Knowledge-based Selection of Association Rules for Text Mining

Dietmar Janetzko, Hacène Cherfi, Roman Kennke, Amedeo Napoli, Yannick Toussaint

Knowledge-based Selection of Association Rules for Text Mining

Dietmar Janetzko[1] and **Hacène Cherfi**[2] and **Roman Kennke**[1] and **Amedeo Napoli**[2] and **Yannick Toussaint**[2]

Abstract. A reoccuring problem in mining association rules is the selection of interesting association rules within the overall, and possibly huge set of extracted rules. The majority of previous work in this area relies on statistical methods for quality estimation and selection of association rules. However, strictly bottom-up approaches are oblivious of knowledge though knowledge may be available (e.g., provided by ontologies), and rule extraction may take advantage of it. In this paper, we conceive of the problem of selecting association rules as a classification task. A framework of a binary probabilistic classifier is introduced that uses ontologies in order to estimate whether and to which degree a rule expresses a mere taxonomic relationship. In so doing, selection of association rules (selection by elimination) is carried out by identifying and discarding trivial association rules.

1 INTRODUCTION

We propose a knowledge-based approach to select interesting association rules extracted from textual databases. Following the classical knowledge discovery schema [6], mining of association rules is carried out in two steps: First, associations between sets of items in databases are discovered (association rule extraction). Second, their interestingness or quality is evaluated by a domain expert (analyst) or by using statistical quality measures. A subset from the overall set of association rules discovered is selected (association rule selection). Association rules have been extracted from market basket databases, but extracting and selecting association rules has also been studied successfully in other domains, especially in text mining, e.g., [7, 5]. A number of algorithms, like Apriori [1] or Close [13] have been designed to tackle the computationally expensive task of association rule extraction. A problem that limits the extraction of association rules is the difficulty to identify and select a subset of interesting rules or, conversely, rules that are trivial. Selection of association rules is usually addressed by using statistical quality measures. Statistical quality measures can easily be applied on the given data. They proceed without using domain knowledge. A number of statistical quality measures have been explored, and their agreement or non-agreement is studied carefully (e.g., [12, 16]). However, quality measures like *support* or *confidence* do not suffice to solve this problem since they often generate contradictory results. A more principled problem of statistical approaches to rule selection is that it makes independent quality measurement of association rules impossible. By following a statistical approach quality measures of association rules rely on the same information that has already been used to discover association rules in the previous step. Apart from the data that feed into the process of rule selection, there are often sources of knowledge available that might be used for the same purpose (e.g., conceptual hierarchies or ontologies). In fact, there are knowledge-based approaches introduced to association rule selection (e.g., [7]), but they are are confined to a small number of information or knowledge types.

A central hypothesis of this paper is the assumption that the key to association rule selection is the usage of knowledge. The knowledge-based approach presented in this paper works by conducting a negative selection or selection by elimination (rejection) of association rules. The overall and possibly large set of rules provided as an output of association rule mining is reduced by those rules that meet a particular criterion. Whether and to which degree an association rule meets a considered criterion (i.e., exemplifying a mere taxonomic relationship) is estimated by a probabilistic approach the details of which are given below. In this way, the initial rule set can be compressed since trivial or non-interesting association rules, i.e., rules that reveal conceptual or taxonomic explanations of concepts, are identified and discarded. Note, however, that by following a rejection-oriented approach we can not take the fact of a non-rejection of a rule as evidence for its quality.

The paper is organised as follows: First, we present the overall formal framework that spells out our approach for selecting association rules. We will present examples that make use of simple ontologies, which are built upon the hypernym-hyponym relationship. Second, a probabilistic framework is introduced and is taken to carry out calculations of the degree of fitness between each rule of a rule set initially mined and a model of the domain chosen (e.g., an ontology). Third, we present an example taken from a text mining experiment that illustrates in which way knowledge, i.e., ontologies, supports the selection of association rules. The paper concludes with a discussion of possible further extensions of the knowledge-based approach to association rule selection.

2 SELECTING ASSOCIATION RULES

Association rules

Let $\mathcal{T} = \{\mathtt{t}_1, \ldots, \mathtt{t}_n\}$ be a non-empty finite set of *texts*. Likewise, let $\mathcal{K} = \{\mathtt{k}_1, \ldots, \mathtt{k}_m\}$ be a non-empty finite set of *keyterms*, i.e., concepts describing the contents of these texts. The set of texts $\mathcal{T}$ and the set of keyterms $\mathcal{K}$ are related through a binary relation $R \subseteq \mathcal{T} \times \mathcal{K}$. An association rule is taken to be an implication of the form $\mathtt{B} \Longrightarrow \mathtt{H}$ where $\mathtt{B}$ stands for *body* or *antecedent*, and $\mathtt{H}$ for *head* or *consequent* with $\mathtt{B} \subseteq \mathcal{K}$, $\mathtt{H} \subseteq \mathcal{K}$ and $\mathtt{B} \cap \mathtt{H} = \emptyset$. Let

[1] GLOBALPARK GmbH, 50354 Hürth (Germany), email: {janetzko,kennke}@globalpark.de

[2] LORIA BP 239 54506 Vandœuvre-lès-Nancy (France), email: {cherfi ,napoli,yannick}@loria.fr. We would like to thank the ECCAI societies for the travel grant given to H. Cherfi .

$\mathtt{B} = \{\mathtt{k}_1, \ldots, \mathtt{k}_p\}$ be the set of keyterms of the body of an association rule r_i and $\mathtt{H} = \{\mathtt{k}_{p+1}, \ldots, \mathtt{k}_q\}$ be the set of terms of the head of r_i. $\mathtt{B} \Longrightarrow \mathtt{H}$ means that all the texts in $\mathcal{T}$ containing the keyterms $\mathtt{k}_1$, $\mathtt{k}_2$, ..., $\mathtt{k}_p$ also contain the keyterms $\mathtt{k}_{p+1}$, $\mathtt{k}_{p+2}$, ..., $\mathtt{k}_q$ with a probability $\mathtt{P}$. The support of r_i is the number of texts containing the keyterms in $\mathtt{B} \cup \mathtt{H} = \{\mathtt{k}_1, \ldots, \mathtt{k}_p, \ldots, \mathtt{k}_q\}$). The confidence of r_i is the ratio of the number of texts containing the keyterm set $\mathtt{B} \cup \mathtt{H}$ and the number of texts containing $\mathtt{B}$ ($\{\mathtt{k}_1, \ldots, \mathtt{k}_p\}$). This ratio is interpreted as the conditional probability $\mathtt{P}(\mathtt{H}|\mathtt{B})$. Support and confidence are two quality measures attached to the association rules [1]. Two user-defined thresholds σ_s for minimal support and σ_c for minimal confidence are used to constrain the process of association rule mining.

We want an association rule to conform to a model provided that the concepts of B and H are adjacent concepts in the considered model. For example, the concepts apple and fruit are adjacent concepts in an ontology that relies on the hypernym-hyponym relationship. Thus, the rule "apple" $\Longrightarrow$ "fruit" is a strong candidate for rejection because it is strictly taxonomic. Conversely, we would like to keep the rule "cherry pie" $\Longrightarrow$ "chocolate", "butter". This rule expresses an interesting combination of "cherry pie", "chocolate" and "butter" that is not taxonomic.

3 MODELS

Models represent the knowledge that is used for selecting association rules. Seen from a formal point of view, we specify models of a domain (e.g., ontologies) by using *relational structures* (e.g., [11], p. 8). Networks of concepts (e.g., hierarchies, ontologies, meronymies, semantic networks) can be redescribed as relational structures. A structure consists of the following ingredients: (i) a non-empty set of items or *concepts* called the universe or domain of the structure, (ii) various operations on the universe, and (iii) various relations on the universe. The operations are optional [3]. A structure made up only of a universe and various relations is usually called a relational structure, which can be specified as follows. Let $C = \{c_1, ..., c_u\}$ be a non-empty finite set of concepts, and let $\overline{R} = \{R_1, ..., R_p\}$ be non-empty finite set of relations, then R_i^C symbolises the set of pairs in $C \times C$ for which the relation R_i holds. Thus, a relational structure is made up of the set of pairs R_i^C together with the set of concepts C. For $i \in \{1, \ldots, n\}$ the n relational structures are expressed by (C, R_i^C). Whenever all the R_i^C for $i \in \{1, \ldots, n\}$ are defined on C, we simply write (C, R). If a relational structure (C, R^C) is used to give a probabilistic account of association rules, we will refer to it as a model M.

On the one hand, we have to distinguish between the set of concepts C that is used to define a model and the set of keyterms $\mathcal{K}$ the concepts of which are used in association rules. This is necessary since we can not rule out the possibility that unsuitable models are used to explain rules. On the other hand, we have to specify a match between a rule r_i and a model. Let the set of concepts common to a rule r_i and a model M be denoted by Z_i. Then, a rule r_i matches a model M if $\forall\, k_j$ a keyterm of a rule r_i $\exists\, c \in C$ such that $Z_i \neq \emptyset$ with $Z_i = \{(k_1 = c_1), \ldots, (k_q = c_u)\}$.

4 MODELS AS PROBABILITY DISTRIBUTIONS

We spell out the probabilistic framework that is used for knowledge-based selection of association rules. This is achieved by calculating the maximum *likelihood* score $P(r_i|M)$, i.e., the probability of an association rule r_i given a model M. To specify $P(r_i|M)$ we have to define a probability distribution P over the concepts of the model considered. In so doing, we make use of the spreading activation theory [4] cast in probabilistic terms. Defining a probability distribution over the concepts of a model M consists of three basic steps:

I. Calculating Minimal Path Lengths. The probability distribution over the concepts of a model is calculated by using the minimal path that has to be traversed within a model M in order to connect the keyterms of B and the keyterms of H of a rule r_i. This is only possible if the concepts of the model M can be related to the concepts of a rule r_i. Thus, $C \cap \mathcal{K} \neq \emptyset$. If there are several alternative paths, the shortest one is chosen. A failure to find a path that connects the concepts of B and H within a model M is denoted by a path length of 0. The length of a path between the concepts c_u and c_v within a model M is denoted by $\ell(c_u, c_v)$. Accordingly, $\min(c_u, c_v)$ is used to refer to the minimal path length between a pair of concepts of B and H. The expression $|(c_u, c_v)|$ is used to refer to the number of paths that relate c_u to all concepts in C and thus in M. Hence, $(c_u, c_v) \subseteq C \times C$ for which the relation R holds. If we want to express that the transition of two concepts c_u and c_v of a model is used to calculate what is called the likelihood of a specific association rule r_i, we indicate this by writing c_u^i and c_v^i.

II. Calculating the Decay Rate Attached to the Minimal Path Lengths. Let us consider the situation that there is in fact a path between c_u and c_v via a model M. A low score will then be assigned to the transition and thus to the rule if the path is long and vice versa. The same can be expressed in terms of the theory of spreading activation [4]. Then, we say that there is a monotonic decay of activation among the concepts of a model. Using more formal expressions, this is specified by introducing a function δ with $\delta : \mathbb{N} \rightarrow \mathbb{R}$. The function δ is a weighting mechanism that punishes long paths between a concept $c_u \in$ B and $c_v \in$ H of a model M. It is calculated by using the reciprocal of the length of a path between c_u and c_v such that $\delta = 1/\ell(c_u, c_v)$.

III. Deriving a Probability Distribution over Models. Probability theory tells us that the scores (one-step transition probabilities) obtained for a transition of a concept c_u of B to each concept of M should sum up to 1. What is needed is a variable to accomplish this standardisation. We take ξ to denote this standardisation score. Note that ξ is calculated $\forall c \in C$. Intuitively, we say that as a result of the ξ-standardisation models with a high number of transitions (branchings) are punished, i.e., in general, their one-step transition probabilities are low. By contrast, models with a low number of transitions are rewarded, i.e., in general, their transition probabilities are high. This is equivalent to the introduction of a weighting of path lengths according to the branching of paths. To achieve the standardisation, we multiply each reciprocal of a path length so that the resulting sum will be 1. This standardisation score is is denoted by ξ. For each c_u, we compute ξ by adding up all scores for the reciprocal of path lengths and calculating the reciprocal of the resulting sum, i.e., for the concept c_u:

$$\xi = \left(\sum_{i=1}^{|(c_u, c_v)|} \frac{1}{\ell(c_u, c_i)} \right)^{-1} \tag{1}$$

Putting all three building blocks for the definition of a probability distribution over the concepts of a model M together, we are in a position to state the probability distribution function P_M as:

$$P_M(c_u, c_v) = \xi \cdot \delta \cdot \min(c_u, c_v) \tag{2}$$

Based on equation 2, we may specify the probability of a rule r_i given

a model M. This score is what is called the likelihood of a rule, which expresses the goodness of fit between an association rule r_i and a model M. Do the likelihood scores reflect correctly the goodness of fit between a rule r_i and a model M? To answer this question we will consider two association rules r_p and r_q and the corresponding likelihood scores given a model M. We will examine two situations: (i) the length of paths to be traversed via a model M is the same for the two rules considered. However, the number of transitions of the concept c_u is larger for rule r_p than for rule r_q. (ii) the length of paths to be traversed via a model is larger for rule r_p than for rule r_q. However, the number of transitions is the same for the two considered rules. Let M be a model and let there be a match between each of two simple association rules r_p, r_q and the model M such that $Z_p \neq \emptyset$ and $Z_q \neq \emptyset$. This match simply indicates that the concepts of the model and the concepts of the rule intersect. If $\ell(c_u^p, c_v^p) = \ell(c_u^q, c_v^q)$, i.e. if we keep δ and thus the lengths of paths constant (situation i), then it follows from equation 2:

$$|(c_u^p, c_v^p)| > |(c_u^q, c_v^q)| \Rightarrow P(r_p|M) < P(r_q|M).$$

5 LIKELIHOOD CALCULATION

We give now give a intuitive example that shows how models are expressed in terms of transition probabilities. In this way, a probability distribution over models is defined that is later used to calculate the likelihood of association rules. A set of texts $\{t_1, \ldots, t_6\}$ is described by a set of keyterms $\{a, \ldots, e\}$ in Figure 1 (a). A model of a simple ontology is depicted in Figure 1 (b), that is read as follows: "a" IS-A "b", "e" IS-A "c", etc. The IS-A relation is reflexive and transitive but this is not indicated on the figure (for simplicity).

Text	Keyterms
t_1	{acd}
t_2	{bce}
t_3	{abce}
t_4	{be}
t_5	{abce}
t_6	{bce}

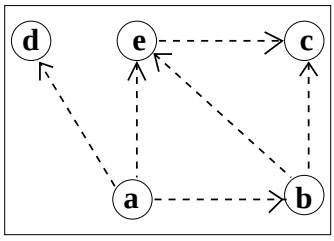

Figure 1. (a) The formal textual database – (b) A simple ontology.

Table 1. Figure 1 written in terms of transition probabilities

↱	a	b	c	d	e	Σ
a	0.22	0.22	0.11	0.22	0.22	1
b	0.05	0.30	0.30	0.05	0.30	1
c	0.10	0.10	0.60	0.10	0.10	1
d	0.10	0.10	0.10	0.60	0.10	1
e	0.06	0.06	0.40	0.06	0.40	1

We will now address the question how the likelihood of association rules is calculated. As shown by the example presented above, we make use of equation 2 to describe a model in terms of transition probabilities. In so doing, for every link that occurs in the graphical representation of the model, we calculate the transition probabilities of the concepts involved. We express the one-step transition probability of $\mathtt{k_p} \in \mathtt{B}$ to $\mathtt{k_{p+\ell}} \in \mathtt{H}$ given a model M by $\mathtt{P_{k_p k_{p+\ell}}}|M$. The probability distribution over all concepts of the ontology shown in Figure 1 (b) leads us to a one-step transition probability matrix presented in Table 1. Pairs of concepts in Figure 1 (b) that are connected by short paths, e.g., (b, c), are described by large transition probabilities (0.30). [3] Conversely, pairs of concepts connected by long paths, e.g., (b, d), are associated with small transition probabilities (0.05). [4]

Once we have calculated the probability distribution of a model M, the determination of $P(r_i|M)$, i.e., the likelihood of a *simple association rule* with $|\mathtt{B}| = |\mathtt{H}| = 1$ proceeds by applying equation 2. Thus, given a probabilistic description of a model we find the likelihood scores for simple association rules in the table of transition probabilities (cf. Table 1). For a rule: "b" $\Longrightarrow$ "e", we have that $P(b \Longrightarrow e|M) = 0.30$. However, if we want to calculate the likelihood of *complex association rules* with $|\mathtt{B}|$ or $|\mathtt{H}| > 1$, we need a more general procedure. We will now delineate this procedure for calculating the likelihood of association rules. Note that this procedure covers likelihood calculation of simple rules as a special case. Since there is a number of transition probabilities involved in complex association rules, we have to calculate an aggregate probability score.

Calculating the likelihood of an association rule starts by the Cartesian product $\mathtt{B_i} \times \mathtt{H_i}$ of the one-step transition probabilities of the concepts involved in an association rule r_i.

$$P(\mathtt{B_i} \times \mathtt{H_i})|M = \prod_{k=1}^{|\mathtt{B_i} \times \mathtt{H_i}|} \{P_{k_1 k_{p+1}}|M, \ldots, P_{k_p k_q}|M\} \quad (3)$$

We then proceed from the set of $|\mathtt{B_i} \times \mathtt{H_i}|$ transition probabilities to the aggregate likelihood score of an association rule. This procedure rests on the assumption that all terms in an association rule r_i are of equal importance. The number of terms in a rule should not impact the likelihood of a rule. For this reason, likelihood calculation is based on the geometric mean of the product of the $|\mathtt{B_i} \times \mathtt{H_i}|$ transition probabilities. This is done by taking the $|\mathtt{B_i} \times \mathtt{H_i}|^{th}$ root of the product of the single transition probabilities. Thus, the aggregated likelihood of rule r_i given model M that expands into $k_1, \ldots, k_p \Longrightarrow k_{p+1}, \ldots, k_q|M$ is calculated by:

$$P(r_i|M) = \sqrt[|\mathtt{B_i} \times \mathtt{H_i}|]{\prod_{k=1}^{|\mathtt{B_i} \times \mathtt{H_i}|} \{P_{k_1 k_{p+1}}|M, \ldots, P_{k_p k_q}|M\}} \quad (4)$$

6 CLASSIFICATION

6.1 A knowledge-based classifier

A high likelihood of an association rule r_i indicates that it conforms to the model M. However, selection depends not only on the likelihood of an association rule considered but also on the model applied. By using both kinds of information a probabilistic binary classifier for association rules is realised. In general, a classifier is defined as a function f that maps an input to a class w_j with $j \in \{1, \ldots, n\}$. The input is usually an ordered n-tuple or vector of attributes. In knowledge-based association rule extraction, the input to the classification task is an association rule r_i and a model M both of which

[3] To deal with the problem of a non-existing path between two concepts, two strategies are used. In simple rules, the probability of 0 is associated to a non-existing path. In complex rules, however, this strategy would turn the aggregate score into 0. Hence, a "penalty score" is used that equals to a minimal probability: $\mathtt{P_M(c_u, c_v)} = \frac{1}{\mathtt{n}+1}$ with $\mathtt{n}$ being the number of concepts of M. In the example, the penalty score was used, and $\mathtt{n} = 5$. For (b,c), we have: $\xi = \left(\left(3 \times \frac{1}{1}\right) + \left(2 \times \frac{1}{6}\right)\right)^{-1} = 0.3$ and $\delta = 1$.

[4] For (b, d): ξ remains the same and $\delta = \frac{1}{6}$.

are used to calculate the likelihood score $P(r_i|M)$. In addition, we make use of a threshold t to carry out a binary classification:

$$f(r_i, M) = \begin{cases} 1 & if\ P(r_i|M) > t \\ 0 & if\ P(r_i|M) \leq t \end{cases} \quad (5)$$

6.2 Classification of association rules

This section presents rule selection for the formal example delineated in Figure 1. Twenty association rules $\{r_1, \ldots, r_{20}\}$ are extracted with minimal support: $\sigma_s = 1$ and minimal confidence: $\sigma_c = 0.1$ (cf. Table 2). For the rule r_7, we have: $P(\text{"}b\text{"} \Longrightarrow \text{"}a\text{"}\ \text{"}c\text{"}\ \text{"}e\text{"}|M) = (P_M(b,a) \times P_M(b,c) \times P_M(b,e))^{1/3} = (0.05 \times 0.3 \times 0.3)^{1/3} = 0.165$. In order to perform an evaluation of our methodology, we classified the rules into eight classes shown in table 2. The left part contains taxonomic rules (T-rules), and the right part contains non-taxonomic rules ($\neg$T-rules) i.e., rules involving at least one path that does not exist in the ontology used. The line blocks group rules according to their structures, i.e., the number of keyterms (one keyterm or more) involved in B and H: line 1 are $(1,1)$ rules; line 2 are $(1,n)$, line 3 are $(n,1)$, and line 4 are (n,m) with $n,m > 1$.

Table 2. Likelihood scores for association rules

#	T	n/d	score	#	$\neg$T	n/d	score
r_1	b⇒e	0/1	0.300	r_3	e⇒b	1/0	0.000
r_2	a⇒c	0/0	0.111	r_4	c⇒a	1/0	0.000
r_5	b⇒c,e	0/2	0.300				
r_6	a⇒b,c,e	0/2	0.176	r_{11}	c⇒b,e	2/0	0.100
r_7	b⇒a,c,e	1/2	0.165	r_{12}	c⇒a,d	2/0	0.100
r_8	e⇒b,c	1/1	0.163	r_{13}	d⇒a,c	2/0	0.100
r_9	a⇒c,d	0/1	0.157	r_{14}	c⇒a,b,e	3/0	0.100
r_{10}	e⇒a,b,c	2/1	0.121				
r_{15}	b,c⇒e	1/1	0.173	r_{16}	c,e⇒b	2/0	0.081
r_{17}	a,b⇒c,e	0/3	0.217	r_{19}	b,c⇒a,e	3/1	0.110
r_{18}	a,e⇒b,c	1/2	0.160	r_{20}	c,e⇒a,b	4/0	0.081

There exists a likelihood threshold $t = 0.110$ that separates the (T)-rules ($P(r_i|M) > 0.110$) from the ($\neg$T)-rules ($P(r_i|M) \leq 0.110$). In $T(1,1)$, both rules are taxonomic. According to the likelihood score definition, the longer the path is (length is 1 for r_1 and 2 for r_2), the lower the score is ($score(r_2) \ll score(r_1)$). Hence, r_2 is less trivial than r_1 (following paragraph 4, property II). There are three taxonomic paths from "a" to "c" (through "b", "e" or "b / e"). The path (a / b / e / c) is discarded as it is the longest path. "b" is a two-branches concept while "e" is only a one-branch concept. Thus, the likelihood score for (a / b / c) is kept instead of the (a / e / c) score as it is lower (illustrating property III). Conversely, in $\neg T(1,1)$ there is no path from "e" to "b" (in r_3) or from "c" to "a" (in r_4) then $score(r_3) = score(r_4) = 0$.

In the T-rules of line-blocks 2, 3, and 4, we can underline two principles: (i) the less non-taxonomic paths between the keyterm in B and the keyterm set in H are, the higher the likelihood score is; (ii) the more direct paths there are, the higher the score is. We indicate this in the n/d column of Table 2. For example in the rule r_{13}, 2/1 means that two paths of the association rule are non-taxonomic and there is 1 direct path linking concepts of B and H.

7 APPLICATION ON TEXTUAL DATA

We will now describe knowledge-based rule selection carried out on a large basis of association rules. The rules were extracted from bibliographical notes of scientific articles on molecular biology. They provided information on contextual data and meta-data encoded in XML elements, e.g., title, author(s), date, publication status (yes,no), keyterms (cf. Figure 2). Overall, our corpus has been composed of a set of 1,361 texts of about 240,000 words (1.6 MBytes). The texts were indexed by $|\mathcal{K}| = 632$ keyterms.

Text: **#391**
Title: Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained In vitro.
Authors: 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 and sparfbxacin 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 of both resistant strains, leading to a Ser83- ->Ile substitution (Escherichia coli numbering) in the corresponding protein. The gyrB, parC, and parE 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.
Keyterms: "characterization""chlamydia trachomatis""determine region""dna""escherichia coli""gyra gene""gyrase""gyrb gene""mutation""ofbxacin""parc gene""pare gene""point mutation""protein""quinolone""sparfbxacin""substitution""topoisomerase"

Figure 2. Example of a bibliographical note (abbreviated).

Initially, two textual fields were extracted: the title and the abstract. Next was an automatic indexing process required for extracting linguistically well-formed keyterms from the texts by using the FASTR tool [10]. Each text could then be represented by a set of keyterms, and text mining according to the classical knowledge discovery schema became feasible. Extraction of association rules was achieved by making use of the *Close* algorithm [13]. Close is based on the closed frequent itemset levelwise search in a boolean table describing a Cartesian product $\mathcal{T} \times \mathcal{K}$. The algorithm starts from the shortest frequent itemset closures and increments efficiently the calculation of larger frequent itemset closures in $\mathcal{K}$. An itemset is closed if it is the maximal set of keyterms shared by a set of texts. An itemset is frequent if it appears more than σ_s-times. Once the closures have been computed, the association rules are derived. We obtained 347 association rules with $\sigma_s = 10$ and $\sigma_c = 0.8$.

Association rule selection is conducted by following the knowledge-based approach described. The ontology used as a model M in knowledge-based rule selection is the UMLS metathesaurus [17]. It contains about 125,000 concepts from about 100 medical and biological thesauri. While the UMLS metathesaurus provides 11 relationships, we restrict our analysis to the IS-A relation ("PAR": parent). Note that the model M derived from the UMLS metathesaurus covers $\mathcal{K}$ only partially. Overall, 438 concepts of M are identical to those in $\mathcal{K}$ ($\approx 70\%$ of the keyterms). Thus, we have an incomplete model. Among the 347 association rules, there are 136 ($\approx 40\%$) that have a likelihood $\neq 0$ (all were complex rules). We have discarded them, and the remaining set is made up of 211 association rules (60%). Among them, there are 46 simple rules and 165 complex rules.

To evaluate the outcome of knowledge based rule selection we describe the behaviour of the binary classifier in terms of the signal detection theory. In a binary classification there are four possible outcomes: *true positives*, *false positives*, (hit rate), *false negatives* and *true negatives* (correct rejections). Remember that we pursue a rejection-oriented approach. Thus, non-rejected rules (positives) are not considered to be exclusively correct positives. When evaluating the outcome of knowledge-based rule selection we address the following questions.

1. Overall rejection rate. First, we have to consider the overall rejection rate. Only if the rate of rejected association rules is suffi-

ciently high, then it makes sense to continue the evaluation of association rule selection. If there is no reduction achieved, then the shrinkage of the rule set failed. In our data set, we achieved a rejection rate of 40%, viz., 136 association rules were rejected since their likelihood score indicated that they were taxonomic rules.

2. Impact of incomplete models on the process of rule selection. The power of models to reject rules (e.g., because they are taxonomic) is hampered by incomplete models, viz., models that do not cover all the concepts used by association rules. This means, that incomplete models lead to an increased rate of false positives that pass unnoticed. In fact, among the two types of errors, that may occur in this type of classification (false positives, false negatives) the proportion of the former was clearly higher (45%) than that of the latter (10%).

3. Validation by expert (analyst) ratings. Among the 136 rules that have been rejected on the basis of knowledge-based rule selection there have been 122 true negatives (correct rejections) (90%) and 14 false negatives (10%). Among the set of 211 association rules that were not rejected, there were 115 true positives (54%) and 96 false positives (45%). We used the 2×2 schema of signal detection theory to compare the expected frequencies (random distribution) with those obtained by the binary classifier. Here, a significant deviation became obvious ($\chi^2 = 87.47$, $\mathtt{df} = 3$, $\mathtt{p} < 0.0001$, two-tailed). This result is mainly due to the high score obtained for correct rejections. As expected, results for positives do not deviate from those of a random distribution. In sum, according to our expectation knowledge based selection of association rules discriminated highly above chance level on the set of negatives.

8 RELATED WORK

Most of the work expended to address the problem of association rule selection made use of statistical approaches without incorporating knowledge. An exception is the work of Basu and his co-researchers [2] who were also following a knowledge-based approach to reduce the overall set of rules extracted. Instead of using probabilistic framework, their work is based on a self-defined measure of semantic similarity between words. The main difference between our approach and the work of Basu and his colleagues relates to the possibilities for extensions. While similarity based approaches are oblivious of base rates, their inclusion is in fact a promising extension of likelihood based approaches once word statistics are available and included into text mining. Quite a different approach to improve the quality of rule selection follows generalised association rules mining [15, 8, 9], which proceeds by using terms that are part of different concept levels of an ontology. When knowing the ancestors of the terms some specific criteria are applied to constrain the rule mining process (e.g., disallowing rules that use both a term and its ancestor). Still, this process remains computationally expensive since using ancestors of each term in a preprocessing step or during the mining steps leads to an even greater number of rules. Similar work that used ontological knowledge, for term classification purposes, is presented in [14].

9 CONCLUSION AND FUTURE WORKS

Rejection-based selection of association rules provides new possibilities to improve rule mining by integrating knowledge. The probabilistic foundations of this approach allow the data mining expert to profit from a broad range of methods drawn from probability theory (e.g., Bayesian techniques). To fully exploit the probabilistic framework for rule selection we plan to extend our work in two directions. First, we will integrate base rates of concepts (e.g., based on word statistics) that provide a weighting schema for the likelihoods already used. By pursuing a Bayesian framework that makes use of priors the usage of word statistics become feasible. Second, the examples presented in the paper made only use of simple ontologies that were built upon the hypernym-hyponym relationship. However, more complex ontologies may also be processed (e.g., incorporating causal relationships).

REFERENCES

[1] R. Agrawal and R. Srikant, 'Fast algorithms for mining association rules in large databases', in *Proc. of the 20^{th} Int'l Conf. on Very Large Databases (VLDB'94)*, pp. 478–499, Santiago, Chile, (1994). Extended version: IBM Research Report RJ 9839.

[2] S. Basu, R. J. Mooney, K. V. Pasupuleti, and J. Ghosh, 'Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge', in *Proc. of the 7^{th} Int'l Conf. on Knowledge Discovery and Data Mining (KDD'01)*, pp. 233–238, San Francisco, USA, (2001). ACM Press.

[3] J. Bell and M. Machover, *A course in mathematical logic*, North-Holland, Amsterdam, 1986.

[4] A. Collins and E. Loftus, 'A Spreading-Activation Theory of Semantic Processing', *Psychological Review*, **82**(6), 407–428, (1975).

[5] M. Delgado, M. J. Martin-Bautista, D. Sanchez, and M.A. Vila, 'Mining Text Data: Special Features and Patterns', in *Pattern Detection and Discovery: Proc. of ESF Exploratory Workshop*, eds., D.J. Hand, N.M. Adams, and R.J. Bolton, volume 2447 of *Lecture Notes in Artificial Intelligence – LNAI*, pp. 140–153, London, (2002). Springer-Verlag.

[6] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'From data mining to knowledge discovery', *AI Magazine*, **17**(3), 37–54, (1996).

[7] R. Feldman and H. Hirsh, 'Exploiting Background Information in Knowledge Discovery from Text', *Journal of Intelligent Information Systems*, **9**(1), 83–97, (1997).

[8] J. Han, 'Mining Knowledge at Multiple Concept Levels', in *Proc. of 4^{th} the Int'l Conf. on Information and Knowledge Management (CIKM'95)*, pp. 19–24, Baltimore, USA, (1995). ACM Press. Invited talk.

[9] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, 'Data mining of association rules and the process of knowledge discovery in databases', in *Data Mining in E-Commerce, Medicine, and Knowledge Management*, 15–36, Springer, Heidelberg, (2002).

[10] C. Jacquemin, 'FASTR : A Unification-Based Front-End to Automatic Indexing', in *Proc. of Information Multimedia Information Retrieval Systems and Management*, pp. 34–47, New-York, (1994). Rockfeller University Press.

[11] D. Krantz, R. Luce, P. Suppes, and A. Tversky, *Foundations of Measurement*, volume 1, Academic Press, San Diego, USA, 1971.

[12] N. Lavrač, P. Flach, and B. Zupan, 'Rule Evaluation Measures: A Unifying View', in *Proc. of the 9^{th} Int'l Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 of *Lecture Notes in Artificial Intelligence – LNAI*, pp. 174–185, Bled, Slovenia, (1999). Springer-Verlag, Heidelberg. Co-located with ICML'99.

[13] 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).

[14] P. Resnik, 'Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language', *Artificial Intelligence Research*, **11**, 95–130, (1999). Morgan Kaufmann Publishers.

[15] R. Srikant and R. Agrawal, 'Mining Generalized Association Rules', in *Proc. of the 21^{st} Int'l Conf. on Very Large Databases (VLDB'95)*, pp. 407–419, Zurich, (1995). Morgan Kaufmann Publishers.

[16] P.-N Tan, V. Kumar, and J. Srivastava, 'Selecting the right interestingness measure for associaton patterns', in *Proc. of the 8^{th} ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pp. 183–193, Edmonton, Canada, (2002). ACM Press.

[17] UMLS. The Unified Medical Language System. National Library of Medicine, 11^{th} edition, 2000.