Maximal Association Rules: a New Tool for Mining for Keyword co-occurrences in Document Collections

Ronen Feldman, Yonatan Aumann,
Amihood Amir, Amir Zilberstein
Department of Mathematics and Computer Science Department
Bar-Ilan University
Ramat-Gan, ISRAEL
{feldman,aumann,amir,zilbers}@cs.biu.ac.il

Willi Kloesgen
GMD
German National Research Center
for Information Technology
D-53754 Sankt Augustin, Germany
kloesgen@gmd.de

Abstract

Knowledge Discovery in Databases (KDD) focuses on the computerized exploration of large amounts of data and on the discovery of interesting patterns within them. While most work on KDD has been concerned with structured databases, there has been little work on handling the huge amount of information that is available only in unstructured document collections. This paper describes a new method for computing co-occurrence frequencies of the various keywords labeling the documents. This method is based on computing maximal association rules. Regular associations are based on the notion of *frequent sets*: sets of attributes, which appear in many records. In analogy, maximal associations are based on the notion of *frequent maximal sets.* Conceptually, a frequent maximal set is a set of attributes, which appear alone, or maximally, in many records. For the definition of "maximality" we use an underlying taxonomy, *T*, of the attributes. This allows us to obtain the "interesting" correlations between attributes from different categories. Frequent maximal sets are useful for efficiently finding association rules that include negated attributes. We provide an experimental evaluation of our methodology on the Reuters-21578 document collection.

Introduction

In this paper we introduce a new data-mining tool which we term *maximal associations rules.* Like the associations rules introduced in (Agrawal et al 1993), *maximal associations rules* are rules of the form $X \Rightarrow Y$, were X and Y are sets of attributes. However, while the regular association rule $X \Rightarrow Y$ says that when one sees X one should also expect to see Y (with some *confidence*), the *maximal* association rule $X \Rightarrow Y$ says that when one sees X *alone* one should also expect to see Y *alone.*

As an example, consider the Reuters-21578 database. This database contains 21578 news articles categorized by country names, topics, people names, organizations and stock exchanges. Suppose that there are ten articles regarding "corn" that are annotated also by USA and Canada, and another twenty articles concerning "fish" and the countries USA, Canada and France. If we now search for (regular) associations with at least 50% confidence, we will only get the rules {USA, Canada}⇒{fish} with 66% confidence and the rule {USA, Canada, France}⇒{fish} with 100% confidence. The information regarding the strong connection between USA, Canada and "corn" is *lost.* In essence, we wish to capture the notion that whenever Canada and USA appear *alone* (without any other country), "corn" also appears. *Maximal Association rules* provide us with the necessary tool: as a *maximal* association, the rule {USA, Canada}⇒{corn} has 100% confidence. In this paper we formally define the notion of maximal associations and present an efficient algorithm for generating all such associations.

Given the algorithm for generating maximal associations, we continue to show how maximal associations may be used to obtain regular associations with negation, called *excluding associations.* An excluding association is a rule of the form $S_1 \cup \neg S_2 \Rightarrow S_3 \cup \neg S_4$ where S_1, S_2, S_3, S_4 are sets of items. The intuitive meaning of such an association rule is that whenever one sees the attributes of S_1 and *not* those of S_2 then one should also expect (with some confidence) to find those of S_3 and *not* those of S_4. For example:

{mining} ⇒ {coal}

is a regular association. However, adding negation we may find the excluding association:

{mining,¬coal} ⇒ {data}.

In general, there can be numerous excluding association, most of which are redundant and noninteresting. We show how to use the concept of maximal associations and frequent maximal sets to generate the "interesting" excluding associations.

Definitions

Let $\mathbb{A} = \{ A_1, \dots, A_n \}$ be a set of attributes with binary domain ({0,1}). A *row, r,* over $\mathbb{A}$ is a tuple $r = \{r[A_1], \dots, r[A_n]\}$, of 0's and 1's. Such a row can also be viewed as a set $\{A_i \mid r[A_i] = 1\}$ of the attributes from $\mathbb{A}$. A *relation, R,* over $\mathbb{A}$ is a multiset of rows over $\mathbb{A}$. A *taxonomy, T,* of $\mathbb{A}$ is

[1] Copyright © 1997, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

a collection of subsets $T=(T_1,...,T_k)$, $T_i \subseteq A$, which together cover A. Each T_i is called a *category*.

We first define regular association, as described by (Agrawal et al., 1993). For a given row, r, and set of items X, we say that r *supports* X if $X \subseteq r$. The *support* of X in a relation R, denoted by $s_R(X)$, is the number of rows $r \in R$ which support X. An *association rule* is a rule of the form $X \Rightarrow Y$, where X and Y are attribute sets. The *support* of the association is the support of $X \cup Y$, and the *confidence* of the association is $s_R(X \cup Y)/s_R(X)$.

Maximal association rules are defined in analogy. A *taxonomy pair* is a pair $X{:}L$, where X is a set of items and L is a category such that $X \subseteq L$. We call L the *boundary* of the pair. For row r and taxonomy pair $X{:}L$ we say that *$X{:}L$ is maximal in r*, if $r \cap L = X$. The *support* of $X{:}L$ in a relation R, denoted by $ms_R(X{:}L)$, is the number of rows $r \in R$ such that $X{:}L$ is maximal in r. Given a row r and taxonomy T, the set of all maximal taxonomy pairs of r is denoted $mp(r)$.

We can extend these notions to sets of taxonomy pairs. Let $V=\{X_1{:}L_1,\ldots,X_k{:}L_k\}$ be a set of taxonomy pairs. We call V a *taxonomy set*. For V as above we also write $V=X{:}L$, where $X=(X_1,...,X_k)$ is the sequence of item sets and $L=(L_1,...,L_k)$ are the boundaries. For row r and taxonomy set V, we say that *V is maximal in r*, if $X_i{:}L_i$ is maximal in r for all i. The *support* of V in a relation R, denoted by $ms_R(V)$, is the number of rows $r \in R$ such that V is maximal in r.

A *maximal association rule* is an expression of the form $V \Rightarrow W$, where V and W are maximal sets. The support of the association is $ms_R(V \cup W)$, and the confidence of the association is $ms_R(V \cup W)/ms_R(V)$.

We search for associations where the support is above some user-defined threshold, which we call the *minimum support* (denoted by σ), and whose confidence is above another user-defined threshold, which we call the *minimum confidence* (denoted by γ). An attribute set with at least the minimum support is a *frequent set.* A taxonomy set V is called a *maximal frequent set* if $ms_R(V) \geq \sigma$.

Example:
Consider the relation R consisting of the following rows:
{Canada, Iran, USA, crude, ship} x 2
{USA, earn}
{USA, jobs, cpi} x 2
{USA, earn, cpi}
{Canada, sugar, tea}
{Canada, USA, trade, acq} x 2
{Canada, USA, earn}

We will use a simple taxonomy that contains 2 categories: "countries" and "topics" (i.e., $T=$ {countries, topics}).

Given that $\sigma = 2$, and $\gamma \geq 0.5$, we have the following frequent maximal sets with respect to $T =$ {countries, topics}. The value of $ms(V)$ is written next to each frequent maximal set V.
{{USA}: "countries"} 4
{{Canada, Iran, USA}: "countries"} 2
{{Canada, USA} : "countries"} 3
{{crude, ship} : "topics"} 2
{{trade, acq} : "topics"} 2
{{Canada, Iran, USA} : "countries" ,{crude, ship} : "topics"} 2
{{USA} : "countries" ,{ cpi, jobs} : "topics"} 2
{{Canada, USA} : "countries" ,{acq, trade} : "topics"} 2

From the above frequent maximal sets we obtain the following maximal associations: (braces were eliminates to simplify the notation):
USA : "countries" ⇒ jobs,cpi : "topics" (2/0.5)
Canada, Iran, USA : "countries" ⇒ crude, ship : "topics" (2/1.0)
Canada, USA : "countries" ⇒ acq, trade : "topics" (2/0.66)

Specifying the boundaries in each expression may be cumbersome. In most cases, for an item set X, we are interested in *all* maximal taxonomy pairs in X. Thus, we adopt the following shorthand notation. Let X be a set of attributes. For taxonomy T, X_T will denote the taxonomy set $X_T=\{ (X \cap L){:}L \mid L \in T, X \cap L \neq \emptyset \}$. Thus, X_T is the set of all maximal taxonomy pairs in X. When clear from the context we shorthand X for X_T.

Computing Frequent Maximal Sets

We now show how to compute all frequent maximal sets in a relation R, given a taxonomy T and support threshold σ. The first step is to transform R into a new relation R', which lends itself to a speedy counting of support for taxonomy pairs. For each $r \in R$, we replace r by $r_T=\{(r \cap T_i){:}T_i \mid T_i \in T, X \cap T_i \neq \emptyset \}$. Let $R'=\{r_T \mid r \in R\}$. Note that the attributes of R' are taxonomy pairs. The following elementary claim establishes the necessary relation between R and R'.

Claim: Let $V=\{X_1{:}L_1,\ldots,X_k{:}L_k\}$, $W=\{Y_1{:}H_1,\ldots,Y_l{:}H_l\}$, be sets of taxonomy pairs. $V \Rightarrow W$ is a *maximal* association rule in R with support α and confidence β iff it is a *regular* association with the same support and confidence in R'.

Thus, it is sufficient to generate the associations is R'. This we do using any of the existing methods (e.g. [Agrawal and Srikant, 1994]).

Example:
T={"countries"={USA, Canada, Iran,....}, "North_America"={USA, Canada, Mexico}, "topics"={acq, jobs, cpi, crude, ship}}

R: {Canada, sugar, tea}, {USA, Canada, Iran, trade, acq},
{USA, Canada, France, earn}

Then R' is:
{{Canada}:"countries",{Canada}:"North_America",{sugar ,tea}:"topics"},
{{USA,Canada,Iran}:"countries",{USA,Canada}:"North_America",{trade,acq}:"topics"},
{{USA,Canada,France}:"countries",
{USA,Canada}:"North_America", {earn}: "topics"}
With support threshold 2, the only frequent maximal set is:
{USA,Canada}:"North_America"}

Excluding Associations

Let S={A_1,A_2,...,A_n} be a set of attributes. We denote the set {$\neg A_1$,$\neg A_2$,...,$\neg A_n$} by $\neg$S. Excluding associations are rules of the form $S_1 \cup \neg S_2 \Rightarrow S_3 \cup \neg S_4$ where S_1,S_2,S_3,S_4 are disjoint sets. The intuitive meaning of such a rule is that whenever one sees the attributes of S_1 and *not* those of S_2 then one should also expect (with some confidence) to find those of S_3 and *not* those of S_4.

In general, the number of excluding associations that exceed the support and confidence threshold is much higher than the number of regular associations (without negation). This is because each association without negation can give rise to many excluding associations with exactly the same support and confidence by simply adding the negation of irrelevant attributes.

These excluding association are certainly of no interest. In order to control the number of excluding associations, we introduce a redundancy filter that is designed to capture "interesting" excluding associations. We define a rule $S_1 \cup \neg S_2 \Rightarrow S_3 \cup \neg S_4$ to be *interesting* if it has the required support and confidence, but the corresponding association without the negations, $S_1 \Rightarrow S_3$, does not have the required confidence (it will always have at least the same support).

Even after introducing this redundancy filter, generating all interesting excluding association is a hard problem. Here we show that maximal association rules, introduced in this paper, give a partial solution to the problem of finding this problem. Using frequent maximal sets, we shall generate all the interesting excluding association with negations *within the taxonomy*. I.e., we generate the rules $S_1 \cup \neg S_2 \Rightarrow S_3 \cup \neg S_4$ where the attributes of S_2 (S_4) are in the same categories as those S_1 (res. S_3). The algorithm we present needs to generate all maximal sets, even those with support 1. To restrict the amount of sets generated we can use small taxonomies. Our experiments prove that even with very small taxonomies we obtain interesting results. The tradeoff between the size of the taxonomy and the number of interesting excluding associations is a matter for future research.

A maximal association rule is, by definition, an excluding association. In order to get only the interesting associations we look for maximal association rules $X{:}L \Rightarrow Y{:}L'$ where the corresponding regular association rule $set(X) \Rightarrow set(Y)$ does not hold (where $set(x)$ is the union of all items in x). Once we have such a rule, we must find negation parts, S_2 and S_4. In general there may be many possible S_2 and S_4. In this case we are interested the minimal S_2, and maximal S_4. However, note that S_4 can only decrease both the support and confidence of the rule. Thus, since we are interested in simple rules with high confidence, we shall drop S_4 altogether.

Consider the frequent maximal set $X{:}L$, with support α. Denote $S_1 = set(X)$. The maximal set $X{:}L$ corresponds to the negation set: $S_1 \cup \neg(set(L)-S_1)$. We seek the minimum cardinality set S_2, $S_2 \subseteq (set(L)-S_1)$, for which $S_1 \cup \neg S_2$ retains the support α. To this end, we consider all maximal sets of the form $Z{:}L$ where $X \subset Z$. We then compute the residue of each of these maximal sets with regard to X, i.e., we compute $set(Z)-set(X)$. Finally, S_2 obtained by computing the minimal set cover of all these residues. We want S_2 to contain at least one element from each of the residues. In that way S_2 covers all possible residues, and eliminates their support from the support of the current rule.

The set covering problem is NP-Complete and hence we must use a polynomial approximation. We will use a greedy covering algorithm to find the minimal cover of all residues. We will compute for each of the elements in the residues the number of residue in which they appear. We then select the element that appears in the highest number of residues and we delete all the residues covered by it. We then update the residue-count for each of the remaining elements and start the process again. The process will terminate when all residues were deleted. The elements that have been accumulated will be the members of the set S_2.

The Document Explorer system using the Reuters-21578 collection found around 30 excluding associations. For instance, the association rule {south_africa, usa}$\Rightarrow${acq}, when seen as a regular association, is supported by 9 documents and has 34% confidence. When we view {south_africa, usa}$\Rightarrow${acq} as a maximal association rule we still have the same support and the confidence of the rule increases to 50%. Using the algorithm in Figure 1 we found that the contents of S_2 is {argentina, kenya, uk, zaire

}. In other words the excluding association rule {south_africa, usa, ¬ argentina, ¬ kenya, ¬ uk, ¬ zaire }⇒{acq} has support 9 and confidence 50%.

Our algorithm finds the minimal S_2 such that the excluding association will have exactly the same support and confidence as the corresponding maximal associations. However, we may delete some of the elements of S_2 if their removal will only change the confidence by a small amount. Such excluding associations will be shorter and hence more meaningful. For example, starting from the excluding association of

{italy, usa, ¬ west_germany, ¬ japan, ¬ iran, ¬ france }⇒{acq}

which has support 10 and confidence 71%, ommiting japan, iran, and france from S_2, we obtain the excluding association

{italy, usa, ¬ west_germany}⇒{acq}

with support 10 and confidence 59% (recall that the association {italy, usa}⇒{acq} has support 10 and confidence 25%). Document Explorer uses a heuristic algorithm for finding such reductions.

Conclusions

We presented a new form of co-occurrence computation suitable for structured databases and in particular useful for document collections. This new form is based on the notion of maximal sets. Maximal sets are defined with respect to a taxonomy. Roughly speaking a frequent maximal set is a set of attributes that in many rows form the largest subset in given categories. The taxonomy is provided by the user. A more complex taxonomy will give rise to a larger number of frequent maximal sets and, as a result, a larger number of maximal association rules.

Frequent maximal sets are useful for a variety of tasks. Foremost, they provide means to capture inference rules otherwise lost using the regular associations. In particular, maximal association rules express the association between subsets of attributes appearing "alone". Maximal associations also help to reduce the number of generated associations and get only associations that are supported by the structure of the database or the document collection. In addition, frequent maximal sets are useful to infer interesting excluding associations, a task that would be very difficult without using frequent maximal sets. Finally, we can use these sets for dynamic browsing of document collections. Here, frequent maximal sets are used here to capture the exact relationship between entities in the collection.

Future work will include improving the usage of frequent maximal sets to find all possible excluding associations, and the generation of frequent maximal sets without the need of user-provided taxonomies. The taxonomy information will be directly inferred from the document collection.

Acknowledgements

This research was supported by grant # 8615 from the Israeli Ministry of Sciences, and by a GIF grant I-429-096.

References

Agrawal R., Mannila H., Srikant R., Toivonen H., and Verkamo I. 1996. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, Eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, pages 307-328, AAAI Press.

Agrawal R. and Srikant R. 1995. "Mining Sequential Patterns", *Proc. of the Int'l Conference on Data Engineering (ICDE)*, Taipei, Taiwan, March 1995.

Agrawal A., Imielinski T., and Swami A. 1993. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 207-216.

Amir A., Aumann Y., Feldman R., and Katz O. 1997. Trie Based Algorithms for Discovering Frequent Sets in Databases. Technical Report, Department of Computer Science, Bar-Ilan University, Israel.

Feldman R., and Hirsh H. 1997. Mining Associations in Text in the Presence of Background Knowledge. In Proceedings of the 2nd International Conference on Knowledge Discovery (KDD-96), Portland, Aug 1996.

Feldman R., Dagan I., and Kloesgen W. 1996. Efficient Algorithms for Mining and Manipulating Associations in Texts. In *Proceedings of EMCSR96*, Vienna, Austria, April 1996.

Feldman R., Amir A., Aumann Y., Zilberstein A., and Hirsh H. 1997. "Incremental Algorithms for Association Generation". In Proceedings of the 1st Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD97), Singapore, 1997.

Feldman R. and Dagan I. 1995. KDT - knowledge discovery in texts. In *Proceedings of the First International Conference on Knowledge Discovery (KDD-95)*, August 1995.

Lewis D. 1997. "The Reuters-21578, Distribution 1.0" http://www.research.att. com /~lewis/reuters21578

Mannila H. and Toivonen H. 1996. Multiple uses of frequent sets and condensed representations - Extended Abstract. In Proceedings of the 2nd International Conference on Knowledge Discovery & Data Mining (KDD-96), Portland, Aug 1996.