Knowledge-based Selection of Association Rules for Text Mining
Dietmar Janetzko, Hacène Cherfi, Roman Kennke, Amedeo Napoli, Yannick Toussaint

To cite this version:

HAL Id: inria-00107787
https://inria.hal.science/inria-00107787
Submitted on 19 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Knowledge-based Selection of Association Rules for Text Mining

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

Abstract. A reoccurring 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 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 \( \mathcal{T} = \{ \tau_1, \ldots, \tau_n \} \) be a non-empty finite set of texts. Likewise, let \( \mathcal{K} = \{ k_1, \ldots, k_n \} \) 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 \( B \Longrightarrow H \) where \( B \) stands for body or antecedent, and \( H \) for head or consequent with \( B \subseteq \mathcal{K}, H \subseteq \mathcal{K} \) and \( B \cap H = \emptyset \). Let

\[ B \subseteq \mathcal{K}, H \subseteq \mathcal{K} \text{ and } B \cap H = \emptyset. \]
\[ \mathcal{B} = \{b_1, \ldots, b_n\} \] be the set of keyterms of the body of an association rule \( r \), and \( \mathcal{H} = \{h_1, \ldots, h_m\} \) be the set of terms of the head of \( r \). \( \mathcal{B} \Rightarrow \mathcal{H} \) means that all the texts in \( \mathcal{T} \) containing the keyterms \( b_1, b_2, \ldots, b_n \) also contain the keyterms \( h_1, h_2, \ldots, h_m \) with a probability \( P \). The support of \( r \) is the number of texts containing the keyterms in \( \mathcal{B} \cup \mathcal{H} = \{b_1, \ldots, b_n, \ldots, b_n\} \). This ratio is interpreted as the conditional probability \( P(\mathcal{H}|\mathcal{B}) \).

We want an association rule to conform to a model provided that the concepts of \( \mathcal{B} \) and \( \mathcal{H} \) 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 \( \text{apple} \Rightarrow \text{fruit} \) is a strong candidate for rejection because it is strictly taxonomic. Conversely, we would like to see rules that express an interesting combination of "cherry pie", "chocolate" and "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., [1], p. 8). Networks of concepts (e.g., hierarchies, ontologies, meronymic, 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 \( C = \{c_1, \ldots, c_n\} \) be a non-empty finite set of concepts, and let \( \mathcal{R} = \{R_1, \ldots, R_p\} \) be non-empty finite set of relations, then \( \mathcal{R}^C \) symbolises the set of pairs \( (C \times C) \) for which the relation \( R_i \) holds. Thus, a relational structure is made up of the set of pairs \( \mathcal{R}^C \) together with the set of concepts \( C \). For \( i \in \{1, \ldots, n\} \) the relational structure is defined by \( (C, R_i) \). Whenever all the \( \mathcal{R}^C \) for \( i \in \{1, \ldots, n\} \) are defined on \( C \), we simply write \( (C, R) \). If a relational structure \( (C, \mathcal{R}^C) \) 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 \( \mathcal{K} \) of the concepts which are used in association rules. This is necessary since we cannot 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 \( r \) and a model. Let the set of concepts common to a rule \( r \) and a model \( M \) be denoted by \( Z_r \). Then, a rule \( r \) matches a model \( M \) if \( \forall j \), a keyterm of a rule \( r \) \( \exists c \in C \) such that \( Z_r \neq \emptyset \) with \( Z_r = \left\{ (k_1 = c_1), \ldots, (k_n = c_n) \right\} \).

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 \( P(r|M) \), i.e., the probability of an association rule \( r \) given a model \( M \). To specify \( P(r|M) \) 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 \( \mathcal{B} \) and the keyterms of \( \mathcal{H} \) of a rule \( r \). This is only possible if the concepts of the model \( M \) can be related to the concepts of a rule \( r \). Thus, \( \mathcal{B} \cap \mathcal{K} \neq \emptyset \). If there are several alternative paths, the shortest one is chosen. A failure to find a path that connects the concepts of \( \mathcal{B} \) and \( \mathcal{H} \) within a model \( M \) is denoted by \( \delta \). On the other hand, we have to specify a match between the path between the concepts \( c_a \) and \( c_v \) within a model \( M \) is defined by \( \ell(c_a, c_v) \). Accordingly, \( \min(c_a, c_v) \) is used to refer to the minimal path length between a pair of concepts of \( \mathcal{B} \) and \( \mathcal{H} \). The expression \( [c_a, c_v] \) is used to refer to the number of paths that go from a concept \( c_a \) to all concepts in \( C \) and thus in \( M \). Hence, \( (c_a, c_v) \subseteq \mathcal{C} \times \mathcal{C} \) for which the relation \( R_i \) holds. If we want to express that the transition of two concepts \( c_a \) and \( c_v \) of a model is used to calculate what is called the likelihood of a specific association rule \( r \), we indicate this by writing \( c_a^r \) and \( c_v^r \).

II. Calculating the Decay Rate Attached to the Minimal Path Lengths. Let us consider the situation that there is in fact a path between \( c_a \) 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 \( \delta : \mathbb{N} \rightarrow \mathbb{R} \). The function \( \delta \) is a weighting mechanism that punishes long paths between a concept \( c_a \in \mathcal{B} \) and \( c_v \in \mathcal{H} \) of a model \( M \). It is calculated by using the reciprocal of the length of a path between \( c_a \) and \( c_v \) such that \( \delta = 1/\ell(c_a, c_v) \).

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 \( c_a \) of \( \mathcal{B} \) to each concept of \( M \) should sum up to 1. What is needed is a variable to accomplish this standardisation. We take \( \xi \) to denote this standardisation score. Note that \( \xi \) is calculated \( \forall c \in C \). Intuitively, we say that as a result of the \( \xi \)-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 denoted by \( \xi \). For each \( c_a \), we compute \( \xi \) by adding up all scores for the reciprocal of path lengths and calculating the reciprocal of the resulting sum, i.e., for the concept \( c_a \):

\[
\xi = \left( \frac{1}{\sum_{i=1}^{l(c_a, c_v)} \frac{1}{\ell(c_a, c_v)}} \right)^{-1}
\]

(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:

\[
P_M(c_a, c_v) = \xi \cdot \delta \cdot \min(c_a, c_v)
\]

(2)

Based on equation 2, we may specify the probability of a rule \( r \) given
a model \( M \). This score is what we call the likelihood of a rule, which expresses the goodness of fit between an association rule \( r_1 \) and a model \( M \). Do the likelihood scores reflect correctly the goodness of fit between a rule \( r_1 \) 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 considered rules. However, the number of transitions of the concept \( c_i \) 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 \( r_p \) and \( r_q \) 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 \( L(c_i^p, c_i^q) = L(c_i^q, c_i^p) \), i.e. if we keep \( \delta \) and thus the lengths of paths constant (situation i), then it follows from equation 2:

\[
[L(c_i^p, c_i^q)] > [L(c_i^q, c_i^p)] \Rightarrow P(r_p | M) < P(r_q | M).
\]

### 5 Likelihood Calculation

We give now give an 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 model of a simple ontology is depicted in Figure 1 (b), that is read as follows: “\( a \) is an \( \alpha \) "b", "c" is an \( \alpha \) "e", etc. The \( \alpha \) \( \alpha \) relation is reflexive and transitive but this is not indicated on the figure (for simplicity).

![Figure 1](image-url)

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

### Table 1

<table>
<thead>
<tr>
<th>Text</th>
<th>Keyterms</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>{acd}</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>{bce}</td>
</tr>
<tr>
<td>( t_3 )</td>
<td>{abc}</td>
</tr>
<tr>
<td>( t_4 )</td>
<td>{be}</td>
</tr>
<tr>
<td>( t_5 )</td>
<td>{abc}</td>
</tr>
<tr>
<td>( t_6 )</td>
<td>{bce}</td>
</tr>
</tbody>
</table>

**Figure 1.** Figure written in terms of transition probabilities

<table>
<thead>
<tr>
<th>( t )</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( d )</th>
<th>( e )</th>
<th>( \Sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a )</td>
<td>0.22</td>
<td>0.22</td>
<td>0.11</td>
<td>0.22</td>
<td>0.22</td>
<td>1</td>
</tr>
<tr>
<td>( b )</td>
<td>0.05</td>
<td>0.30</td>
<td>0.30</td>
<td>0.05</td>
<td>0.30</td>
<td>1</td>
</tr>
<tr>
<td>( c )</td>
<td>0.10</td>
<td>0.10</td>
<td>0.60</td>
<td>0.10</td>
<td>0.10</td>
<td>1</td>
</tr>
<tr>
<td>( d )</td>
<td>0.10</td>
<td>0.10</td>
<td>0.60</td>
<td>0.10</td>
<td>0.10</td>
<td>1</td>
</tr>
<tr>
<td>( e )</td>
<td>0.06</td>
<td>0.06</td>
<td>0.40</td>
<td>0.06</td>
<td>0.40</td>
<td>1</td>
</tr>
</tbody>
</table>

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 \( b_k \in B \) to \( e_j \in E \) given a model \( M \) by \( P_{b_k|e_j} | M \). 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). Conversely, pairs of concepts connected by long paths, e.g., \((b, d)\), are associated with small transition probabilities (0.05). Once we have calculated the probability distribution of a model \( M \), the determination of \( P(r_1 | M) \), i.e., the likelihood of a simple association rule with \( [B] = [H] = 1 \) 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: “\( b \) \( \Rightarrow \) "c", we have that \( P(b \Rightarrow c | M) = 0.30 \). However, if we want to calculate the likelihood of complex association rules with \( [B] \) or \( [H] > 1 \), 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 \( B_1 \times B_1 \) of the one-step transition probabilities of the concepts involved in an association rule \( r_1 \).

### 6 Classification

#### 6.1 A knowledge-based classifier

A high likelihood of an association rule \( r_1 \) 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 \( r_1 \) and a model \( M \) both of which...
are used to calculate the likelihood score \( P(r|M) \). In addition, we make use of a threshold \( t \) to carry out a binary classification:

\[
f(r_t, M) = \begin{cases} 
1 & \text{if } P(r|M) > t \\ 
0 & \text{if } P(r|M) \leq t 
\end{cases}
\]

(5)

### 6.2 Classification of association rules

This section presents rule selection for the formal example delineated in Figure 1. Twenty association rules 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, 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-

---

**Text:** Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained in vitro.

**Title:** Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained in vitro.

**Author:** Dessus-Babus-S, Bebear-CM, Charron-A, Bebear-C, de-Barbeyrac-B.

**Abstract:** The L2 reference strain of Chlamydia trachomatis was exposed to sub-hibitory concentrations of ofloxacin and sparfloxacin to select flothroquino-lone-resistant mutants. In this study, two resistant strains were isolated after four rounds of selection \( \{ A, B \} \) point mutation was found in the gyrA gyrB quinolone-resistance-determining region of both resistant strains, leading to a Ser83 -&gt; 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 Ch. trachomatis, DNA gyrase is the primary target of ofloxacin and sparfloxacin.

**Keywords:** "characterization" "Chlamydia trachomatis" "determine region" "DNA" "Escherichia coli" "gyrA gene" "gyrB gene" "mutation" "ofloxacin" "parC gene" "parE gene" "point mutation" "protein" "quinolone" "sarfloxacin" "substitu-
tion" "topoisomerase".

![Example of a bibliographical note (abbreviated).](image)

---

**Table 2.** Likelihood scores for association rules

<table>
<thead>
<tr>
<th>#</th>
<th>( \tau )</th>
<th>nil</th>
<th>score</th>
<th>#</th>
<th>(-\tau)</th>
<th>nil</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_1 )</td>
<td>b&amp;e</td>
<td>0/1</td>
<td>0.300</td>
<td>( r_3 )</td>
<td>e&amp;b</td>
<td>1/0</td>
<td>0.100</td>
</tr>
<tr>
<td>( r_3 )</td>
<td>b&amp;c,e</td>
<td>0/2</td>
<td>0.300</td>
<td>( r_4 )</td>
<td>e&amp;b</td>
<td>0/1</td>
<td>0.000</td>
</tr>
<tr>
<td>( r_5 )</td>
<td>a&amp;b,c,e</td>
<td>0/2</td>
<td>0.176</td>
<td>( r_11 )</td>
<td>c&amp;b,e</td>
<td>2/0</td>
<td>0.100</td>
</tr>
<tr>
<td>( r_7 )</td>
<td>b&amp;a,c,e</td>
<td>1/1</td>
<td>0.163</td>
<td>( r_12 )</td>
<td>d&amp;a,c,d</td>
<td>2/0</td>
<td>0.100</td>
</tr>
<tr>
<td>( r_8 )</td>
<td>e&amp;a,b,c</td>
<td>1/1</td>
<td>0.163</td>
<td>( r_13 )</td>
<td>d&amp;a,c,e</td>
<td>2/0</td>
<td>0.100</td>
</tr>
<tr>
<td>( r_9 )</td>
<td>a&amp;c,d</td>
<td>0/1</td>
<td>0.157</td>
<td>( r_14 )</td>
<td>e&amp;a,b,e</td>
<td>3/0</td>
<td>0.100</td>
</tr>
<tr>
<td>( r_{13} )</td>
<td>b&amp;a,c,e</td>
<td>1/1</td>
<td>0.173</td>
<td>( r_{16} )</td>
<td>c&amp;b,e</td>
<td>2/0</td>
<td>0.081</td>
</tr>
<tr>
<td>( r_{17} )</td>
<td>a,b&amp;c,e</td>
<td>0/3</td>
<td>0.217</td>
<td>( r_{19} )</td>
<td>b&amp;a,c,e</td>
<td>3/1</td>
<td>0.110</td>
</tr>
<tr>
<td>( r_{18} )</td>
<td>a&amp;b&amp;c,b</td>
<td>1/2</td>
<td>0.160</td>
<td>( r_{20} )</td>
<td>c&amp;e,a,b</td>
<td>4/0</td>
<td>0.081</td>
</tr>
</tbody>
</table>

There exists a likelihood threshold \( t = 0.110 \) that separates the \( \{\tau\} \)-rules \( (P(r|M) > 0.110) \) from the \( \{-\tau\} \)-rules \( (P(r|M) \leq 0.110) \). In \( \{\tau\}, \) both rules are taxonomic. According to the likelihood score definition, the longer the path is (length is \( n \)) the more direct paths there are, the higher the score will be. Conversely, in \( \{-\tau\} \) the less non-taxonomic paths between the keyterm in \( \{\tau\} \)-rules and line-blocks group rules according to their structures, i.e., the number of keyterms (one keyterm or more) involved in the \( B \) 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 \( n, m > 1 \).

---

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

---

**Figure 2.** Example of a bibliographical note (abbreviated).
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 ($\chi^2 = 87.47$, df = 3, $p < 0.0001$, 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


