# Maximal Association Rules: A Tool for Mining Associations in Text

**Article** *in* Journal of Intelligent Information Systems · November 2005

DOI: 10.1007/s10844-005-0196-9 · Source: DBLP



**4 authors**, including:

Yonatan Aumann
Bar Ilan University
**113** PUBLICATIONS **2,208** CITATIONS
SEE PROFILE

Ronen Feldman
Hebrew University of Jerusalem
**129** PUBLICATIONS **4,113** CITATIONS
SEE PROFILE


Journal of Intelligent Information Systems, 25:3, 333–345, 2005


# Maximal Association Rules: A Tool for Mining Associations in Text

AMIHOOD AMIR amir@cs.biu.ac.il
YONATAN AUMANN aumann@cs.biu.ac.il
RONEN FELDMAN feldman@cs.biu.ac.il
MOSHE FRESKO freskom1@cs.biu.ac.il
*Department of Computer Science, Bar Ilan University, Ramat Gan, 52900, Israel*

*Received September 24, 2003; Revised September 26, 2004; Accepted September 28, 2004*

**Abstract.** We describe a new tool for mining association rules, which is of special value in text mining. The new tool, called *maximal associations*, is geared toward discovering associations that are frequently lost when using regular association rules. Intuitively, a maximal association rule $X \stackrel{\max}{\Longrightarrow} Y$ says that whenever $X$ is the only item of its type in a transaction, than $Y$ also appears, with some confidence. Maximal associations allow the discovery of associations pertaining to items that most often do not appear alone, but rather together with closely related items, and hence associations relevant only to these items tend to obtain low confidence. We provide a formal description of maximal association rules and efficient algorithms for discovering all such associations. We present the results of applying maximal association rules to two text corpora.

**Keywords:** text mining, association rules, data mining

## 1. Introduction

***Motivation.*** Knowledge Discovery in Databases (KDD) has been defined as: "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al., 1996). Accordingly, a key tool in KDD is the *association rule*, aimed at discovering and representing hidden associations in the data. Association rules have been researched and applied extensively, in diverse domains and applications (Lee et al., 1999; Bench-Capon et al., 2000; Michail, 2000; Satou et al., 1997; Brijs et al., 1999; Ma et al., 2000; Dong et al., 2000). However, while association rules provide means to discover many interesting associations, they fail to discover other, no less interesting associations, also hidden in the data. In this paper, we consider a new type of association rule, which we call *maximal association rule*, which allows to discover interesting associations otherwise not captured by regular association rules.

***Example.*** As an example, consider a hospital database of diseases and symptoms, wherein each record corresponds to a single patient in a single admission. Suppose that there is one disease, say $A$, that is very common, appearing in 50% of the records, and another much less common disease, say $B$, appearing in only 10% of the records. Suppose further that $A$ is marked by multiple symptoms, $x$, $y$ and $z$, and that $B$ is marked by either $x$ or $y$, but

not both. If we were to search for regular associations, we may get the association between $A$ and $x, y, z$, but we would miss the crucial diagnostic association linking $x$ *alone* and $y$ *alone* to $B$. The reason is that the many instances of $A$ with $x$, $y$ and $z$, reduce the confidence of the rules "$x \Rightarrow B$" and "$y \Rightarrow B$". In order to obtain the association "$x \Rightarrow B$" we need to capture the notion that whenever $x$ appears *alone* then $B$ also appears, with high confidence (note that the rule "$B \Rightarrow x$" need not be true, since $B$ frequently appears with $y$). Regular association rules fail to capture such associations. *Maximal Association rules*, introduced in this paper, are designed to capture these cases.

***Maximal association rules and text mining.*** One application where maximal association rules are most useful is in text mining. Text mining is concerned with applying data mining techniques to text. Association rules have been applied to text in the following way. Each document or sentence of the text is associated with a relevant set of *terms*. These terms are either categories associated with the text (Feldman and Hirsh, 1996), or words and phrases from the text, obtained either by taking all the words or phrases of the text (see Rajman and Besancon, 1997), or the outcome of some type of linguistic and/or statistical analysis (see Rajman and Besancon, 1997; Feldman et al., 1998). The sets of terms thus obtained define the underlying transaction database (one transaction for each document or sentence). Association rules are then obtained from this database.

In text documents, closely related items frequently appear together. For example, "Linux" may appear most often together with "Windows", and Condoleezza Rice together with George Bush. The result is that associations specifically relevant to one item, but not the other, tend to have low confidence. For example, an association between "Linux" and "open source" would have low confidence, because of the many transactions where "Linux" appears together with "Windows" (and without "open source"); or an association between Condoleezza Rice and Stanford University, would be lost because of the large number of transactions where Ms. Rice appears together with George Bush (without Stanford). Maximal associations provide means to overcome this problem. Maximal Associations are designed to pinpoint associations relevant to one item but not the other, i.e. associations pertaining to "Linux" or Ms. Rice *alone*. In Section 4 we provide results of applying maximal association rules to two sample text corpora.

***Maximal associations and regular associations.*** We note that maximal association rules are not designed to replace regular association rules, but rather to complement them. Using only maximal association rules, many interesting regular associations may and will be lost. In Section 4 we show that maximal association rules discover interesting associations that would not have been found using regular associations. The converse is also true. In one of the data sets, for example, more than half of the regular association rules were not found as maximal associations (with the same support and confidence).

***Outline.*** The rest of this paper is organized as follows. In the remainder of this section we consider related work. In Section 2 we provide the formal definition of maximal association rules. In Section 3 an efficient algorithm for finding all maximal association rules is provided. In Section 4 we present the results of applying maximal association rules to two text corpora.

### 1.1. Related work

Association rules were introduced by Agrawal et al. (1993), and have been extensively researched ever since. Most of research is focused on developing fast and efficient algorithms for mining of the associations; the reader is referred to Hipp et al. (2000) for a comprehensive survey.

Variants of the basic definition of association rules have also been considered in several papers. We briefly review some of these works. (Srikant and Agrawal, 1997) introduce the notion of mining *generalized* association rules, and provide efficient algorithms for mining such associations. A generalized association rule is a rule that, given a taxonomy (a *is-a* hierarchy) over the items, provides associations between elements from any level of the taxonomy (e.g. "customer buys outerwear ⇒ customer buys hiking boots" rather than "customer buys down jacket ⇒ customer buys hiking boots"). Multiple-level association rules are also discussed by Han and Fu (1999). Cai et al. (1998) and Tao et al. (2003) consider the problem of mining associations rules where items may have different *weights* (i.e. different levels of importance), and provide efficient algorithms for discovering such rules.

Wang et al. (2000) consider a different notion of *weight*, where the weight of an item in a transaction determines the "strength" of the item in the transaction (e.g. *number* of beer bottle purchased in the transaction). Liu et al. (1999) consider the problem of mining association rules with multiple support thresholds. Tung et al. (1999) consider associations that span across multiple transactions, rather than in the same transaction (e.g. "if the movie has many viewers in NY then it has few viewers in Paris"). Brin et al. (1997) consider the generalization of association rules to general *correlations*. Mannila et al. (1995) consider the problem of discovering frequent *episodes*, where an episode is a collection of events occurring within a given time window.

Mining for association rules in text was first considered in Feldman and Dagan (1995), Feldman and Hirsh (1996) and Feldman et al. (1998). Rajman and Besancon (1997) consider different methods for associating terms with the text, and the introduction of natural langauge processing techniques to association rule mining in text. Feldman et al. (1998) show how to associate terms with the text based on extracting key terms and phrases from the text. Ahonen et al. (1997) describe a general framework for text mining of frequent episodes from the text. Wong et al. (1999) describe a system for visualizing association rules for textual data. A general overview on text mining can be found in Hearst (1999).

An early version of this paper appeared in Feldman et al. (1997).

## 2. Definitions

Let $S = \{a_1, \ldots, a_n\}$ be a set of literals, called *items*. A *transaction*, $t$, over $S$ is a subset $t \subseteq S$. A *database, D, over S* is a multiset of transactions over $S$. A *grouping, G, of S* is a division of $S$ into disjoint sets, $G = \{g_1, \ldots, g_k\}$. We call the elements of $G$ *categories*. For an item $a$, we denote by $g(a)$ the category that contains $a$. Similarly, if $X$ is a set of items all of which are from a single category, then we denote this category by $g(X)$.

We first review the definition of regular association rules, as defined in Agrawal et al. (1993). For a transaction, $t$, and a set of items $X$, we say that $t$ *supports* $X$ if $X \subseteq t$. The *support* of $X$ in a database $D$, denoted by $s_D(X)$, is the number of transactions $t \in D$ that support $X$. Support can also be measured on a percentage basis. The *percentage support* of $X$ in $D$ is the percentage of transactions $t \in D$ that support $X$.

An *association rule* is a rule of the form $X \Rightarrow Y$, where $X$ and $Y$ are disjoint sets of items from $S$. The *support* of the association $X \Rightarrow Y$, denoted by $s_D(X \Rightarrow Y)$, is the support of $X \cup Y$, and the *confidence* of the association, denoted by $c_D(X \Rightarrow Y)$, is $s_D(X \cup Y)/s_D(X)$.

In a maximal association rule $X \overset{\max}{\Longrightarrow} Y$ we are interested in capturing the notion that whenever $X$ appears *alone* then $Y$ also appears, with some confidence. For this, we must first define the notion of *alone*. We do so with respect to the categories of $G$, as follows.

For a transaction $t$, a category $g_i$ and an itemset $X \subseteq g_i$, we say that $X$ *is alone in* $t$ if $t \cap g_i = X$. That is, $X$ is alone in $t$ if $X$ is the largest subset of $g_i$ which is in $t$. In this case we also say that $X$ is *maximal in* $t$, and that $t$ M-*supports* $X$. For a database $D$, the M-*support of* $X$ *in* $D$, denoted $s_D^{\max}(X)$ is the number of transactions $t \in D$ that M-support $X$.

A *maximal association rule*, or M-association, is a rule of the form $X \overset{\max}{\Longrightarrow} Y$ where $X$ and $Y$ are subsets of distinct categories, $g(X)$ and $g(Y)$, respectively. The M-*support* of the maximal association $X \overset{\max}{\Longrightarrow} Y$, denoted by $s_D^{\max}(X \overset{\max}{\Longrightarrow} Y)$, is defined as:

$$s_D^{\max}(X \overset{\max}{\Longrightarrow} Y) = |\{t : t \text{ M-supports } X \text{ and } t \text{ supports } Y\}|$$

That is, $s_D^{\max}(X \overset{\max}{\Longrightarrow} Y)$ is the number of transactions in $D$ that M-support $X$ and also support $Y$ in the regular sense. The intuitive meaning of the rule $X \overset{\max}{\Longrightarrow} Y$ is that whenever a transaction M-supports $X$, then $Y$ also appears in the transaction, with some probability. However, in measuring this probability we are only interested in those transactions where some element of $g(Y)$ (the category of $Y$) appears in the transaction. Accordingly, we define confidence as follows. Let $D(X, g(Y))$ be the subset of the database $D$ consisting of all the transactions that M-support X and contain at least one element of $g(Y)$. The M-*confidence* of the rule $X \overset{\max}{\Longrightarrow} Y$, denoted by $c_D^{\max}(X \overset{\max}{\Longrightarrow} Y)$, is defined as:

$$c_D^{\max}(X \overset{\max}{\Longrightarrow} Y) = \frac{S_D^{\max}\left(X \overset{\max}{\Longrightarrow} Y\right)}{|D(X, g(Y))|}$$

We search for associations where the M-support is above some user-defined *minimum* M-*support*, denoted by $\hat{s}$, and the M-confidence is above some user-defined *minimum confidence*, denoted by $\hat{c}$. A set $X$ with M-support at least $\hat{s}$ is said to be M-*frequent*.

Any maximal association rule is also a regular association, with perhaps different support and confidence. We define the M-*factor* of the rule $X \overset{\max}{\Longrightarrow} Y$ to be the ratio between the M-confidence of the maximal association $X \overset{\max}{\Longrightarrow} Y$ and the confidence of the *corresponding* regular association $X \Rightarrow Y$. Specifically, let $D'$ be the sub-set of transaction that contain at

least one item of $g(Y)$. Then, M-factor of the association $X \overset{\max}{\Longrightarrow} Y$ is defined as

$$\textit{M-factor}(X \overset{\max}{\Longrightarrow} Y) = \frac{C_D^{\max}\left(X \overset{\max}{\Longrightarrow} Y\right)}{c_{D'}(X \Rightarrow Y)}$$

Note that in defining the M-factor the denominator is the confidence on the rule $X \Rightarrow Y$ *with respect to* $D'$. The reason is that since the M-confidence is defined with respect to $D'$, the comparison to regular associations must also be with respect to this set. In general, we seek M-associations with a higher M-factor, which are the more interesting rules.

***Example.*** Consider the following database $D$ consisting of the 10 transactions:

| ID | Transaction |
|---|---|
| 1 | A,B,x,y,1 |
| 2 | A,B,D,u,z,1,2,3 |
| 3 | A,B,C,z,1 |
| 4 | A,B,x,y,z,2,3,4 |
| 5 | C,z,2,3 |
| 6 | A,B,u,1,3 |
| 7 | C,D,z,1,2 |
| 8 | A,B,u,x,y,4 |
| 9 | A,D,z,2,4 |
| 10 | A,B,x,y,z,1 |

We group the elements into three categories:

- "Capitals"={A,B,C,D},
- "Lowercase"={u,x,y,z},
- "Digits" = {1,2,3,4}.

With $\hat{s} = 3$ and $\hat{c} = 75\%$, we have the following maximal associations:

1. A,B $\overset{\max}{\Longrightarrow}$ x (M-support = 4, M-confidence = 80%).
2. 1 $\overset{\max}{\Longrightarrow}$ A,B (M-support = 3, M-confidence = 100%).
3. z $\overset{\max}{\Longrightarrow}$ C (M-support = 3, M-confidence = 75%).

Note that as regular associations the above rules (1–3) have confidence: 57, 83 and 38%, respectively. Thus, with a 50% confidence thresholds, only the second rule would be obtained as a regular association rule.

***Additional types of maximal associations.*** Note that in the definition of maximal association rules the antecedent is maximal, but the consequent need not be maximal. Thus, a

maximal rule $X \stackrel{\max}{\Longrightarrow} Y$ says that if $X$ appears *alone*, then $Y$ also appears, not necessarily alone. We note that alternative definitions are also possible, i.e. requiring maximality for both sides, or just for the consequent. Any of these alternative definitions would constitute a mathematically valid definition. In this paper we chose the definition as provided above since, in our experience, it provides the most significant and interesting rules.

## 3. Computing maximal association rules

Computing maximal association rules is, in general, much faster than computing regular associations (provided that the number of categories is not too big). The reason is that for each category, any transaction M-supports at most one item-set. Thus:

1. All M-frequent sets can be generated in a small number of passes over the database (see below).
2. For each category, the M-frequent sets *partition* the database into small sub-databases. Any M-association rule is fully supported by transactions from only one of these sub-databases. Thus, each sub-database can be considered separately (once the M-frequent sets have been generated).

A detailed pseudo-code description of the algorithm for discovering all M-associations is provided in figure 1. All integer variables are assumed to be initialized to zero, and all sets to the empty set. The algorithm starts by generating all M-frequent sets, using the `M-Frequent-Sets()` procedure (line 1). For each M-frequent set $X$ and category $g$, `M-Frequent-Sets()` also generates the sub-databases $D(X, g)$. Recall that the sub-database $D(X, g)$ consists of the projection on category $g$ of the transactions M-supporting $X$ (i.e. for each transaction $t$ M-supporting $X$, $D(X, g)$ includes the portion of $t$ that is in $g$). The M-frequent sets constitute the antecedent (left-hand-side) of the rules.

Next, the algorithm computes the consequence of the rules (the right-hand-side). Consider an M-frequent set $X$, and suppose $X \stackrel{\max}{\Longrightarrow} Y$ is an M-association , where $Y \subseteq g$. Let $D'$ be the sub-database $D(X, g)$. Clearly, only the transactions *within* $D'$ can possibly support $X \stackrel{\max}{\Longrightarrow} Y$. Moreover, suppose that the support of $Y$ within $D'$ is $s_{D'}(Y)$. Then the M-support of $X \stackrel{\max}{\Longrightarrow} Y$ is $s_{D'}(Y)$, and the M-confidence of the rule is $s_{D'}(Y)/|D'|$. Thus, in order to find all M-associations with minimum M-support $\hat{s}$ and minimum M-confidence $\hat{c}$, we need only search within $D'$ for all sets $Y$ with support $\max(\hat{s}, \hat{c} \cdot |D'|)$ (lines 5–6). Generating all such frequent sets in $D'$ can be performed using any known algorithm for computing frequent sets.

The procedure `M-Frequent-Sets()` finds all M-frequent sets using two passes over the database. The first pass is aimed at reducing the number of candidate M-frequent sets, while the second pass generates the sets, and the corresponding sub-databases $D(X, g)$. The first pass employs a hash table that hashes item-sets $X$ into an array of integers, initialized to 0. The hash table is of size to fit in memory, and thus several distinct item-sets $X$ may need to map to the same location. In the first pass (lines 2–8) the database is scanned in

**Maximal Associations**$(D, G, \hat{s}, \hat{c})$
$D$ - Database,
$G$ - grouping of literals to categories,
$\hat{s}$- minimum M-support threshold,
$\hat{c}$- minimum M-confidence threshold.

```
1   M ← M-Frequent-Sets(ŝ)
2   foreach X ∈ M do
3       foreach g ∈ G do
4           D' ← D(X, g)
5           s̄ ← max(ŝ, ĉ · |D'|)
6           F ← Frequent-Sets(D', δ)
7           foreach Y ∈ F do
8               Output X ⟹^max Y
9               Output M-support = s_D'(Y), M-confidence = s_D'(Y) / |D'|)
10          end foreach
11      end foreach
12  end foreach
```

`M-Frequent-Sets(`$\hat{s}$`)` (find all sets with M-support at least $\hat{s}$)

```
1   Large ← ∅
2   foreach t ∈ D do
3       foreach g ∈ G do
4           X ← t ∩ g
5           if X ≠ ∅ then
6               Hash(X) + +
7       end foreach
8   end foreach
9   foreach t ∈ D do
10      foreach g ∈ G do
11          X ← t ∩ g
12          if X ≠ ∅ and Hash(X) ≥ ŝ then
13              s^max(X) + +
14              if s^max(X) ≥ ŝ then
15                  Large ← Large ∪ {X}
16              foreach g' ∈ G if g' ∩ t ≠ ∅ then
17                  D(X, g') ← D(X, g') ∪ {t ∩ g'}
18      end foreach
19  end foreach
20  Return Large
```

9

`Frequent-Sets(`$D'$, $\bar{s}$`)`
(finds and computes support for all item sets with support $\geq \bar{s}$ in $D'$)

- Use any algorithm for discovering frequent sets.

*Figure 1*. Algorithm for discovering maximal associations.

sequence, and for each transaction $t$ we consider all item-sets $X$ M-supported by $t$ (there is one such item set for each category). For each such $X$ we increment the value of Hash($X$) by one. Clearly, by the end of the this pass, only those item sets $X$ for which Hash($X$) $\geq \hat{s}$ can possibly be M-frequent. Provided that the number of M-frequent sets in not too large (if this is not the case then $\hat{s}$ was set too low), the number of candidate M-frequent sets dramatically reduces after the first pass. In the second pass (lines 9–19), the database is again scanned in sequence. For each transaction $t$ and item-set $X$, M-supported by $t$, if Hash($X$) $< \hat{s}$ then $X$ is ignored, as it is clearly not M-frequent. Otherwise, it is a candidate M-frequent set, and its $s^{\max}(X)$ count is incremented by one. In addition, the appropriate portions of $t$ are added to the sub-databases $D(X, g')$. By the end of the second pass, those item sets with $s^{\max}(X) \geq \hat{s}$ are the M-frequent sets. We note that if the database is small enough, the first pass can be omitted.

## 4. Experimental results

We applied the maximal association rules tool to two text corpora.

### *4.1. The process*

Each text dataset was comprised of a large set of documents, which were further split into sentences. Each sentence was processed using an Information Extraction (IE) engine (ClearForest's ClearTags server), extracting the key items mentioned in the sentence. Thus, each sentence was converted into a transaction comprising of the set of key items mentioned in the sentence. These transactions formed the underlying database for which M-associations were calculated.

### *4.2. The DJ dataset*

The first data set, which we denote by DJ, consisted of 10,000 Dow Jones financial news briefs, all from September 23–24, 2001. The news briefs are in plain, natural language English text. Each article contains an average of 13 sentences and 436 words. The Information Extraction engine was configured to identify items of the following categories: companies and organizations, people, and products. After conversion into a transaction database, there were a total of 132,227 transactions, of which 99,674 were non-empty (i.e. the corresponding sentences contained at least one extracted term). The average number of items per non-empty transaction was 1.45. However, the text also included some sentences with a very large number of terms. There were 89 sentences containing 15 terms or more. Each individual term appeared on average in 11.6 sentences. Due to the low number of appearances of terms, and the low average number of terms per sentence, we used a relatively low minimum M-support threshold of $\hat{s} = 5$.

The total number of maximal association rules with M-confidence at least 50% was 1,386. If we limit ourselves to maximal rules with at least a 1.2 M-factor, the number of rules was 65.

*Table 1.* M-associations from the DJ dataset.

| | M-confidence (%) | Regular confidence (%) |
|---|---|---|
| Kim Beazley $\overset{\max}{\Longrightarrow}$ Labor Party<br>(Mr. Beazley was the leader of the Australian Labor party at the time) | 100 | 51 |
| Pharmacia Corp $\overset{\max}{\Longrightarrow}$ Ambien<br>(Ambien is Pharmacia's #2 selling drug) | 85 | 44 |
| Monsanto Corp. $\overset{\max}{\Longrightarrow}$ Fred Hassan<br>(Mr. Hassan is the CEO of the parent company of Monsanto) | 86 | 50 |
| Amazon.com $\overset{\max}{\Longrightarrow}$ Warren Jenson<br>(Mr. Jenson is the CFO of Amazon) | 55 | 37 |
| Qantas Airlines $\overset{\max}{\Longrightarrow}$ Australian Council of Trade Unions<br>(Qantas had a big dispute with trade unions in October 2001) | 54 | 34 |
| Isuzu, GM $\overset{\max}{\Longrightarrow}$ Richard Wagoner<br>(Mr. Wagoner is the CEO of GM, who have a partnership with Isuzu) | 86 | 64 |
| BASF AG $\overset{\max}{\Longrightarrow}$ Jurgen Strube<br>(Mr. Strube is Chairman of BASF) | 75 | 50 |
| Del Webb Corp. , Pulte Homes $\overset{\max}{\Longrightarrow}$ Mark O'Brien<br>(Mr. O'Brien is the President of Pulte which acquired Del Webb Corp.) | 100 | 75 |
| Windows $\overset{\max}{\Longrightarrow}$ Microsoft | 83 | 54 |

The M-association with highest M-confidence and M-factor was:

Federal Reserve $\overset{\max}{\Longrightarrow}$ Alan Greenspan (M-confidence $= 100\%$, M-factor $= 4.2$)

As a regular association, this rule has a mere confidence 23%, and would thus be under the 50% cut-off point. A sample of other M-associations discovered in the dataset is provided in Table 1.

A close examination of the sixty five M-associations with M-factor 1.2 or more shows that essentially all of them represent true associations in the real world. For example, thirty eight of these M-associations are of the type where one side of the rule is a company or an organization and the other side is a person. Of these, in 84% of the cases the person is an employee of the company. There is an interesting difference, however, between rules of the type "Company $\overset{\max}{\Longrightarrow}$ Person" and those of the type "Person $\overset{\max}{\Longrightarrow}$ Company". In rules of the type "Company $\overset{\max}{\Longrightarrow}$ Person" in 70% of the cases the person is an executive of the company (e.g., CEO, Chairman, CFO). In rules of the type "Person $\overset{\max}{\Longrightarrow}$ Company" only in 40% of the cases the person is an executive. This is consistent with the intuitive meaning of association rules: a company is chiefly associated with its executives, while any employee can be associated with the company. In rules of the type "Company $\overset{\max}{\Longrightarrow}$ Product" in 86% of the cases the product is a product of the company.

*Table 2*. M-associations from the TR dataset.

| | M-confidence (%) | Regular confidence (%) |
|---|---|---|
| Abu Bakar Baashir $\stackrel{\max}{\Longrightarrow}$ Bali bombing<br>(Abu Bakar is the #1 suspect for the 2002 Bali bombing) | 100 | 53 |
| Al Qaeda's $\stackrel{\max}{\Longrightarrow}$ Al Farooq Training Camp<br>(al Farooq was Al Qaeda main training camp) | 61 | 29 |
| United States, Malaysia $\stackrel{\max}{\Longrightarrow}$ Yazid Sufaat<br>(Yazid is an ex Malaysian army captain wanted by the US in connection to the world trade center terror attack) | 55 | 32 |
| Hamas, Hizballah $\stackrel{\max}{\Longrightarrow}$ Iran<br>(Hamas and Hizballah are anti-Israeli terrorist organizations supported by Iran) | 50 | 30 |
| Afghanistan, Egypt $\stackrel{\max}{\Longrightarrow}$ Mohammed Atef<br>(Atef, born in Egypt, is Al Qeada's second in command) | 50 | 25 |
| Riduan Isamuddin $\stackrel{\max}{\Longrightarrow}$ Indonesia, Malaysia<br>(Isamuddin, born in Indonesia and then exiled to Malaysia, is the head of a south east Asia Islamic terror organization suspected for the 2002 Bali bombing) | 50 | 33 |
| Al Qaeda, Abu Sayyaf $\stackrel{\max}{\Longrightarrow}$ Philippine<br>(Abu Sayyaf is a terrorist group, linked to Al Qaeda, operating in the Philippines) | 78 | 47 |

### *4.3. The TR dataset*

The second dataset, which we denote by TR, consisted of 1,970 news briefs on issues relating to terror and defence, from the period of one year starting September, 11, 2001. The news briefs are in plain, natural language English text. Each article contains an average of 120.5 sentences and 992 words. The Information Extraction engine was configured to identify items of the following categories: countries, companies and organizations, people, and facilities. After conversion into a transaction database, there were a total of 237,399 transactions, of which 167,116 were non-empty (i.e. the corresponding sentences contained at least one extracted term). The average number of terms per transaction was 1.32. Again, we set the minimum M-support threshold to $\hat{s} = 5$, and minimal M-confidence to $\hat{c} = 50\%$.

The total number of M-association was 1,109. If we limit ourselves to rules with at least a 1.2 M-factor, the number of M-associations is 128. A sample of some of these M-associations is provided in Table 2. Note that with a minimum confidence of 50%, all but the first of these rules would not have been obtained as a regular association rule.

### *4.4. Quantitative analysis*

Figure 2 shows the number of M-associations in the collections as a function of the minimum M-confidence, and for various M-factor values. As expected, the number of rules decreases for with the M-confidence threshold and the M-factor.

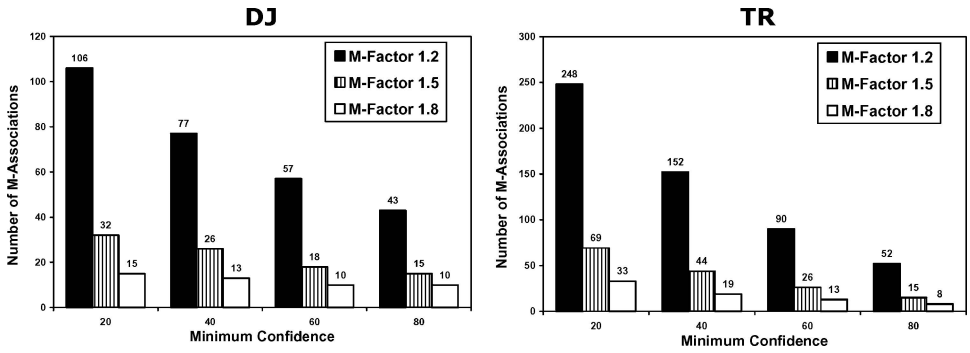


*Figure 2*. Number of M-Associations for the DJ and TR datasets. Minimum M-support is 5.

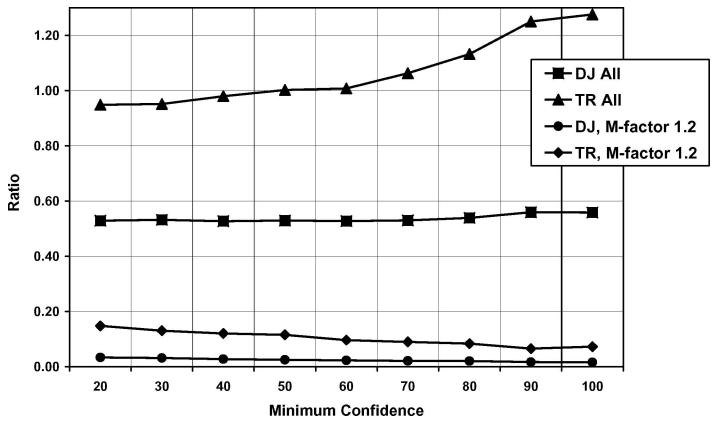


*Figure 3*. The ratio between the number of M-Associations to the number of regular associations, for a given confidence threshold. Minimum support (and M-support) is 5.

We also compared the number of M-associations to the number of regular associations, for various confidence thresholds. Since M-associations are always from one category to another, in the comparison we only considered regular associations where all elements of the left-hand-side are from one category, and all elements of the right-hand-side are from another category. Clearly, this can only reduce the number of regular associations. Figure 3 depicts the ratio between the number of M-associations to the number of regular associations. In general, the TR collection has more M-associations than DJ (see figure 2) and relatively less regular associations. Interestingly, for high confidence levels in TR there are more M-associations than regular associations. If we also consider the M-factor, the number of M-associations reduces dramatically. In the DJ collection, the number of M-associations with M-factor at least 1.2 is only 1–3% of the number of regular associations. In the TR collection the percentage is between 7 and 17%.

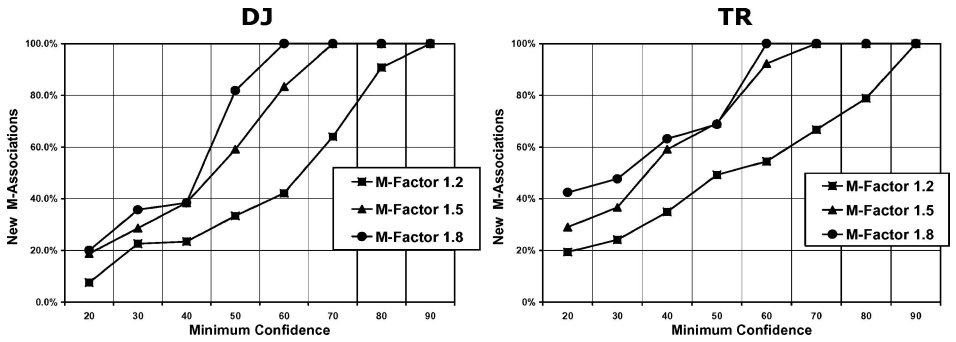


*Figure 4.* Percentage of M-Associations that would have been rejected as regular associations, for various minimum confidence thresholds and M-factors. Minimum support (and M-support) is 5.

Maximal associations are designed to capture associations not otherwise captured by regular associations, or to increase the confidence of rules. Figure 4 provides the percentage of M-associations that would have been rejected as regular associations, for a given confidence threshold. As can be seen, the percentage of M-associations that would not have been obtained as regular associations increases with the confidence threshold, and approaches 100% at a confidence threshold of 90%. This confirms our motivation that M-association capture associations that would otherwise be lost. The examples of the previous sections show that these associations include meaningful and important associations.

The computing time of regular and maximal association rules for the DJ set where 4:14 and 3:15 minutes, respectively, and for the TR set 10:05 and 3:21 minutes, respectively. The computations where performed using an Intel Pentium-4 2.4 GHz processor with 768 M memory. We note that only minimal optimization was performed, for both types of rules.

## References

Agrawal, R., Imielinski, T., and Swami, A.N. (1993). Mining Association Rules Between Sets of Items in Large Databases. In Buneman, Peter and Jajodia, Sushil (Eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, (pp. 207–216). Washington, D.C.

Ahonen, H., Heinonen, O., Klemettinen, M., and Verkamo, I. (1997). Applying Data Mining Techniques in Text Analysis. Technical Report C-1997-23, University of Helsinki.

Bench-Capon, T.J.M., Frans, Coenen, and Leng, P. (2000). An Experiment in Discovering Association Rules in the Legal Domain. In *Proceeding of the Workshop on Legal Information Systems and Applications (LISA)* (pp. 1056–1060).

Brijs, Tom, Swinnen, Gilbert, Vanhoof, Koen, and Wets, Geert. (1999). Using Association Rules for Product Assortment Decisions: A Case Study. In *Knowledge Discovery and Data Mining* (pp. 254–260).

Brin, Sergey, Motwani, Rajeev, and Silverstein, Craig. (1997). Beyond Market Baskets: Generalizing Association Rules to Correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 265–276).

Cai, C.H., Fu, Ada Wai-Chee, Cheng, C.H., and Kwong, W.W. (1998). Mining Association Rules with Weighted Items. In *International Database Engineering and Application Symposium* (pp. 68–77).

Dong, Jianning, Perrizo, William, Ding, Qin, and Zhou, Jingkai. (2000). The Application of Association Rule Mining to Remotely Sensed Data. In *SAC (1)* (pp. 340–345).

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Knowledge Discovery and Data Mining* (pp. 82–88).

Feldman, R., Aumann, Y., Amir, A., Zilberstein, A., and Klosgen, W. (1997). Maximal Association Rules: A New Tool for Mining for Keyword Co-Occurrences in Document Collections. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 167–170).

Feldman, R. and Dagan, I. (1995). KDT-Knowledge Discovery in Texts. In *Proceedings of the First International Conference on Knowledge Discovery (KDD)* (pp. 112–117).

Feldman, R., Dagan, I., and Hirsh, H. (1998). Mining Text Using Keyword Distributions. *Journal of Intelligent Information Systems*, 10(3), 281–300.

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., and Zamir, O. Text Mining at the Term Level. In *Principles of Data Mining and Knowledge Discovery* (pp. 65–73).

Feldman, R. and Hirsh, H. (1996). Mining Associations in Text in The Presence of Background Knowledge. In *Knowledge Discovery and Data Mining* (pp. 343–346).

Han, J. and Fu, Y. (1999). Mining Multiple-Level Association Rules in Large Databases. *Knowledge and Data Engineering*, 11(5), 798–804.

Hearst, M. (1999). Untangling Text Data Mining. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 3–10).

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining—A General Survey and Comparison. *SIGKDD Explorations*, 2(1), 58–64.

Lee, W., Stolfo, S.J., and Mok, K.W. (1999). A Data Mining Framework for Building Intrusion Detection Models. In *IEEE Symposium on Security and Privacy* (pp. 120–132).

Liu, B., Hsu, W., and Ma, Y. (1999). Mining Association Rules with Multiple Minimum Supports. In *Knowledge Discovery and Data Mining* (pp. 337–341).

Ma, Y., Liu, B., Wong, C.K., Yu, P.S., and Lee, S.M. (2000). Targeting the right students using data mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)* (pp. 457–464).

Mannila, H., Toivonen, H., and Verkamo, A.I. (1995). Discovering Frequent Episodes in Sequences. U.M. Fayyad and R. Uthurusamy (Eds.), *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada. AAAI Press.

Michail, Amir. (2000). Data Mining Library Reuse Patterns Using Generalized Association Rules. In *International Conference on Software Engineering* (pp. 167–176).

Rajman, M. and Besancon, R. (1997). Text Mining: Natural Language Techniques and Text Mining Applications. In *Proceedings of the seventh IFIP Working Conference on Database Semantics*.

Satou, K., Shibayama, G., Ono, T., Yamamura, Y., Furuichi, E., Kuhara, S., and Takagi, T. (1997). Finding Association Rules on Heterogeneous Genome Data. In *Proceedings of the Second Pacific Symposium on Biocomputing (PSB)* (pp. 397–408).

Srikant R. and Agrawal, R. (1997). Mining Generalized Association Rules. *Future Generation Computer Systems*, 13(2/3), 161–180,

Tao, F., Murtagh, F., and Farid, M. (2003). Weighted Association Rule Mining Using Weighted Support and Significance Framework. In *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (To appear).

Tung, A.K.H., Lu, H., Han, J., and Feng, L. (1999). Breaking the Barrier of Transactions: Mining Inter-Transaction Association Rules. In *Knowledge Discovery and Data Mining* (pp. 297–301).

Wang, W., Yang, J., and Yu, P.S. (2000). Efficient Mining of Weighted Association Rules (WAR). In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 270–274).

Wong, P.C., Whitney, Paul, and Thomas, Jim. (1999). Visualizing Association Rules for Text Mining. In *IEEE Symposium on Information Visualization (INFOVIS)* (pp. 120–123).