Knowledge Discovery in Textual Databases: A Concept-Association Mining Approach

Mutlu Mete [2], Nurcan Yuruk[2], Xiaowei Xu [1]*, and Daniel Berleant[1]

*Corresponding author

[1]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR 72204-1099, USA

[2]Department of Applied Science, University of Arkansas at Little Rock, Little Rock, AR 72204-1099, USA

Abstract

Concepts are often related to short sequences of words that occur frequently together across the text collections. Such concepts convey much of the meaning in any language. Association rule mining is a powerful technique for extracting relations among concepts. From a text mining perspective, association rules have mainly been used in a traditional support-confidence framework. This approach suffers from a tendency to generate a prohibitively large number of candidate rules. Recent works try to eliminate irrelevant rules from the result set. Recently also, the measures of *bond* and *all-confidence* were proposed instead of the more traditional support and confidence framework.

This chapter describes a new approach for mining associations among concepts from text collections. Our goal is to extract interesting associations among concepts from their co-occurrences within text collections. The proposed approach consists of three steps. We first extract the concepts themselves from the texts. Then we extract interesting associations among the concepts using association rule mining. Finally we construct a directed graph from the mined associations.

Introduction

The number of scientific publications is exploding as online digital libraries and the World Wide Web grow. MEDLINE, the premier bibliographic database of the National Library of Medicine (NLM), contains about 18 million records from more than 7,300 different publications dating from 1965; it is growing by about 400,000 citations each year. The explosive growth of information in textual documents creates great need for techniques for knowledge discovery from text collections.

Data mining, also known as knowledge discovery in databases, has been defined as the nontrivial extraction of implicit, previously unknown and potentially useful information from given data [1]. Different techniques including clustering, classification and association rule mining have been used to extract knowledge from text collections.

Association rule mining, first introduced by Agrawal et al. [2], can be summarized for our purposes here as follows. Suppose we have n tuples over a set A of attributes A_1, A_2, A_3, … A_n. Each tuple represents a text passage. Each attribute represents a word, and the attributes are Boolean in the sense that a value of true indicates that the word is present in the passage and a value of false indicates that the word is not. Thus each tuple represents the bag of words associated with the passage. Let I and J be two disjoint subsets of attributes in A. We say that $I \rightarrow J$ is an association rule if the following two conditions are satisfied.

- Support: At least a fraction s of the tuples contain attributes I or J.
- Confidence: Among the tuples in which I appears, at least a fraction c also have J appearing in them.

The goal is to identify all valid association rules for a given relation. An attribute set is called a frequent item set if its attributes are in enough tuples. Frequent item sets form the basis of association rule mining. Exploiting the monotonicity property of frequent item sets (each subset of a frequent item set is as frequent or more), and using data structures that support counting, the set of all frequent item sets can be efficiently determined even for large databases. Different algorithms have been developed for that task, e.g. Agrawal and Srikant (1994 [3]). Please see [4] for a review of association rule mining.

Traditionally, association rule mining algorithms use the support-confidence framework to find interesting association rules in the following two steps.

- All itemsets that have support above the user specified minimum are generated. These itemsets are called the *large itemsets*.
- For each large itemset, all the rules that have a minimum confidence are generated as follows: for a large itemset X and any $Y \subset X$, if support(X)/support(X-Y) $\geq$ *minimum_confidence*, then X-Y $\rightarrow$ Y is a valid rule.

The support-confidence framework cannot find association rules between rare items, i.e. items that do not satisfy the minimum support condition.

Recent works [5] [6] [7] [8] deal with finding rules based on other metrics besides support and confidence. In [6], the authors mine association rules that identify correlations and consider both the absence and presence of items as a basis for generating the rules. In [7], the authors use support as part of their measure of interest of an association. However, when rules are generated, instead of using confidence, the authors use a metric they call conviction, which is a measure of implication and not just co-occurrence. In [8], the authors present an approach addressing the rare item problem. In [5], the authors also look at alternative measures of interest, named the gini index, entropy gain, and chi-squared.

Most recently, Omiecinski [9] describe all-confidence and bond as interestingness measures for association rules. These new measures are not based on support and can find dependence between itemsets. Another advantage of these new measures is that they satisfy downward closure as support does. Therefore, there exist efficient algorithms to find association rules that satisfy the measures.

Although originally association rule mining was proposed for mining consumer purchasing patterns in retail stores, applications extend far beyond this specific setting. For example, Morishita and Sese [10] applied association rule mining for genome mining. Association rule mining techniques are also used for classification [11] and clustering tasks [12].

The work most relevant work to this chapter is association rule mining based knowledge discovery in textual databases. Feldman et al [13][14][15] used an approach to association rule mining techniques for knowledge discovery in text collections based on statistical analysis for discovering associations among individual keywords assigned to texts. There is no description about how keywords were assigned to texts, suggesting that the assignment may be performed manually. Lin [16] et al. used a similar technique. The main difference was that Lin et al. used key terms automatically extracted from text collections. Recently Loh et al. [17] proposed a concept based approach to text data mining. Their approach combines an automatic categorization step with a data mining step. Categorization identifies concepts presented inside texts. Data mining then discovers patterns by analyzing and relating concept distributions in the collection. Another classification step is then needed to create concept definitions.

In this chapter we use an *n-gram* based approach to extract concepts from textual databases. Our approach does not require a specific domain, unlike Weeber et al. [18], who mapped sentences to predefined UMLS concepts.

Once concepts are extracted, we mine for associations in text collections. This is motivated by shortcomings of previous work which takes bags of words as input to the association rule mining algorithms (e.g. [3]), and finds associations among single isolated words. There are two pitfalls in that approach. One is that some concepts consist of multiple words. These multiple word concepts, such as *lung*

cancer, cannot be found as a unit in the association rules. The other is that the number of associations is overwhelmingly large. This means that it is difficult to find interesting rules from such a large number of associations.

Attempts to mine rules using only single words appear to be rooted in the fact that they can introduce significant ambiguity, since it is the context within which word patterns appear that identifies the real meaning. We show examples in Table 1. For example, if a searcher is looking for information about lung cancer, the concept would be *lung cancer*. Instead of only single isolated words we permit multi-word concepts as input to the associate rule mining algorithm. Therefore, our approach is able to find associations among such concepts. This also reduces the number of associations. For example, if we use multi-word concepts as input we might find the rules "smoking → lung cancer" and "lung cancer → smoking." Using the single isolated words "lung" and "cancer," however, we might find the rules "smoking → lung," "smoking → cancer," "lung → cancer," "lung → smoking," "cancer → smoking," "cancer → lung," "smoking, lung → cancer," "smoking, cancer → lung," "cancer, lung → smoking," "cancer, smoking → lung," "lung, smoking → cancer," and "lung, cancer → smoking." In the second case there are twelve rules while with multi-word concept extraction only two rules would be found. In general, an isolated word based approach will generate many more redundant rules than a concept based approach.

Table 1. Isolated Words vs. Concepts

Isolated Words	Concepts
New	New York
Lung	Lung cancer
Data	Data mining

Graph Representation

To build on the concept associations that are extracted, a graph representation of them is useful because it shows indirect associations: if A and B are associated, and B and C are associated, then A and C are indirectly associated. This is not always obvious from a long list of associations, but is easily seen when associations are represented visually in a graph. These transitive associations can lead to new knowledge. For example, Hearst shows in [19], when investigating causes of migraine headaches, that the following associations can be found:

- stress is associated with migraines
- stress can lead to loss of magnesium
- calcium channel blockers prevent some migraines
- magnesium is a natural calcium channel blocker
- spreading cortical depression (SCD) is implicated in some migraines

- high levels of magnesium inhibit SCD
- migraine patients have high platelet aggregability
- magnesium can suppress platelet aggregability

These transitive associations suggest that magnesium deficiency may play a role in some kinds of migraine headache.

Method

An experimental evaluation on real textual datasets was conducted. We compared our concept based approach with the isolated word based approach. Different interestingness measures were compared. Finally we show how the results can be used to generate a directed concept association graph.

Concept Based Association Rule Mining Approach

Our concept based text mining approach consists of three parts: concept extraction, concept association mining and concept association graph generation. They are described in the next sections respectively.

Concept Extraction

In general, it is well-known that the input heavily determines the quality of the outputs; *garbage in, garbage out* is a famous computer axiom meaning that if invalid data is entered into a system, the resulting output will also be invalid. Particularly in association rule mining, more accurate input produces more interesting rules that may lead to discovery of unknown associations. Instead of only single isolated words, concepts that are closely related multi-word groups that have semantic coherence seem more beneficial to use as inputs to association rules. The concept extraction approach we use here is different from usual concept extraction studies that focus on digging up the most representative words of documents. In such studies (e.g. [20]), finding common *themes* in a given document is the main objective and these common themes or patterns are considered as concepts. In this study we define and use concept in a different manner. Here, a concept is a single word or group of consecutive words that occurs frequently enough in the entire document collection. Each concept candidate was expected to satisfy a predefined support threshold equivalent to 10 occurrences, equal to the threshold used to prune the least frequent words during preprocessing.

We further elaborate our method for concept extraction as follows. It is necessary to preprocess the datasets before extracting the concepts. Only letters, digits, '/', and '-' were kept in words. Other preprocessing steps included stop word

elimination, stemming, and pruning the least and most frequent words. Since the most frequent words can be an important part of a multi-word concept, and therefore removing them may lead to missing some meaningful rules, they are not pruned based on an upper threshold. Additional preprocessing performs stemming, which can affect concept extraction significantly. In general, stemming is finding the base form of the word to improve concept analysis. For instance, the stemmer converts both *addressing* and *addressed* words to the same stem term *address*. The Krovetz Stemmer [21] was used for this purpose. After cleaning the raw dataset, it remains with only stemmed words that are frequent enough. We used an *n*-gram based approach for mining concepts. This approach needs two parameters *max_ngram_size* and *min_sup*. The former indicates the maximum number of words in a concept, and the latter is a frequency threshold for concepts.

One of the crucial steps that affect results is specification of key parameters. Assume that a dataset has a number of 4-gram concepts (i.e., concepts that are four words long). If *max_ngram_size* is set to 2, these concepts will be divided into two parts. Thus using a small *max_ngram_size* tends to both lose actual concepts, and also unnecessarily increase the number of concepts. It is also easy to see that the number of concepts can exceed the number of words in the original pre-processed dataset. Table 2 shows example textual data. The algorithm for extracting concepts consists of two steps. In the first step, candidate concepts are found, and then counted in order to check the support of each of them. The higher order *n*-grams are generated first in order not to split words apart from their neighbour. For each *n*, where *n* is the number of words in the concept, the algorithm passes over the dataset once. The result is a candidate concept table that will be used in the next step. In the second step, the concept candidates are pruned based on the threshold *min_sup*. All the candidates with less support than that are eliminated from the candidate table. The concepts in Table 3 show the IDs and supports for concepts extracted from the dataset shown in Table 2. Note that in Table 3 concept {B} does not appear because all of occurrences of B except in D5 are included in the 2-gram concepts {A B} and {C B} as shown in Table 2. Since the remaining B in D5 is not frequent enough to satisfy *min_sup*, it does not appear in Table 3. Also, since each occurrence of word E is contained in {C D E} or {E F}, there is no concept {E} in Table 3.

Table 2. Sample dataset with six documents and six words

Documents	Words
D_1	A B C D E F A
D_2	E F C B
D_3	A B E F
D_4	A C B F
D_5	C D C D E B A B
D_6	D A C D E C

Table 3. Concepts extracted using dataset from Table 2 with max_ngram_size = 3 and min_sup = 2

Concept_ID	Concept	Support
1	{C D E}	3
2	{A B}	3
3	{C B}	2
4	{E F}	2
5	{A}	3
6	{C}	2
7	{D}	2
8	{F}	2

After concept extraction each document is represented as a *bag of concepts*. Table 4 shows the *bag of concepts* representation for the original dataset in Table 2. We will use *bags of concepts* instead of bags of words as input for our next step of mining concept associations.

Table 4. Dataset from Table 2 with extracted concepts

Documents	Items
D_1	2 1 8 5
D_2	4 3
D_3	2 4
D_4	5 3 8
D_5	6 7 1 2
D_6	7 5 1 6

Mining Concept Associations

After concept extraction, each document is represented by a set of concepts. It is not the concepts themselves, but the associations between them that represent the knowledge buried in the documents that it is our goal to extract. One major difference between our approach and earlier methods is that we use multi-word concepts instead of single isolated words as items for association rule mining.

Because we are interested in the dependence between concepts in text, we decided to use all-confidence and bond as our measure of interestingness for association rules. We also compared these two measures with use of support and confidence measures for mining textual datasets to reveal differences in the two approaches. Our experimental evaluation will show that the rules generated using bond and all-confidence are more interesting and easy to examine than those generated using support and confidence.

Generating a Directed Graph of Concept Associations

The approach used to construct a directed concept association graph is not complicated. We need one pass over the rules generated previously. The method of generating a directed graph needs a confidence parameter *min_conf* to decide the type of line and the direction. The algorithm reads a rule and its confidence value. Each side of the association is created as a new node if it is not yet represented. If the confidence value is at least *min_conf*, the edge between two nodes is drawn as a solid line but otherwise a dashed line is used to show the weakness of the relation. It is clear that some rules that are paired (e.g. A → B and B → A) are read twice. In this case, the rule with the greater confidence value determines the direction of the edge. If both are equal, the edge will have two arrowheads. Also, if both confidence values of the pair of rules are smaller than *min_conf*, the edge between them will be a bidirectional dashed line. The directed graph of the dataset from Table 2 is shown in Figure 1 with *min_conf* = 0.6. It is obvious that directed graphs are more informative than undirected graphs. For example, the confidence values of 3 → 4 and 4 → 3 are both 0.5, smaller than *min_conf*, 0.6. Therefore, in Figure 2 the edge between {C B} and {E F} is a bidirectional dashed line. On the other hand, the edge that connects {C} and {D} is a bidirectional solid line because their confidence values are equal and above *min_conf*.

Note how directed graphs highlight transitive rules that are otherwise obscured in the result set. X → Y is a transitive rule if X implies Z, Z implies Y, and X does not imply Y directly. For example, it is too time consuming to figure out that the association between {C} and {A} can be interesting because {C} implies {C D E} and {C D E} implies {A} but {C} does not imply {A} directly. More concretely, based on the above discussions, directed graphs depict both direct associations between concepts as well as transitively justified indirect rules.

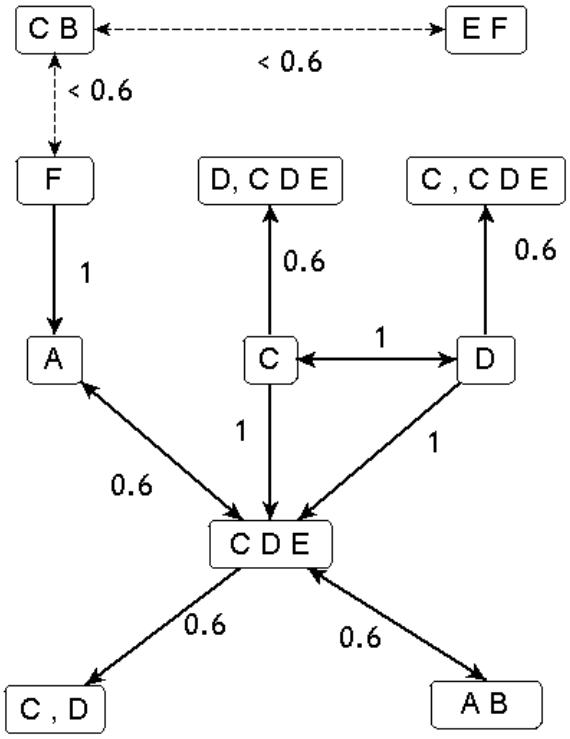

Fig. 1. Directed concept association graph based on dataset from Table 2.

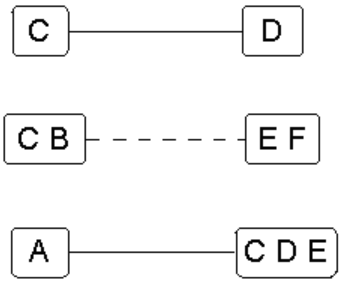

Fig. 2. Three sample of undirected concept association graph.

Experiments and Results

To demonstrate robustness and coherence of the approach, we chose two textual datasets. The first dataset, which we call PubMed-abstracts, consists of 9795 biomedical abstracts downloaded from PubMed Central [22] with the keyword *mRNA*. For the second dataset, we did experiments with the BioMed Central text corpus [23] that includes 4581 published articles of peer-reviewed biomedical research.

To show the effectiveness and efficiency of our algorithm, we look at the results in three different ways. Our first focus is to demonstrate how results from concept based association rule mining using multi-word phrases outperform the results from using only single words. Second, we must determine how to rate the *interestingness* of association rules that are extracted. For this purpose, we compare traditional support and confidence properties with new interestingness measures, *bond* and *all-confidence*, introduced in [9]. Finally, we evaluate our method for constructing concept association graphs - directed graphs for interesting concept associations based on the all-confidence property. A graph representation is employed to help users interpret the results and infer new rules.

Isolated words vs. multi-word concepts

In this section, we highlight the concept extraction technique that we have developed to generate accurate input to association rule mining algorithms. First, recall that in the case of single words, the number of association rules that is generated is considerably higher than when multi-word concepts are used. This large number of rules increases the human effort required to interpret them and identify the interesting ones. Our rule mining software is a modified version of the fast apriori implementation by Bodon [24] and was run on a Sun machine with 4GB of physical memory.

Regarding PubMed-abstracts, using the traditional support and confidence interestingness measures, with a support threshold of 0.005 and a confidence threshold of 0.7, multi-word concept based rule mining produced only four rules. However, using isolated words caused the program to crash without yielding any rules. It started with 2105 frequent 1-itemsets and consumed all memory after generating 1,541,371 4-itemsets.

We tuned the support and confidence thresholds in order to decrease the number of frequent itemsets so that the process running the program can fit into memory. We set support to 0.007 and confidence to 0.5. The isolated word based approach returned 1,993,922 rules with the largest rule containing 11 words. The multi-word concept based approach produced only 5 rules, all of them with three words. All five seemed interesting. Certainly, five are manageable to interpret compared to 1,993,922 rules. From the isolated word based approach, 63,816 rules had a confidence value of 1.0, which was unexpected. A confidence value of 1.0

describes the strongest associations which therefore might be expected to be interesting. However, one can see that they are not actually meaningful and interesting by looking at the top 20 results shown in Table 5. Table 6 illustrates rules from the multi-word concept based approach. It is obvious that the concept based approach does not have redundant rules, while the word based approach does. For instance, the rules "transcription-polymerase → chain", "transcription-polymerase → chain, reaction" and "transcription-polymerase → reaction" in Table 5 are redundant.

Table 5. Top Association Rules from isolated words (support = 0.007, confidence = 0.5)

Rules	Support
method, conclusion, level, cell, expression → result	488
method, conclusion, effect, cell→result	469
rt-pcr, method, conclusion, cell→result	433
method, conclusion, decrease→result	406
reverse, chain, method, expression → reaction	390
reverse, chain, method, conclusion → reaction	381
objective, method, significant → result	381
objective, method, conclusion, significant → result	378
transcription-polymerase → chain	375
transcription-polymerase → chain, reaction	375
transcription-polymerase → reaction	375
background, method conclusion, study → result	374
objective, method, level → result	373
objective, method, conclusion, level → result	368
china, objective → result	361
method, conclusion, induce, cell → result	354
method, conclusion, level, protein, cell → result	346
method, conclusion, show, cell → result	345
line, human, study, mrna → cell	344
objective, method, conclusion, increase → result	340

Table 6. Association rules from concepts (support = 0.007, confidence = 0.5)

Rules	Support
vascular endothelial growth factor → vegf	124
cardiac → heart	89
background, method → result	88
induce apoptosis → apoptosi	78
apoptotic → apoptosi	71

New Metrics vs. the Traditional Support & Confidence

Table 7 is the summary of the results from the PubMed-abstracts dataset. They were acquired from multi-word concept based rule mining using the traditional support and confidence interestingness measure, the all-confidence measure, and

the bond measure, respectively. It shows the total number of association rules for each interestingness measure.

Table 7. Number of association rules for support and confidence, all-confidence and bond

Threshold	Support (0.002) &Conf.	All Conf.	Bond
0.5	752	529	232
0.7	143	62	38
0.8	67	26	15
0.9	21	8	6
1.0	0	2	2

Initially, for both the PubMed-abstracts and BioMed Central datasets, *max_ngram_size* was defined as 4, so concept phrases up to four words long were sought, and *min_sup* was defined as 10, meaning the concept phrase appeared in at least ten documents. In the traditional model, for high support thresholds, such as 0.5 and 0.2, the software returned no rules. In order to make this model comparable with the new measures, we decreased the support threshold to 0.002, thus reducing the pruning effect of a high support as much as possible. The 0.002 value was determined by memory limitations because in case of an even lower threshold, 0.001, the memory needed for candidate generation greatly exceeded the available memory.

As noted in [9], there is no certain relationship between the traditional and the new measures. From Table 7, it can be observed that the set of association rules is larger for the traditional measures but does not necessarily include all of the rules yielded by using bond or all-confidence. In the following we show some example rules from results obtained using the all-confidence measure with threshold value 0.5. Note that none of these rules are listed in the result set obtained using traditional interestingness measures (support = 0.002, confidence = 0.5). These example show that support pruning may cause some interesting rules to be lost in the pruning process. After each rule, a descriptive note with its source is attached to confirm the association.

- skin disease → psoriasis

… a chronic (long-lasting) *skin disease* of scaling and inflammation that affects 2 to 2.6 percent of the United States population, or between 5.8 and 7.5 million people. Although the disease occurs in all age groups, it primarily affects adults. It appears about equally in males and females. *Psoriasis* occurs when skin cells quickly rise from their origin below the surface of the skin and pile up on the surface before they have a chance to mature [25].

- plasmodium → malaria

Four species of *Plasmodium* infect humans and cause *malaria* [26].

- spinal ligament → ectopic bone formation

… of the posterior longitudinal *ligament of the spine* (OPLL) is a common

form of human myelopathy caused by a compression of the spinal cord by *ectopic ossification* of spinal ligaments [27].

- vein wall → thrombu

Deep *vein thrombosis* (DVT), a form of venous thromboembolic disease, refers to the formation of a *thrombus* (blood clot) within a deep *vein*, commonly in the thigh or calf. Although venous thromboembolic disease can develop after any major surgery, people who have orthopaedic surgery on the lower extremities are especially vulnerable. Three factors contribute to formation of clots in veins:
Stasis, or stagnant blood flow through veins. This increases the contact time between blood and *vein wall* irregularities [28]...

When we compare the bond and all-confidence measures, the bond measure is more restrictive than all-confidence. With the same threshold value it returns fewer rules. Its result set is a subset of the results yielded by all-confidence using the same threshold. Nevertheless, by changing the threshold it is possible to obtain approximately the same set of rules. For example, experiments with the PubMed-abstracts dataset show that results from bond = 0.7 were the same as results using all-confidence = 0.8. Therefore, all-confidence and bond are equivalent in being able to find the most interesting rules. We show some rules generated by using bond and all-confidence for the BioMed Central dataset in Table 8.

Table 8. Number of association rules extracted from BioMed Central

Threshold		All Conf.	Bond
0.5	MF: 10	2968	626
	MF: 20	3141	285
0.7	MF: 10	226	120
	MF: 20	72	41
0.9	MF: 10	45	36
	MF: 20	20	18
1.0	MF: 10	32	23
	MF: 20	17	15

MF Minimum frequency threshold for concepts in dataset

Directed Graphs

Depending on the formal definition of bond and all-confidence, rules that are returned by these measures contain no information about direction of the association. After mining a sample dataset, Table 9 lists rules whose all-confidences are at least 0.5. Based on the aforementioned association rules, three examples of undirected graphs between some concepts are shown in Figure 2. Although graph representations of concept associations assist users to see relations quickly and easily, undirected graphs are less informative that one might like. Therefore, since each rule has a confidence value, we integrated the graph representation with the confidence value to give directions to edges.

Table 9. List of rules satisfying all-confidence=0.5 after concept mining of dataset from Table 2.

Rule	Confidence
8 → 3	0.5
3 → 8	0.5
8 → 5	1.0
5 → 8	0.6
7 → 6	1.0
6 → 7	1.0
7 → 6, 1	1.0
6 → 7, 1	1.0
1 → 7, 6	0.6
7 → 1	1.0
1 → 7	0.6
6 → 1	1.0
1 → 6	0.6
4 → 3	0.5
3 → 4	0.5
5 → 1	0.6
1 → 5	0.6
2 → 1	0.6
1 → 2	0.6
8 → 3	0.5
3 → 8	0.5
8 → 5	1.0
5 → 8	0.6

We further apply the traditional confidence metric to find the direction of each rule. Figures 3-6 are sample graphs that exhibit some interesting associations. The first two are constructed from the PubMed-abstracts dataset, and the last two are from the BioMed Central dataset. All the associations shown in the figures have all-confidence values greater than 0.5. Figure 5 shows interesting associations between *chlorthalidone* and the others. The dashed lines between concepts indicate weak confidence values (<0.6). The descriptive notes below explain the relationships between concepts in Figure 5 and Figure 6 respectively.

- ALLHAT (Antihypertensive and Lipid-Lowering Treatment to Prevent *Heart Attack Trial*) was the largest antihypertensive trial and the second largest lipid-lowering trial and included large numbers of patients over age 65, women, African-Americans, and patients with diabetes, treated largely in community practice settings…This trial is comparing treatment of hypertension with a diuretic (*chlorthalidone*) against newer types of antihypertensives - an alpha-adrenoceptor blocker (*doxazosin*), an ACE inhibitor, and a calcium antagonist - in a high-risk patient group (all over 55 years with one or more cardiovascular disease (CVD) risk factors) [29].
- Osteoporosis is a significant public health problem associated with increased mortality and morbidity. Our aim in this cross-sectional study was to investigate the relationship between lifetime physical activity and *calcium*

intake and bone mineral density (BMD) and BMC (bone mineral content) in 42 regularly menstruating Caucasian women (age 21.26+/-1.91 years, BMI 23.83+/-5.85). BMD and BMC at the lumbar spine (L2-L4), hip (femoral neck, trochanter, total), and total body were assessed by dual energy x-ray *absorptiometry* (DXA). [30]

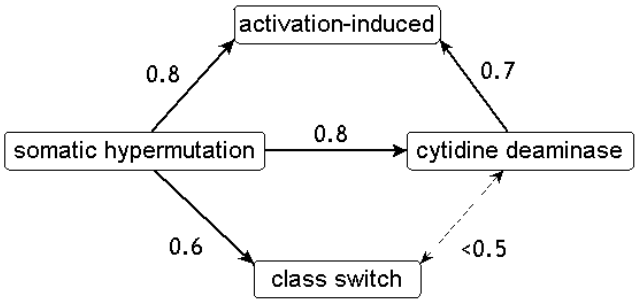

Fig. 3. Graph representation of a set of transitive association rules from PubMed (all-confidence = 0.5)

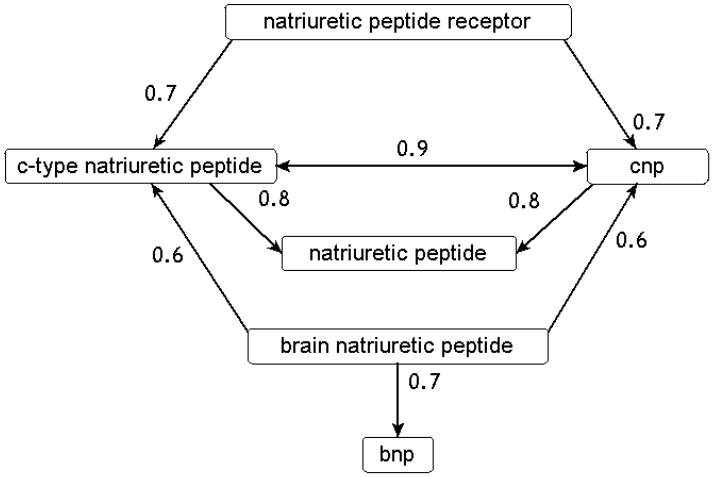

Fig. 4. Graph representation of a set of transitive association rules from PubMed (all-confidence = 0.5).

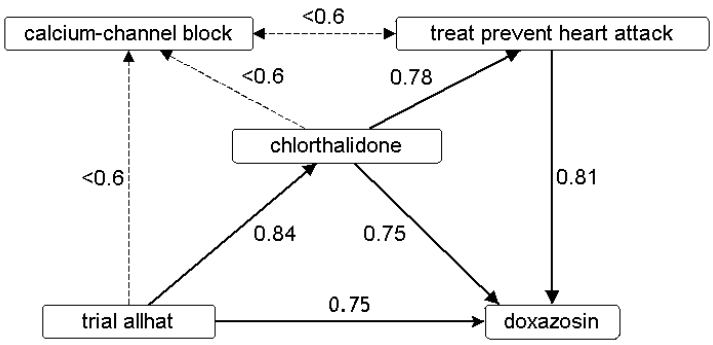

Fig. 5. Graph representation of a set of transitive association rules from BioMed Central with all-confidence = 0.6

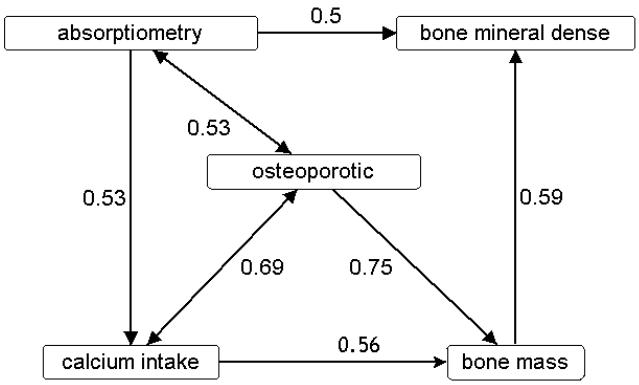

Fig. 6. Graph representation of a set of transitive association rules from BioMed Central with all-confidence = 0.5

Conclusions

In this chapter we explore a new approach for knowledge discovery from biomedical text databases. A concept is a keyword or multi-word phrase that describes the subject about which a user is seeking information. For example, if a searcher is looking for information about lung cancer, the concept would be *lung cancer*. Our goal is to extract interesting associations among concepts that co-occur within a text collection.

After concept extraction, the documents are each represented as a *bag of concepts*. It is not the concepts, however, but the associations between them that represent much of the knowledge buried in the documents. Our text data mining

task is to dig out these buried associations as nuggets from text collections. One contribution of this chapter is the use of concepts as inputs for the associate rule mining algorithm. Another contribution is a graph representation of mined associations.

We evaluated our techniques by using two real textual datasets. The evaluation results show our system can automatically find interesting concepts and their associations from unstructured text data. The experimental results also show that our approach can significantly reduce the number of uninteresting associations. Considering the results from the real-world dataset we used, we conclude that the all-confidence measure is quite useful in generating interesting associations. Furthermore, the generated directed graph that is generated for concept associations not only shows the directed associations represented by association rules but also all transitive associations. The concept association graph can be used to infer new association rules. Therefore it can lead to the discovery of new knowledge from textual databases.

In the future, we will investigate other concept extraction algorithms for knowledge discovery from textual databases. We will also develop more efficient algorithms for generating concept association graphs.

Examples

As additional examples, we mine Reuters-21578 Text Collection [31]. Some interesting relations are visualized in Figure 7-8. Former one depicts associations between a mine company, a union and others. Latter shows interesting associations between *energy information administrate eia* and the others. The dashed line between *oil demand* and *gasoline demand* indicates weak confidence value. The concept *energy information administrate eia* contains also abbreviation *eia* for energy information administrate. The abbreviation *spr* stands for Strategic Petroleum Reserve in energy related news.

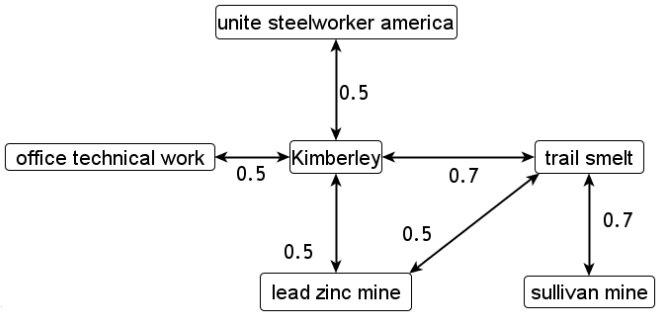

Fig. 7. Graph representation of a set of transitive association rules from Reuters-21578 Text Collection (all-confidence = 0.5)

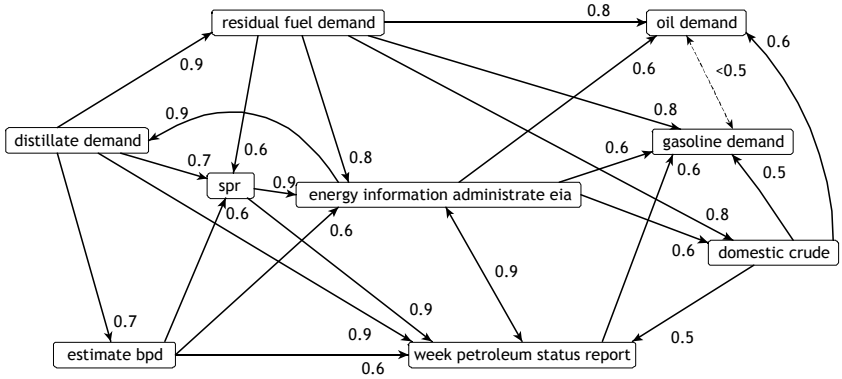

Fig. 8. Graph representation of a sample set of transitive association rules from Reuters-21578 Text Collection (all-confidence = 0.5)

Chapter Questions

1- What is support and confidence of an association rule?
2- Formulate bond and all-confidence measures.
3- Discuss relations between all-confidence, bond, and support for a dataset.
4- What is a concept and how it is related to a *n*-gram?

References

1. Fayyad U, Piatetsky-Shapiro G, and Smyth P (1996) Knowledge Discovery and Data Mining: Towards a Unifying Framework, Second Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, pp. 82-88
2. Agrawal R, Imielinski T, and Swami A (1993) Mining Association Rules Between Sets of Items in Large Database. ACM SIGMOD Conference, pp. 207-216
3. Agrawal R and Srikant R (1994) Fast Algorithms for Mining Association Rules. 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, pp 487-499
4. Hipp J, Guntzer U, Nakhaeizadeh G (2000) Algorithms for Association Rule Minin – A General Survey and Comparison, ACM SIGKDD Explorations, Vol.2, pp. 58-64
5. Manning C and Schütze H (1999) Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA.

6. Brin S, Motwani R, Silverstein C (1997) Beyond Market Basket: Generalizing Association Rules to Correlations. *ACM SIGMOD Conference*, Tucson, Arizona, pp. 265-276
7. Brin S, Motwani R, Ullman J, Tsur, S (1997) Dynamic Itemset Counting and Implication Rules for Market Basket Data *ACM SIGMOD Conference*, Tucson, Arizona, pp. 255-264
8. Liu B, Hsu W, Ma YM (1997) Mining Association Rules with Multiple Minimum Supports. *The Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 337-341
9. Omiecinski E (2003) Alternative Interest Measures for Mining Associations. *IEEE Trans. Knowledge and Data Engineering*, 15(1), pp. 57-69
10. Morishita S, Sese J (2000) Traversing Itemset Lattices with Statistical Metric Pruning. In *Proc. of the 19th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*,. ACM Press, pp. 226-236
11. Liu B, Hsu W, and Ma YM (1998) Integrating Classification and Association Rule Mining. *The Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, New York City, pp. 80-86
12. Beil F, Ester M, and Xu X (2002) Frequent Term-Based Text Clustering. 8th *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, pp 436-442
13. Feldman R, and Dagan I (1995) Knowledge Discovery in Textual Databases (KDT). *First international conference on knowledge discovery (KDD'95)*, Montreal, pp 112-117
14. Feldman R, and Hirsh H (1996) Mining Associations In Text In The Presence Of Background Knowledge. *2nd International Conference on Knowledge Discovery and Data Mining*, pp. 343-346
15. Feldman R, Dagan I, Hirsh H (1998) Mining Text Using Keyword Distributions. *Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, 10(3), pp. 281-300
16. Lin SH, Shih CS, Chen MC (1998) Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach. *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp 241-249
17. Loh S, Wives LK, Oliveia JPM (2000) Concept Based Knowledge Discovery from Texts Extracted from the Web. *ACM SIGKDD Explorations*, vol. 2, pp. 29-40
18. Weeber M, Vos R, Klein H, de Jong-van den Berg LTW (2001) Using Concepts In Literature-Based Discovery: Simulating Swanson'S Raynaud Fish Oil And Migraine Magnesium Discoveries. Journal of American Society for Information Science and Technology; 52 (7), pp.548–557
19. Hearst MA (1999) Untangling Text Data Mining. *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland, pp 3-10

20. Lewis D (1992) An Evaluation Of Phrasal And Clustered Representations On A Text Categorization Problem. *ACM-SIGIR Conference on Information Retrieval*, Copenhagen, Denmark, pp 37-50
21. Krovetz R (1993) Viewing Morphology as an Inference Process. *16th ACM SIGIR Conference*, Pittsburgh, pp 191-202
22. PubMed Central, http://www.pubmedcentral.nih.gov/
23. BioMed Central text corpus, http://www.biomedcentral.com/info/about/datamining/
24. Bodon F (2003) A Fast APRIORI Implementation. *IEEE ICDM Workshop on Frequent Itemset Mining Implementations*, Melbourne, Florida, pp 56-65
25. Health Information Main Page, http://www.niams.nih.gov/hi/topics/psoriasis/psoriafs.htm
26. College of Biological Sciences, http://www.biosci.ohio-state.edu/~parasite/plasmodium.html
27. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Abstract&list_uids=9662402
28. Hospital for Special Surgery, Orthopedic Surgery, http://orthopaedics.hss.edu/services/conditions/hip/dv_thrombosis.asp
29. Information For Health Professionals http://allhat.sph.uth.tmc.edu
30. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12150501&dopt=Abstract
31. Reuters-21578 Text Categorization Text Collection http://www.daviddlewis.com/resources/testcollections/reuters21578/