our point of view.

7.2 The Interpretation of the Analyst and the Role of the Quality Measures

The interpretation step involves the analyst, who has been asked to interpret and comment each of the 347 rules with respect to knowledge of the domain.

A rule is supposed to be *interpretable* for the analyst when he can relate all the terms included in both `B` and `H` parts. The relations between terms can be of various types, *e.g.* specialisation/generalisation, composition, causality, synonymy, hyperonymy, ...

The task of the analyst consists in explaining why two terms are related with respect to the domain knowledge. It can be noticed that the extracted association rules include the most important concepts of the domain. Moreover, the most interesting rules for the analyst are the rules relying on cases (c) and (b) in figure 3.

As mentioned before, we are mainly interested by rules having rare terms in `B` and `H`, *i.e.* case (c) in figure 3. For example, the experiment that we have done shows that the two rules #270 and #202 (see Appendix) have a high interest values (respectively 80.059 and 40.245) meaning that they are relevant for the analyst with respect to the domain knowledge.

The symmetrical behaviour of the interest measure can be useful in the interpretation. For example, the two rules #108 ("dalfopristin" $\Longrightarrow$ "quinupristin") and #332 ("quinupristin" $\Longrightarrow$ "dalfopristin") have the same high value of interest, namely 75.611. This emphasizes a similarity of behaviour of populations of bacteria resistant to the two antibiotics; this has been, in addition, confirmed by the analyst.

Moreover, it appears that the two genes `ParC` and `GyrA` occur often together in the same rules. This fact is justified by the analyst as that the two genes have a similar behaviour in the mutation mechanism. But, some rules including ...`ParC`... $\Longrightarrow$ ...`GyrA`... have been extracted with high conviction values (stressing the "direction

of the implication"). Thus, the two genes do not play exactly the same role, and the analyst explained that GyrA has been discovered before ParC, and thus there are more texts dealing with GyrA only (the more recent texts deal with both GyrA and ParC).

The analyst states that the rule #9 ("aztreonam" "clavulanic acid" "enzyme "$\Longrightarrow$ "β-lactamase") is more meaningful than the rule #11 ("aztreonam" "enzyme" $\Longrightarrow$ "β-lactamase"), even if the second has a greater support (16) than the first (11). Thus also shows that extracted knowledge units depend on the analysed corpus, and that background knowledge is necessary. Moreover, the analyst states that the rule #11 is more informative than the rule #181 ("enzyme" "β-lactamase" $\Longrightarrow$ "β-lactams"), even if "β-lactams" is a hyperonym of "aztreonam".

The two exact rules #219 ("gyrA gene" "resistance mechanism" $\Longrightarrow$ "quinolone") and #326 ("quinolone" "resistance mechanism" $\Longrightarrow$ "gyrA gene") are in relation with the same 11 texts. The interest and satisfaction values allow to rank them in a high position. Ten texts out of eleven confirm the phenomenon of resistance due to the mutation of the GyrA gene, but the last text (#1032) is in contradiction with the interpretation of the two rules: "No changes in the *quinolone-resistance* determining regions of parC, parE, *gyrA*, or gyrB were found in this mutant". This shows that the negation should be taken into account to make more precise the TM process.

Feedback on indexing.

The TM process depends on the text indexing process (as mentioned above). If an index term is missing or is not accurately used, then the TM process produce rules that are not of optimal quality. This is often the case when index term is peripheral. In turn, the results of the TM process can have a direct influence and may be used to improve the text indexing.

For example, considering a rule such as "mycobacterium tuberculosis" $\Longrightarrow$ "tuberculosis", the analyst states that the index term "tuberculosis" is not relevant here, and can mislead the interpretation. Thus, the TM process is indeed interactive and iterative: the extracted association rules can be used to filter noisy index terms in accordance to the knowledge domain.

7.3 Related Approaches

Several works are in relation with the present work. Among them, in [9], the association rule exploration and refinement is done according to specific pattern included in the antecedents and consequents of the rules. This allows to study rules even if they have low degree of confidence. In [2], the rule extraction is constrained by using a maximal antecedent and a minimal consequent (only one term). This is a way of reducing the set of extracted rules, not relying only on support and confidence. In [7], hierarchical classification and grammatical relations in texts are used for extracting the so-called diagrams of subcategorisation, describing a specific term in its context.

8 Conclusion

This article presents a practical methodology for text mining based on association rules extraction, quality measures for classifying the rules and interactive evaluation with an analyst, expert of the domain.

The text mining process relies on a classical knowledge discovery loop. First, texts are prepared, *i.e.* indexed according to certain techniques, and then they are processed. Texts are represented within boolean tables texts $\times$ keyterms. Then, the *Close* algorithm for extracting closed frequent itemsets, and the association rule extraction programme are run on these tables. A number of quality measures allow the analyst to interpret rules with more indicators than the classical support and confidence measures. Indeed, this more complete set of quality measures can be used to study association rules under new and interesting points of view.

This is still under development, and research work must be completed and enriched. In particular, a careful study of measures for their own and compared to each other has to be achieved, as well as the parallel exploitation of a domain knowledge. Furthermore, techniques such as iceberg concept lattices should be used, with profit, in text mining.

References

[1] P. Anick and J. Pustejovsky. *An Application of lexical Semantics to Knowledge Acquisition from Corpora*. In Proceedings of COLING'90: The 13th International Conf. on Computational Linguistics, volume 3, pages 712, Helsinki, 1990.

[2] J. Azé and Y. Kodratoff. *A Study of the Effect of Noisy Data in Rule Extraction Systems*. In Proc. of EMCSR 2002: 16th European Meeting on Cybernetics and Systems Research, Vienna, 2002. 6 pages.

[3] E. Brill. *Unsupervised learning of disambiguation rules for part of speech tagging*. In Proc. of Joint SIGDAT ACL'99 Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-99), College Park, University of Maryland, 1999.

[4] M. Delgado, M. J. Martin-Bautista, D. Sanchez, and M.A. Vila. *Mining Text Data: Special Features and Patterns*. In D.J. Hand et al., editor, Pattern Detection and Discovery: Proc. of ESF Exploratory Workshop, volume 2447 of LNAI, pages 140153, London, 2002. Springer-Verlag.

[5] J. Ducloy, J. C. Lamirel, and E. Nauer. *A workbench for bibliographical or factual data handling*. In Proc. of 14th Int'l Conference CODATA - The Information Revolution: Impact on Science and Technology, pages 6370, Chambéry, FR, 1994.

[6] V. Duquenne. *Latticial structures in data analysis*. Theoretical Computer Science, 217:407436, 1999.

[7] D. Faure, C. Nédellec, and C. Rouveirol. *Acquisition of Semantic Knowledge using Machine learning methods: The System ASIUM*. Technical Report ICS-TR-88-16, LRI Université Paris-Sud, January 1998.

[8] U.M. Fayyad, G.F. Piatetsky-Shapiro, and P. Smith. *From data mining to knowledge discovery*. AI Magazine, 17(3):3754, 1996.

[9] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. *Text mining at the term level*. In LNAI: Principles of Data Mining and Knowledge Discovery, 1510(1):6573, 1998.

[10] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin, 1999.

[11] C. Jacquemin. *FASTR : A Unification-Based Front-End to Automatic Indexing*. In Proc. of Computer-Assisted Information Retrieval (RIAO'94), pages 3447, New-York, 1994. Rockfeller University.

[12] Y. Kodratoff. *Knowledge Discovery in Texts : A definition, and Applications*. In LNAI: Proc. of the 11th Int'l Symp. ISMS'99, volume 1609, pages 1629, Warsaw, 1999. Springer.

[13] N. Lavrač , P. Flach, and B. Zupan. *Rule Evaluation Measures: A Unifying View*. In Proc. of ILP'99: 9th International Workshop on Inductive Logic Programming, volume 1634 of LNAI, pages 174185, Bled, Slovenia, 1999. Springer-Verlag.

[14] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. *Efficient mining of association rules using closed itemset lattices*. Information Systems, 24(1):2546, 1999.

[15] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. *Computing iceberg concept lattices with titanic*. Journal of Data and Knowledge Engineering, 42(2):189222, 2002.

Appendix: Details on the Extracted Association Rules

The theme of the texts, phenomenon of gene mutation in antibiotic-resistant bacteria, is specific and the interpretation relies on a high level of domain expertise. More precisely, antibiotic resistance occurs when bacteria change in some way that reduces or eliminates the effectiveness of drugs, chemicals, or other agents designed to cure or prevent infections. Next, we present some comments on the extracted rules.

Number: **120**
Rule: **"determine region" "gyrA gene" "Gyrase" "mutation"** $\Longrightarrow$ **"Quinolone"**
pB: "0.008" *pH*: "0.059" *pBH*: "0.008"
Support: "11" *Confi dence*: "1.000" *Interest*: "17.012" *Conviction*: "undefi ned" *Dependency*: "0.941" *Novelty*: "0.008" *Satisfaction*: "1.000"

According to the rule #120, the analyst underlines that a "mutation" of "gyrA gene" in a "determine region" of a DNA-fragment (which controls the "Gyrase" enzyme behaviour) causes a resistance to any antibiotic from the "Quinolone" family. To have the complete pattern of resistance mechanism, this rule misses the bacteria name. Actually, different bacteria were involved in the 11 texts quoted in this rule.

Number: **279**
Rule: **"mutation" "parC gene" "Quinolone"** $\Longrightarrow$ **"gyrA gene"**
pB: "0.015" *pH*: "0.046" *pBH*: "0.014"
Support: "21" *Confi dence*: "0.952" *Interest*: "20.574" *Conviction*: "20.028" *Dependency*: "0.906" *Novelty*: "0.014" *Satisfaction*: "0.950"

The rule #279 emphasizes the fact that the gene "parC" was discovered more recently than the gene "gyrA". These two genes are mutationally dependent (by combination) and resist to the "Quinolone" antibiotics family.

Number: **202**
Rule: **"grlA gene"** $\Longrightarrow$ **"mutation" "Staphylococcus Aureus"**
pB: "0.009" *pH*: "0.023" *pBH*: "0.008"
Support: "12" *Confi dence*: "0.917" *Interest*: "40.245" *Conviction*: "11.727" *Dependency*: "0.894" *Novelty*: "0.008" *Satisfaction*: "0.915"

Number: **270**
Rule: **"mecA" "meticillin"** $\Longrightarrow$ **"mecA gene" "Staphylococcus Aureus"**
pB: "0.009" *pH*: "0.012" *pBH*: "0.009"
Support: "12" *Confi dence*: "1.000" *Interest*: "80.059" *Conviction*: "undefi ned" *Dependency*: "0.988" *Novelty*: "0.009" *Satisfaction*: "1.000"

The rules #202 and #270 stress on that "Meticillin" inhibits the "mecA gene" and cure infections, due to "mutation" of the "grlA gene", caused by the "Staphylococcus Aureus" bacterium.

Number: **293**
Rule: **"mycobacterium tuberculosis"** $\Longrightarrow$ **"tuberculosis"**
pB: "0.053" *pH*: "0.067" *pBH*: "0.053"
Support: "72" *Confi dence*: "1.000" *Interest*: "14.956" *Conviction*: "undefi ned" *Dependency*: "0.933" *Novelty*: "0.049" *Satisfaction*: "1.000"

Number: **175**
Rule: **"dna" "tuberculosis"** $\Longrightarrow$ **"mycobacterium tuberculosis"**
pB: "0.0152" *pH*: "0.053" *pBH*: "0.0149"
Support: "21" *Confi dence*: "0.952" *Interest*: "18.003" *Conviction*: "19.889" *Dependency*: "0.899" *Novelty*: "0.014" *Satisfaction*: "0.950"

We present some rules that the analyst has not accepted. As we pointed out before, the automatic indexing by FASTR collect both the term and all its sub-terms if they are registered as entries of the vocabulary. The rules #293 and #175 given above are identified as an artifact of the indexing phase. 108 rules (31.1%) relate unfiltered terms.
Some rules relate synonyms to preferential terms, or to hyperonyms (*i.e.* more general terms), or to hyponyms (*i.e.* more specific terms) . These rules show that the authors describe the same concept with different terms, and the mining process reveal such usage:

Number: **183**
Rule: **"epidemic strain"** $\Longrightarrow$ **"outbreak"**
pB: "0.012" *pH*: "0.057" *pBH*: "0.012"
Support: "16" *Confi dence*: "1.000" *Interest*: "17.449" *Conviction*: "undefi ned" *Dependency*: "0.943" *Novelty*: "0.011" *Satisfaction*: "1.000"

Number: **2**
Rule: **"agar dilution"** $\Longrightarrow$ **"dilution method"**
pB: "0.019" *pH*: "0.025" *pBH*: "0.019"
Support: "26" *Confi dence*: "1.000" *Interest*: "40.029" *Conviction*: "undefi ned" *Dependency*: "0.975" *Novelty*: "0.019" *Satisfaction*: "1.000"

The rule #183 confirms that an "epidemic strain" is an "outbreak", and the next one #2 states that "agar dilution" is a kind of "dilution methods". 15 rules ($\approx$ 4.5%) over the total number of rules indicate such relations.