# How far Association Rules and Statistical Indices help Structure Terminology?

**Hacène Cherfi and Yannick Toussaint** [1]

**Abstract.** Automatic or semi-automatic structuring of terminology extracted from large corpora still remain a bottleneck issue for managing the fast growing textual sources. This paper aims at defining a methodology to tackle this point using a text mining process for association rules extraction. We show the ability of the rules to enhance the quality of the terminology by filtering the ambiguous, noisy terms of a domain of speciality. However, the mining process often generates a huge number of rules. This issue leads us to raise the question of how can we find a subset of rules that constitutes a *valid* relational structure according to the knowledge domain. We use statistical indices to rank the rules that are more capable of reflecting the complex semantic relations between terms. We also study how far some rules can help the expert with identifying synonymical / hyperonymical relations or with filtering terms.

## 1 INTRODUCTION

This article deals with data mining applied to textual sources. We present a full application for the selection and indexing of technical texts using NLP technique; followed by a process of text mining in order to discover a relational knowledge structure for this domain. Above these tasks, we point out on the role of the cognitive interpretation of the results. The user is a specialist of a specific domain (*i.e.* an expert). The results give a synthetic sight of the contents of a collection of thousands texts (*i.e.* a corpus); exhibiting genericity, similarity or causality relationships within a single text or between some of them. The text mining process enables the expert to find the well-known concepts in his domain and may also rise up trends highlighting new relations between the concepts. Here, the set of concepts delimits a set of terms (*i.e.* a terminology) of the domain, and they are related by means of the cooccurrence of terms in the texts. As it is assumed by the *lexical semantics* community, the cooccurrence relation, possibly, denotes a semantic link between terms. Hence, we evaluate the text mining's ability to produce a conceptual model of the domain represented by a set of relations. These terminological relations are built by extracting association rules. We choose the paradigm of the symbolic representation to extract the association rules. From this point of view, our work is closely related to [18] or [13] issues who also extract association rules but on non-textual data. The number of rules is exponential in the number of terms. So, another point we address in this paper is: how can we select the most significant subset from these rules that reflect the knowledge domain?

In order to achieve our objectives, we proceed in two steps:

(i) The expert identifies, within the set of extracted rules, a subset which has a special interest for him (*i.e.* the meaningful rules);
(ii) We evaluate this subset's accuracy to produce a conceptual model of the domain (*i.e.* a subset of meaningful rules) with the help of statistical indices which rank the rules from the most to the least meaningful.

Section 2 describes the characteristics that the corpus should have and enlightens the text representation format provided to the mining process. In section 3, the mining process and the association rules are defined. Section 4 relates to the *statistical indices* used to rank the rules. We introduce the criterion of interpretability of the rule, and ask the expert to evaluate each rule according to his preference. This analysis is given in section 5. In section 6, an evaluation of the adequacy of the text mining results to the needs of the expert is given, by means of the above items (i) and (ii) of our objectives. The confrontation of the formal results (calculation of the association rules, computation of the indices) with the reality of the domain (the interpretation of the expert) is an original contribution to text mining.

## 2 DATA DESCRIPTION

The very first step in the text mining process is the selection of the texts and the representation of their contents. This representation must be independent from their syntax and should reflect, mainly, their semantics. Thus, it is necessary to identify and connect the concepts quoted in the texts. The representation is based on a terminological network and on the list of the terms extracted from the texts.

**Definition 1 (Term)** *A* term *consists in one or more words considered together as a single syntactic construction (*i.e. *an indivisible unit). This term makes sense only in the context in which it is used (trade association, technical or scientific domains, etc.). This context is called* domain of speciality. *A term denotes an object (abstract or concrete) of the domain of speciality.*

"*Complex words (i.e. multi-word terms) may often reduce the ambiguity and rise up precision*" [7]. Hence, the indexing using terms instead of words is more accurate to characterise a text and to grab its contents.

What are the characteristics to choose a corpus as an input to our text mining process?

- All the texts must reflect a coherent or a homogeneous contents in a speciality domain. Narrowing the topic of the texts makes possible the use of a restricted terminology. [10] has shown that specialised corpora are characterised by a specific vocabulary.
- Each text must be written with a high density of terms. The more there are terms in a text, the more the terminological network reflecting the contents will be exhaustive. Thus, we prefer an abstract of a scientific article to a thesis for example.

[1] both are at: LORIA - INRIA Lorraine - Campus scientifique - BP 239 - Vandœuvre-lès-Nancy F-54506 cedex - France. {cherfi, yannick}@loria.fr

These are the two principal criteria which make our collection of texts a *corpus*. To enhance the identification of the terms in the texts, we use a nomenclature of terms and collect both the preferential term and its morpho-syntactic variants.

The texts are indexed using the FASTR software [11]. It is a parser based on unification grammars and, more precisely, on the logical form of Tree Adjoining Grammars (TAG). FASTR extracts from the texts all the terms it can identify within a *nomenclature*. In order to avoid too much dispersion in data, the variant forms of the terms are also collected and they are brought back to their preferential form. For instance, "*transfer of capsular biosynthesis genes*" which is not registered as a term in the nomenclature is turned into its registered form "*gene transfer*".

> ***Document***: 391
> ***Title***: Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis.
> ***Author(s)***: Dessus-Babus-S; Bebear-CM; Charron-A; Bebear-C; de-Barbeyrac-B
> ***Full abstract***: The L2 reference strain of Chlamydia trachomatis was exposed to subinhibitory concentrations of ofloxacin (0.5 microg/ml) and sparfloxacin (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 ofloxacin and sparfloxacin.
> ***Key term(s)***: "determine region" "escherichia coli" "gyra gene" "gyrase" "gyrb gene" "mutation" "ofloxacin" "parc gene" "pare gene" "point mutation" "protein" "quinolone" "sparfloxacin" "substitution" "topoisomerase"

**Figure 1.** An excerpt of the document #391 (shorten abstract).

Our corpus is composed of a set of $1,361$ documents of about $200,000$ words. It is about $6$ Mø bytes large. A document is composed of an identifier (*i.e.* a number), a title, authors, an abstract (text in natural language), and a list of key terms (see Figure 1). These texts, come from the domain of molecular biology, and deal with the gene mutations in antibiotic-resistant bacteria.

## 3 ASSOCIATION RULE EXTRACTION

### 3.1 Mining process

**Definition 2 (Text mining)** *Our text mining process consists in:*

*(a) a formal method to extract association rules;*
*(b) the computation of statistical indices that can be used to rank the rules;*

The association rules in (a) are extracted in two steps. First, we generate the frequent closed sets using the *Close* algorithm [16]. Afterward, we mine for the association rules from these sets.

Several approaches propose to deal with a high number of rules. A first one simply reduces their number by calculating a minimal set of rules, so one can infer (or retrieve) the whole set of rules. This pruning is built-in during the mining process, and after organising the data in a hierarchical structure like a Closed sets lattice [22, 20]. A second way consists in using the formalism of "rule templates". Each side of the implication rule is assigned to a type coming from the domain ontology so that one can filter them [12, 9]. Incremental techniques [5] generate the rules one-by-one when new elements are added to the database. Moreover, a maintenance criterion is used to control the rule generation process. [2] search for the "best" rules by defining two partial orders combining support and confidence. The authors claim that these two partial orders enable to grab the most interesting ones. These methods reduce the set of association rules, discarding or minimising the "redundant" ones. However, we cannot ensure that only non-redundant rules are meaningful for the expert. He may sometimes prefer one to another even if the former can be deduced from the latter. Thus, we defend another approach in which all the rules are kept. The expert can access to the most interesting ones by means of the statistical indices mentioned in (b). The indices can be seen as a weighting system used to rank the rules.

### 3.2 Association rule definition

Association rules were initially used in data analysis [15]; then in data mining in order to find regularities or correlations in large relational databases [1]. Thereafter, they were applied to text mining [8, 14].

**Definition 3 (Association rule)** *An association rule is defined as:*

$$\mathtt{R} : \mathtt{t}_1 \wedge \ldots \wedge \mathtt{t}_i \Longrightarrow \mathtt{t}_{i+1} \wedge \ldots \wedge \mathtt{t}_n$$

A rule consists in a conjunction of terms on the left hand side (called $\mathtt{B}$) implying a conjunction of terms on the right hand side (called $\mathtt{H}$). It will thus be referred as: $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$. The intuitive interpretation (*i.e.* semantics) of $\mathtt{R}$ is: if a document owns the terms $\{\mathtt{t}_1, \ldots, \mathtt{t}_i\}$ as key terms, then it also tends to own $\{\mathtt{t}_{i+1}, \ldots, \mathtt{t}_n\}$.

## 4 STATISTICAL CHARACTERISATION OF THE ASSOCIATION RULES

A mapping of the "probability theory" and the "set theory" is established as an interpretation function of the validity of the rule $\mathtt{R}$. Indeed, we represent the results of an experience using sets in a given *possibility space* $\mathtt{S}$. When $\mathtt{S}$ is finite, one can associate each element of this space to a positive quantity called a "probability" [19].

Let us consider the rule $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$. Intuitively, the informative value of $\mathtt{R}$ depends on the distribution of $\mathtt{B}$ and $\mathtt{H}$ among documents. Let $\mathtt{S}_\mathtt{B}$, $\mathtt{S}_\mathtt{H}$, and $\mathtt{S}_{\mathtt{B}\wedge\mathtt{H}}$ be the sets of documents that have the respective terms $\mathtt{B}$, $\mathtt{H}$, and $\mathtt{B} \wedge \mathtt{H}$. Three probabilities have a determinative impact for all the index values of a rule: $\mathtt{P}(\mathtt{B})$, $\mathtt{P}(\mathtt{H})$, and $\mathtt{P}(\mathtt{B} \wedge \mathtt{H})$ with

$$\mathtt{P}(\mathtt{X}) = \frac{\mathtt{number\ of\ documents\ having\ X}}{\mathtt{total\ number\ of\ documents}}$$

Figure 2 illustrates three different types of distributions of major interest in our case. The fourth possible one ($\mathtt{H}$ rare and $\mathtt{B}$ frequent) does not happen in this context since we deal with rules having high confidences[2].

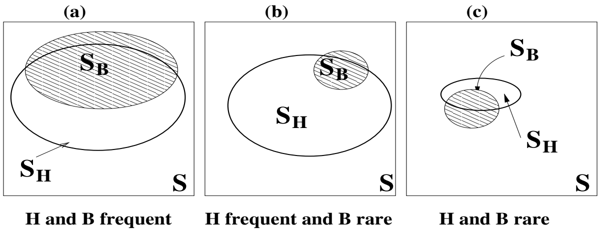

**Figure 2.** Three major cases illustrating the variability of $\mathtt{S}_\mathtt{B}$, and $\mathtt{S}_\mathtt{H}$ –S represents the possibility space–.

From the knowledge discovery point of view, the higher $\mathtt{P}(\mathtt{H})$ is, the more expected and the less significant the rule is.

[2] *cf.* sections 4.1 for `minconf` definition and 5.1 for the values of it we used in our experiments.

- In (a), the probability distributions of $B$ and $H$ are both high, that is, the terms of the rule are widespread in the corpus. This kind of rule is meaningless;
- As $P(B)$ is low, case (b) seems more interesting. Since $P(H)$ still remains high, the drawback is that any document that owns $B$, tends to also own $H$;
- Finally, the most interesting case is (c). The terms are rare and occur often together (*i.e.* $P(B \wedge H)$ is high). So they are more probably related. The experiments show that, in this case, we deal with complex rules gathering more terms on the both sides $B$ and $H$.

As a matter of fact, the two next subsections show that the values of some indices are capable of reflecting the three cases (a), (b), and (c) of Figure 2.

## 4.1 Support and confidence indices

$P(B \wedge H)$ is the well-known index called **support** of the rule $R$. The higher the support is, the larger $S_{B\wedge H}$ is (*i.e.* the more numerous are the documents in $S$ which contribute to the extraction of the rule).

The conditional probability $P(H|B) = \frac{P(B\wedge H)}{P(B)}$ is called the **confidence** of $R$. Confidence measures the degree of validity of a rule: the lower confidence is, the more numerous the counterexamples are (*i.e.* documents having the terms of $B$ but not the terms of $H$). If confidence equals $1$, the rule is called *valid* otherwise it is *approximative*.

Association rule extraction algorithms, usually, use a threshold on the support and the confidence (resp. minsup and minconf) in order to gain in efficiency. However, support and confidence are not able to make a difference between the three cases (a), (b), and (c). Support focuses on the intersection $S_B \cap S_H$, so it would distinguish on one hand (a), and on the other hand (b) and (c). The confidence highlights how far $S_B$ is included in $S_H$ and may stay rather constant in the three cases.

## 4.2 Related indices

We present some other statistical indices, found in the litterature, which give additional information and allow different "rankings" on the rules.

### 4.2.1 Interest index

The **interest** index (*a.k.a.* IBM's *lift*) measures how far from the independence $B$ and $H$ are. The interest is:

$$\mathrm{int}\,[B \implies H] = \frac{P(B \wedge H)}{P(B) \times P(H)} \tag{1}$$

Nevertheless, the interest index is completely symmetrical (*i.e.* $\mathrm{int}\,[B \Longrightarrow H] = \mathrm{int}\,[H \Longrightarrow B]$), so it cannot be used to underline the implication direction from $B$ to $H$. **Interest** varies in $[0, +\infty[$. Following the definition of section 4, the association rule context ensures that $P(B \wedge H) \leq P(B)$ and $P(B \wedge H) \leq P(H)$. Hence, the interest value increases when $S_B$ and $S_H$ are small sets. Moreover, when $S_B \cap S_H \approx S_B$ and $S_B \cap S_H \approx S_H$, this reinforces the high value of this index. The experiment confirms that rules of high interest concern terms which are rare in the corpus (*cf.* Figure 2 (c)).

### 4.2.2 Conviction index

The **conviction** index, proposed by [4], measures the independence of $B$ and $\neg H$.

$$\mathrm{conv}\,[B \implies H] = \frac{P(B) \times P(\neg H)}{P(B \wedge \neg H)} \tag{2}$$

When $B \wedge \neg H$ is true, it is a counterexample for the rule since it is the unique case where the logical implication is wrong. So the conviction measures the validity of the implication direction from $B$ to $H$. This index top-ranks rules that have low $P(H)$ or high $P(B)$ combined with highly confidences (*i.e.* $P(B \wedge H) \approx P(B)$). It has no sense for approximative rules since the denominator equals to $0$. It varies in $[0, +\infty[$, and it is not symmetrical.

It can be noticed that when $B$ and $H$ are independent, $\mathrm{int}\,[B \implies H] = \mathrm{conv}\,[B \implies H] = 1$.

### 4.2.3 Dependency index

The **dependency** index, frequently used in statistics, measures how do the fact that $B$ is known influences the probability that $H$ occurs. It is defined by:

$$\mathrm{dep}\,[B \implies H] = |P(H|B) - P(H)| \tag{3}$$

So, the more $H$ depends on $B$, the higher this index is. The major factor which increases the dependency is the size of $S_H$. This may lead to get similar dependency values in Figure 2 (a) and (b). Especially, for valid rules where $\mathrm{dep}\,[B \implies H] = 1 - P(H)$ does not depend on $B$. In order to correct this, the two following indices, which still are dependencies, are defined.

### 4.2.4 Novelty and satisfaction indices

The **novelty** index [17] is defined by:

$$\mathrm{nov}\,[B \Longrightarrow H] = P(H \wedge B) - P(B) \times P(H) \tag{4}$$

The absolute value of this index equals to $\mathrm{dep}\,[B \implies H] \times P(B)$. The lower $P(B)$ is, the lower novelty is. Hence, Figure 2 (b) and (c) cases are lower ranked. So, we are interested in low values of this index. It varies in $]-1, 1[$ and is negative when $P(B \wedge H) \approx 0$. This index is symmetrical. Therefore, the **satisfaction** index was defined as:

$$\mathrm{sat}\,[B \implies H] = \frac{(P(\neg H) - P(\neg H|B))}{P(\neg H)} \tag{5}$$

which also satisfies: $|\mathrm{sat}\,[B \Longrightarrow H]| = \frac{\mathrm{dep}\,[B \Longrightarrow \neg H]}{P(\neg H)}$. The lower $P(B)$ is, the higher this index is. It is not significant for valid rules since it equals to $1$. When $B$ and $H$ are independent, $\mathrm{dep}[B \Longrightarrow H] = \mathrm{nov}[B \Longrightarrow H] = \mathrm{sat}[B \Longrightarrow H] = 0$.

Generally speaking, these two indices must be *jointly* examined when we are in (a) or (b) (*i.e.* rules that have, more probably, low dependencies). The lower novelty and the higher satisfaction are, the more the rule is meaningful.

# 5 EXPERIMENTS

Two experiments were conducted using a corpus on molecular biology. First one, with a non-supervised indexing by FASTR, yielded $22,885$ terms corresponding to $3,337$ different terms. Among these terms, $1,762$ (*i.e.* $52.8\%$) were terms that correspond to only one document (*i.e.* hapax). The diversity of the terms in the corpus is a well-known phenomenon in "information analysis" due to *peripheral* terms used in texts. A second test was set to terms manually filtered by the expert. This filtering makes it possible to eliminate most of the noise. The corpus was indexed by a total of $14,374$ terms. There are $632$ different terms (*i.e.* $18.94\%$ of the different terms in $1^{st}$ experiment). We note that there is no terms which frequencies are lower than $5$ times, and $49\%$ of the terms occur between $5$ and $15$ times.

## 5.1 Description of the results

When `minsup` was set to $0.7\%$ and `minconf` to $100\%$ (*i.e.* only valid rules), $1,202$ rules were generated. $713$ of them have a support in $[0.07, 0.11]$ corresponding to a range of $[10, 15]$ documents.

However, the rules were so numerous that the expert cannot analyse them precisely. In the $2^{nd}$ experiment on filtered terms, `minsup` was kept unchanged and `minconf` was set to $80\%$ (*i.e.* almost valid), we obtained $347$ rules, $128$ of them were valid. This $347$-set is manageable for the next step.

Among these rules, over $10\%$ of them fall in Figure 2 case (a), $28.5\%$ in the case (b), and the remaining $61.5\%$ almost represent the case (c).

## 5.2 Interpretation step

The interpretation step involves the expert who was asked to comment each rule in order to link it up to the knowledge domain. We show through the interpretation results that the most important concepts of the domain emerge from the association rules. Moreover, three different sorts of relations, that can be used to structure the knowledge domain, were highlighted by the expert.

**Definition 4 (Interpretability)** *A rule is* interpretable *if the expert can link together all the terms involving in* B *and* H*. The task of the expert consists in explaining why it is normal, from his point of view, that one term appears with another.*

The domain of the experiment is molecular biology, more precisely, the phenomenon of gene mutation in antibiotic-resistant bacteria. This topic in a non-trivial one and it needs a certain level of expertise that we tried to get at with the help of the expert.

**Identifying complex relations:** The rule that the expert explained the most easily was about the resistance phenomenon:

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

According to this rule, the expert 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.

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

This rule emphasises that the gene "parC" was discovered more recently than the gene "gyrA". These two genes are mutationally dependent (by combination) and resist to "Quinolone" antibiotics.

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

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

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

**Synonymical / hyperonymical relations:** Some rules relate synonyms to preferential terms, or to hyperonyms (*i.e.* generic terms). These rules show that authors describe the same concept with different terms, and the mining process can reveal such usage:

*Number*: 183
*Rule*: "epidemic strain" $\Longrightarrow$ "outbreak"
*pB*: "0.012" *pH*: "0.057" *pBH*: "0.012"
*Support*: "16" *Confidence*: "1.000" *Interest*: "17.449" *Conviction*: "undefined"
*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" *Confidence*: "1.000" *Interest*: "40.029" *Conviction*: "undefined"
*Dependency*: "0.975" *Novelty*: "0.019" *Satisfaction*: "1.000"

The rule $\#183$ confirms the fact that an "epidemic strain" is an "outbreak", and the next one $\#2$ states that "agar dilution" is one kind of "dilution methods". $16$ rules ($4.6\%$) over the total number of rules indicate such relations.

**Unfiltered term relations:** Next, we present some rules that the expert denied as non-reflecting semantic relations. As we pointed out before, the automatic indexing by FASTR collect both a term and all its sub-terms if they are registered as entries of the nomenclature. $108$ rules ($31.1\%$) relate unfiltered terms. The following two rules are identified as an artifact of the indexing phase:

*Number*: 293
*Rule*: "mycobacterium tuberculosis" $\Longrightarrow$ "tuberculosis"
*pB*: "0.053" *pH*: "0.067" *pBH*: "0.053"
*Support*: "72" *Confidence*: "1.000" *Interest*: "14.956" *Conviction*: "undefined"
*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" *Confidence*: "0.952" *Interest*: "18.003" *Conviction*: "19.889"
*Dependency*: "0.899" *Novelty*: "0.014" *Satisfaction*: "0.950"

# 6 EVALUATION OF THE RULES QUALITY

## 6.1 Confronting indices to expert evaluation

The rules that reflect Figure 2 case (c) gather the most complex relations between terms on the both sides B and H. These kinds of rules are of major importance for the expert.

By definition, the interest index best ranks the rules that have rare terms in B and H. These rules are possibly meaningful from the point of view of the expert. They constitute $29.4\%$ of the total generated rules. The rules $\#270$ and $\#202$ that illustrate case (c) have the respective high values of interest $80.059$ and $40.245$.

The conviction index reinforces the implication direction from B to H, as it was emphasised. Again, about $30\%$ of the rules have a high index value of conviction. The above rule $\#279$, which indicates a time precedence of the genes quoted in, has the highest conviction ($20.028$). However, the rule $\#215$ ("gyrA gene" "parE gene" $\Longrightarrow$ "parC gene" "Quinolone"), which indicates the other implication direction from "gyrA" to "parC", falls down according to the conviction index ($11.735$). We point out on that conviction may help to distinguish between the rules $\#279$ and $\#215$, they are well ranked by the interest index since they both represent case (c).

The dependency increases as $P(H)$ decreases. The rules that correspond to Figure 2 (c) are the most probably meaningful in the domain. Hence, the dependency index reflects such case like in the rule $\#279$ and more than half of the rules ($53.3\%$) have a value of dependency index greater than the average value.

The two following rules illustrate the role of the novelty and satisfaction indices: The meaningless rule $\#273$ ("meticillin" $\Longrightarrow$ "staphylococcus Aureus") corresponds to Figure 2 (a) and the more meaningful one #265 ("mecA gene" "meticillin" $\Longrightarrow$ "Staphylococcus Aureus") because of the presence of the gene corresponds to Figure 2 (b). However, they have both low dependencies. The value of the novelty index ranks #273 before #265, and the satisfaction index ranks them conversely. Thus, the two indices can distinguish the case (a) from (b).

### 6.2 Adequacy of term extraction to expert evaluation

The important fact, for rules #293 and #175, is that there are no texts in the corpus which concern the "tuberculosis" as a disease. Nevertheless, both "Mycobacterium tuberculosis" and "Tuberculosis" are registered in the nomenclature, and thus "Tuberculosis" is collected too. As a matter of fact, non specialists may easily deduce wrong implication between the bacteria and the disease. This is exactly the kind of tricky deadlock to resolve with the help of human expertise in the indexing and in the interpreting steps. In particular, the rule #175 is more confusing since it has a high conviction value (19.889).

Indexing the texts using a nomenclature minimises the "noise" effect compared to other term extraction tools (Acabit [6], Lexter [3], Xerox Shallow Parser [21], etc.). As it is usually the case when we parse technical texts with FASTR, we build a first nomenclature from existing sources: different general medical thesauri in our case.

Rules such as #293 and #175 reveal the quality of the nomenclature. Conversely, the combination of automatic indexing with a good manual filtering enhance the quality of the rules. We are, currently, exploring how far association rules can help filtering term candidates provided by other term extraction tools.

An unexpected fact, from the knowledge discovery point of view, is that the expert prefers precise rules rather than generic ones. Following the interpretation of the expert, the rule #9 ("aztreonam" "clavulanic acid" "enzyme "$\Longrightarrow$ "$\beta$-lactamase"), is more meaningful than #11 ("aztreonam" "enzyme" $\Longrightarrow$ "$\beta$-lactamase") even if the latter has a higher support than the former (16 *vs.* 11). On the other side, he prefers the rule #11 to #181 ("enzyme" "$\beta$-lactamase" $\Longrightarrow$ "$\beta$-lactams") even if "$\beta$-lactams" is a hyperonym of "aztreonam". The documentalists often add the generic term of an entity, in the description of the documents, for information retrieval purposes. We think that it is a drawback for the interpretation phase following an automatic mining process and for any relational structure extraction from texts.

## 7 CONCLUSION

This article relates to a complete experiment in automatic processing of a technical corpus associated to a text mining process; including the interpretation of the results by an expert on real world data. Although that means a part of subjectivity inherent in any human interaction, we evaluate positively the results. For the use of statistical indices, we found that a combination of *interest* and *conviction* group the rules reflecting case (c). These rules are the most meaningful. The *novelty* and *satisfaction* values could be informative for low dependency rules when we are in Figure 2 (a) or (b).

From the knowledge structuring point of view, we insist on the quality of the indexing phase. Our approach is based on a boolean description (presence *vs.* absence) of the terms in the document. In this way, our method shows a very high sensitivity to the indexing quality. The association rules can enhance the quality of the indexing by filtering the ambiguous or noisy terms detected in the two sides of the rules. By doing this, we ensure that the subset of rules extracted is valid and may constitute a relational structure for characterising the knowledge domain.

## REFERENCES

[1] R. Agrawal and R. Srikant, 'Fast algorithms for mining association rules in large databases', in *Proc. of the 20th Int'l Conf. on Very Large Data Bases (VLDB'94)*, pp. 478–499, Santiago, Chile, (1994).

[2] R. J. Bayardo and R. Agrawal, 'Mining the most interesting rules', in *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 145–154, (1999).

[3] D. Bourigault, I. Gonzalez-Mullier, and C. Gros, 'LEXTER, a Natural Language Tool for Terminology Extraction', in *7th EURALEX Int. Congress*, pp. 771–779, Göteborg, Sweden, (1996).

[4] S. Brin, R. Motwani, J. Ullman, and S. Tsur, 'Dynamic Itemset Counting and Implication Rules for Market Basket Data', in *Proc. of the ACM SIGMOD'97 Conference on Management of Data*, volume 36, pp. 255–264, Tucson, USA, (1997).

[5] D.W. Cheung, J. Han, V. Ng, and C.Y. Wong, 'Maintenance of discovered association rules in large databases: An incremental updating technique', in *Proc. 12th IEEE Int'l Conf. on Data Engineering (ICDE-96)*, New Orleans, USA, (1996).

[6] B. Daille, 'Study and Implementation of Combined Techniques for Automatic Extraction of Terminology', in *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, 49–66, MIT Press, (1996).

[7] N. Faraj, R. Godin, R. Missaoui, S. David, and P. Plante, 'Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte', *Canadian Journal of Information and Library Science*, **21**(1), 1–21, (1996).

[8] R. Feldman and I. Dagan, 'Knowledge Discovery in Textual Databases (KDT)', in *Proceedings of the 1st Int'l Conf. on Data Mining and Knowledge Discovery*, Montreal, CA, (1995). AAI Press.

[9] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir, 'Text mining at the term level', *LNAI: Principles of Data Mining and Knowledge Discovery*, **1510**(1), 65–73, (1998).

[10] Z. Harris, *Mathematical Structures of Languages*, Wiley-Interscience, New-York, 1968.

[11] C. Jacquemin, 'FASTR : A Unification-Based Front-End to Automatic Indexing', in *Proc. of Computer-Assisted Information Retrieval (RIAO'94)*, pp. 34–47, New-York, (1994). Rockfeller University.

[12] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo, 'Finding interesting rules from large sets of discovered association rules', in *Proc. of the 3rd Int'l Conf. on Knowledge Management*, pp. 401–407, Gaithersburg, USA, (1994). ACM Press.

[13] Y. Kodratoff, 'Knowledge Discovery in Texts : A definition, and Applications', in *LNAI: Proc. of the 11th Int'l Symp. ISMS'99*, volume 1609, pp. 16–29, Warsaw, (1999). Springer.

[14] B. Lent, R. Agrawal, and R. Srikant, 'Discovering trends in text databases', in *Proc. of the Third Int'l Conf. on Knowledge Discovery and Data Mining: KDD-97*, Newport Beach, U.S.A., (1997). ACM Press.

[15] M. Luxenburger, 'Implications partielles dans un contexte', *Mathématiques, Informatique et Sciences Humaines*, **29**(113), 35–55, (1991).

[16] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, 'Efficient mining of association rules using closed itemset lattices', *Information Systems*, **24**(1), 25–46, (1999).

[17] G. Piatetsky-Shapiro, *Discovery, Analysis, and Presentation of Strong Rules*, AAAI/MIT Press, 1991. Chapter 13.

[18] A. Simon and A. Napoli, 'Building Viewpoints in an Object-Based Representation System for Knowledge Discovery in Databases', in *1st Int'l Conf. on Inform. Reuse and Integration (IRI'99)*, pp. 104–108, (1999).

[19] M. R. Spiegel, *Theory and Problems of Statistics*, Mc-Graw Hill, 1988. 2nd Edition.

[20] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal, 'Intelligent structuring and reducing of association rules with formal concept analysis', in *LNAI: Proc. of KI 2001 Advances in Artificial Intelligence*, volume 2174, pp. 335–350. Springer, (2001).

[21] Multi Lingual Theory and Technology (MLTT) group. XeLDA: Xerox Linguistic Development Architecture. see http://www.xrce.xerox.com/ats/xelda/.

[22] Y. Toussaint and A. Simon, 'Building and interpreting term dependencies using association rules extracted from galois lattices', in *Proc. of Content-Based Multimedia Information Access RIAO'00*, volume 2, pp. 1686–1693, Paris, (2000).