Generating Association Graphs of Non-Co-occurring Text Objects using Transitive Methods

Niranjan Jayadevaprakash
Purdue University School Of Science, Indianapolis
723 W. Michigan Street SL 280
Indianapolis, IN 46202.
+1 (317) 274-9727
njayadev@cs.iupui.edu

Snehasis Mukhopadhyay
Purdue University School Of Science, Indianapolis
723 W. Michigan Street SL 280
Indianapolis, IN 46202
+1 (317) 274-9727
smukhopa@cs.iupui.edu

Mathew Palakal
Purdue University School Of Science, Indianapolis
723 W. Michigan Street SL 280
Indianapolis, IN 46202
+1 (317) 274-9727
mpalakal@cs.iupui.edu

ABSTRACT

In this paper we discuss text data mining (TDM) mainly in the context of the biomedical domain, where we extract associations from MEDLINE text articles and construct association graphs. We explore two techniques, the co-occurrence method and transitive method. We propose a novel transitive method of finding associations that does not rely on meta-data, and compare the results with another known transitive method that uses meta-data in text, to find a link/relationship between objects of interest. Co-occurrence of these terms (objects) is not required in the transitive methods to find out that they are associated. The results show that our proposed new method is as accurate as the known method that uses meta-data. This, in turn, implies that relationships can be discovered even when meta-data is not available or incomplete. A case study of a transitive association between a pair of genes (BRCA1—STAT1) is also carried out to illustrate the effective hypothesis generating ability of our method. Based on the results, we conclude that our method can be used effectively for association extraction and also for hypothesis generation, which can later be validated through biological experimental analysis.

Keywords

Bio-Informatics, Text Mining, Data Mining, Association rules, metadata, knowledge discovery

1. INTRODUCTION

Discovering non-trivial trends and patterns or associations amongst objects of interest (which may be any entity) in varied or single domain from huge collections of data (semi) automatically can help to improve decision-making speed and accuracy [3], and is a challenging current problem. The act of finding associations or relationships amongst various entities of interest is called a form of knowledge extraction or data mining. The data could be well structured as in a database [5], or semi-structured as in a web page with meta-data, or it could be totally unstructured as in a simple text document. The extraction of knowledge from unstructured or semi-structured text categorizes as Text Data Mining (TDM). A substantial amount of research work has been done towards text mining in recent years [4].

One simple way to capture associations from a database of text documents would be to simply read through the relevant documents and construct the associations manually. This would be really impractical in the case of a large volume of information especially in a domain like biomedical research. Reading through this many documents to extract knowledge would obviously be impractical in terms of time and effort, and would be prone to errors. Associations among biological objects such as genes, proteins, diseases, etc. are one such form of knowledge, which would be of great interest to biomedical researchers. To facilitate such automatic knowledge extraction various association discovery algorithms have been proposed [13] [12].

An association is formally defined as a bi-directional implication between two objects in the form of A->B, where A and B are the two objects of interest and A is an antecedent and B is a consequent, or vice-versa. The definition for association rules as applied to mining structured data is given in [12]. These objects are words/phrases in the context of text mining. In the biomedical domain the objects could be connected/associated by words/phrases like 'binds', 'activates', 'function as' and also names of other biological objects. In this paper we do not discuss the nature of the association nor do we discuss the directionality of the relationship.

Associations can be discovered either from structured data as those stored in a database [15] or from unstructured data such as free text documents. In this paper we derive associations from unstructured or semi-structured text data. Machine learning or statistical techniques can be applied in performing such text data

SAC'05, March 13-17, 2005, Santa Fe, New Mexico, USA.
Copyright 2005 ACM 1-58113-964-0/05/0003…$5.00.

mining. [7] discusses a machine learning approach and deals with the biomedical domain.

In this paper we discuss a very effective statistical method to discover associations, known and novel, in unstructured text data. We use the tf-idf model to find these associations mainly in the biomedical domain using the MEDLINE database of documents. The objects of interest in this case are gene terms and we intend to extract already known associations as well as to generate hypothetical associations, which could help in scientific discoveries. A set of genes is used to query terms to retrieve documents from MEDLINE. A statistical model for association discovery is then used and a weight is attached to every discovered association and an association graph is built.

We use the principle of co-occurrence to find these associations at first and later explore transitive methods [8] in which co-occurrence (which is a restriction applied to the former method) is not a necessity. A comparison is performed between our new transitive method not using meta-data, and a known transitive method [1][2], which uses metadata present with the text documents in the MEDLINE database. The results presented illustrate how the new method can make novel association discoveries. This was evaluated by utilizing some already established associations and then using a set of documents dated before the earliest date of documents in which the object pair co-occurs. This is shown for the BRCA1 and STAT1 genes.

The paper is structured as follows: In section 2 we discuss co-occurrence method (non-transitive method) for mining associations using the tf-idf model. In section 3 we discuss the disadvantages of the co-occurrence method and discuss two transitive methods not requiring co-occurrence (one new method using unstructured text without meta-data and another known method using meta-data). We also include the results of the comparison between the two transitive methods. Section 4 discusses the conclusions and some directions of future work.

2. CO-OCCURRENCE BASED STATISTICAL METHOD FOR NON-TRANSITIVE ASSOCIATION EXTRACTION

We use a statistical method to extract basic non-transitive associations from unstructured abstracts from MEDLINE.

2.1 Model

The statistical model we use here is the tf-idf vector space model [9] (term frequency * inverse document frequency model), which is identified as one of the more effective text mining models compared to other models such as the Log Level Likelihood ([17] pages 172-174) and the Odds ratio models. The tf-idf model uses the concept of relevance and co-occurrence of terms. The relevance of a term ‘j’ w.r.t. a document ‘i’ is as

$$w_{ij} = t_{ij} * \lg(\frac{N}{N_j})$$

w_{ij} = relevance of term ‘j’ in document ‘i’, t_{ij} = term frequency of term ‘j’ in document ‘i’, N_j = document frequency for term j, N = total number of documents.

A particular term is more relevant w.r.t. a document if it appears more frequently in the document and appears in fewer numbers of documents in the document set.

An association weight is attached with every association between a pair of terms [6]. This is given by A_{jk}

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg(\frac{N}{N_j}) * t_{ik} * \lg(\frac{N}{N_k})$$

2.2 Experimental setup

The biomedical domain was chosen for the experiments. The text documents were from the MEDLINE (PubMed) database of biomedical articles. The query terms were a list of genes. The goal was to determine associations between all pairs of query terms (genes) using the tf-idf method. The associated weight with each pair of terms indicates the strength of the relationship between the terms. The strength of the association between two objects is directly proportional to the association weights between them.

The documents were downloaded based on the list of gene terms. The abstracts in each of the documents were extracted and the tf-idf method was applied to the set of extracted abstracts. We obtain a set of associations from which we construct an underlying non-transitive association graph. The initial set of data used involved a set of gene terms, which belonged to the same series. The relationships between these terms were known in advance and were well established.

We obtained a set of associations showing strong relationships between the gene terms. This meant that these gene terms co-occurred in many of the documents in the data set. Since these terms co-occurred, they must be strongly related as they were mentioned.

3. TRANSITIVE METHODS FOR ASSOCIATION EXTRACTION

Transitivity is defined in the context of relations. A relation R is transitive *if* R(x,y) *and* R(y,z) ➔ R(x,z). For example, if nutrient A deficiency affects physiological process B and B causes disease C, then nutrient A deficiency and disease C are associated. This example illustrates one level of transitivity (indirection), similarly, we can have multiple levels of indirection which links two objects of interest (A and C in the above example).

The non-transitive method previously discussed in section 2 is shown to be fairly effective, but it has an obvious limitation. It relies heavily on co-occurrence of a pair of terms in the same documents to establish a relationship amongst them. Two terms may not co-occur in any of the documents but they may yet be related. This relationship may have been discovered at a recent time or the association may be novel.

To overcome the above-mentioned shortcomings, we propose to use transitive methods to mine associations. These methods basically relate two terms of interest present in separate documents through some commonly occurring properties/concepts in documents containing the terms of interest.

We discuss 2 methods that involve transitive associations amongst terms of interest. The first method is a known one [1][2] using meta-data (the Medical Subject Heading or MeSH terms) present in documents (we call this method the MeSH method) and the other is a newly proposed one based on unstructured text without containing and using any meta-data terms (we call this method the Unstructured Text method).

3.1 Meta-data based transitive mining (the MeSH method)

In this method, meta-data present in documents is utilized to find transitive relations between query terms. The meta-data are representative of the key concepts associated with the document, e.g., keywords in technical articles. The MeSH (medical subject headings) meta-data of MEDLINE articles are used to test the effectiveness of this approach [1].

The MeSH headings in document contain the profile of the document, which is basically a set of concepts related to the main subject of the document like genes, diseases, etc. The MeSH headings are assigned from a standard list by trained indexers. The basic principle of working in MeSH method is as follows. If there are certain concepts common to the documents containing two objects of interest (in the abstract), then the objects are related. The weight associated is directly proportional to the number of concepts common to the documents containing the two objects of interest. The base model used to assign the strength of the association is a variation of the tf-idf model.

The procedure is as follows:

1. The relevant documents (along with their MeSH data), which have the objects of interest (gene terms) is downloaded from MEDLINE.
2. The unique set of MeSH terms is extracted from the set of downloaded documents.
3. A score is calculated for all pairs of terms (objects of interest) and MeSH terms using the scoring function

   $$W_{xy} = \text{Score}_{TFIDF}(C_x, M_y) = a * \log(\frac{N}{Q})$$

   C_x is the query term, M_y is the metadata term (one out of the entire MeSH term set in the document set), a = number of documents having C_x and M_y, N = total number of documents, Q = a + number of documents having M_y, but not C_x

4. The direct association strength between query terms which have non-zero values (common MeSH terms) is then calculated using the function **Association (C_i , C_j)** =

$$\sum_{r=1}^{p}(w_{ir} * w_{jr}) / \sqrt{\sum_{r=1}^{p} w_{ir}^2 * \sum_{r=1}^{p} w_{ir}^2}$$

where p is the total number of unique MeSH terms collected from the document set.

The data set used was again the set of gene terms described in section 2.2 in which about 90 facts are known (i.e., the existence of associations and non-associations). The results are shown in figure 1. The overall recall was 89% and further a large number of novel associations were also discovered. This is a fairly accurate result.

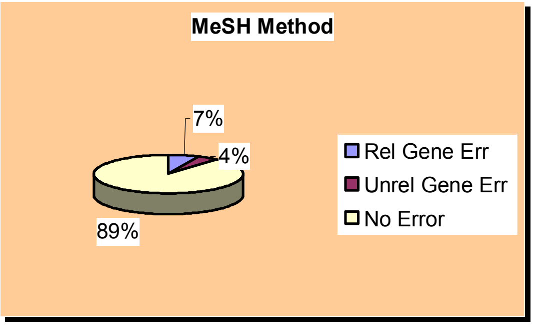

Figure 1:Accuracy for MeSH method

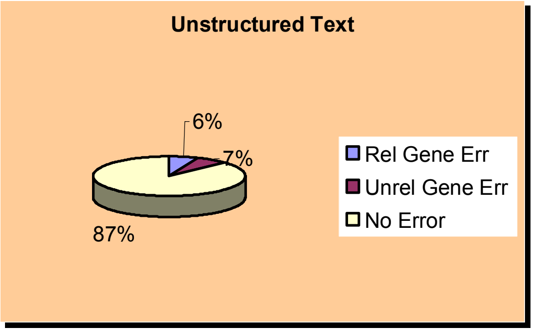

Figure 2: Accuracy for transitive unstructured text method

3.2 Transitive Method on Unstructured Text without metadata (the Unstructured Text Method)

In the method discussed in the previous section, there are several disadvantages. The method relies on metadata, which is assigned manually to every document by expert human indexers. This takes a lot of time and effort and can be prone to error. Besides, a large number of documents in the pubmed database are not assigned MeSH terms. Another disadvantage is that the MeSH vocabulary is subject to revision quite often. The previous method cannot identify associations even among terms that co-occur in the same document, if there are no MeSH terms. [8] also discusses a transitive closure method (called TransMiner) for finding novel associations. The difference between our proposed method and the method used in [8] is that the TransMiner method uses only the query terms as intermediate terms for transitivity whereas our method generates relevant terms (vocabulary discovery as explained below) from the document set and uses them as connecting concepts between the objects of interest (query terms). Using a fixed set of terms to find a transitive connection is a disadvantage because it is not always necessary that the terms co-occur.

To achieve transitivity, we perform a vocabulary discovery on the document set. This step identifies other important and relevant terms (not listed in the initial query) in the entire set of documents. Vocabulary discovery can be viewed as an automated method for generating “meta-data” for the unstructured text. The discovered set of words substitutes for the MeSH data in the previous method. The vocabulary discovery method used works as follows (based on [16]):

1. Collect all unique words in the set, eliminating common words (like etc, and, is) using a stop-word list.
2. The generated words are cleaned by applying word stemming (morphing) techniques. The WORDNET morphing library [14] was used for this purpose
3. Apply the tf-idf method $w_{ij} = t_{ij} * \lg(\frac{N}{N_j})$ (the terms are as explained in section 2.1) using the unique words from step 1. The output is a matrix of documents against the unique words.
4. Rank the words in decreasing order of their weights w.r.t the documents.
5. Extract the tokens that are ranked between 1 to R in at least D documents based on the rank and distribution proportion selected by the user.
6. A small value of R ensures selection of highly weighted tokens, and a relatively large value of D ensures that the same token is highly weighted in significant proportion of the training documents.

The result of the vocabulary discovery is an extended thesaurus (list of terms) which not only contain the original query terms, but also additional significant terms generated in the manner described above through vocabulary discovery.

The tf-idf method as explained in section 2 is then applied and associations are found between all pairs of terms. The various types of associations obtained would be:

1. Associations between initial set of query terms.
2. Associations between one initial query term and a newly discovered term.
3. Associations between two newly discovered query terms.

A graph, G, is then constructed from the associations. This graph is an undirected graph with the objects in the thesaurus as nodes and the edges indicate an association between them. The graph can be a disconnected graph. The advantage of having a graph structure is to facilitate the application of the transitive closure algorithm. We apply a transitive closure on the graph G containing the set of initial and discovered terms combined, to obtain all associations between the initial set of terms, either directly or through one or more discovered or original terms. The transitive closure algorithm used was the Floyd-Warshall’s algorithm for transitive closure. This has a time complexity of $O(n^3)$. The output of the transitive closure of G is another graph G^* containing G and closed under transitivity. When associations between initial set of words through discovered terms is found, the minimum association weight in the chain of associations is chosen to be the association strength between the initial set of words. This idea here is that the strength of association between the mail terms is as strong as the “weakest link” (the minimum weight association in the chain of associations) about 13% (figure 2), which compares well with the MeSH method used in section 3.1. given that it does not depend on metadata, which may not be available or which may keep changing and appended often. The associations/non-associations amongst genes in a row in the table is known and the associations amongst two genes in different rows are not well known.

Query Gene	Rel. Gene1	Rel. Gene2	Unrel. Gene1	Unrel. Gene2
IFNGR1	IFNGR2	STAT1	POU1F1	ABCC6
MAPK-8IP1	NUEROD1	GCGR	ABCA4	CNGB1
MSH2	BRCA1	PIK3CA	BLMH	ADRB2

Figure 3: Sample Query Genes

Another interesting comparison was to see the number of associations each of the methods (MeSH and Unstructured) returned (including known and novel ones), how many of them were common to both, and the number of associations that were exclusive to each method. The results are as shown in figure 4. Most of the known associations were discovered by both methods. Amongst the unknown associations from the original dataset, about 84.3% (541 out of 645) of the associations returned exclusively by the transitive unstructured text method were found to be potential associations from manual evaluation and nearly 80.18% (about 180 out of 225) of the associations returned exclusively by the MeSH method were found to be potential associations. This suggests that our method has a better recall than the MeSH method. This verification was performed by manually reading through the abstracts of MEDLINE articles, which contained terms of the associations. Some of the discovered associations are discussed in the next section.

3.3 The BRCA1—STAT1 association

In this section we pick the gene pair BRCA1 and STAT1 to illustrate the findings of our method. The association between BRCA1 and STAT1 gene is evident as they co-occur in recent research reports like [10][11]. This association was appropriately discovered as a direct association by both, MeSH method (section 3.1) and our method (section 3.2), when applied on a document set (from MEDLINE) which was up to date. To show the ability of our approach to find new associations through transitivity, we considered only those documents, which exclusively had either STAT1 or BRCA1 but not both. The first mention of both genes in the same document was only as recently as the year 2000, though the STAT1 gene was first mentioned in articles which were published in the year 1994 and BRCA1 in the year 1993. Therefore we restricted our document set to those that were published between the years 1st Jan 1900 to 31st Dec 1999. As expected, our approach showed a one level transitive association between STAT1 and BRCA1, through the words like “tumor suppressor”, “cytokine”, “IFN” i.e.

BRCA1--- tumor suppressor ---STAT1.
BRCA1--- cytokine---STAT1

The MeSH method failed to detect this association, due to the lack of a non-overlapping set of MeSH data. This is an obvious disadvantage of the MeSH method as not all documents have MeSH data attached to them. Another interesting observation was made when the document set considered for the above associations was considered only until the year of 1995, which was only slightly greater than the years during which the articles containing STAT1 and BRCA1 were published. The association between STAT1 and BRCA1 was discovered by our approach (section 3.2) through a two level transitivity through the words "tumor suppressor" and "cytokine"

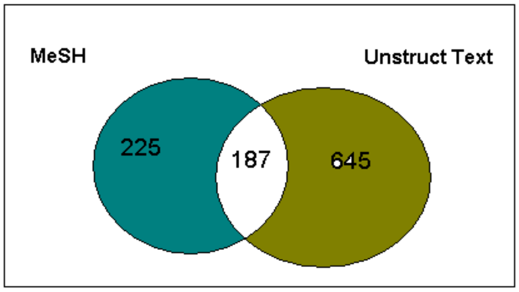

Figure 4: Number of associations returned by MeSH and Unstructured text method

BRCA1--- tumor suppressor--- cytokine---STAT1

This was also observed for many other associations like BRCA1 and MSH2. From these observations we can say that our method can be used in early discovery of associations, and the transitive associations are more than one level when there is very little information/data available. The MeSH method has the disadvantage that there can be only one level of transitivity and the objects can be connected only through MeSH terms.

4. CONCLUSION AND FUTURE WORK

To conclude, the non-transitive method for association extraction relies on the co-occurrence of terms in documents. To overcome this drawback, we apply transitive methods on text documents to mine established and novel associations. We have realized from the experiments that the transitive unstructured text mining method without using meta-data is nearly as effective as the MeSH method which uses meta-data containing related concepts to the subject of the document. The meta-data is manually assigned by trained human experts and is not assigned to all documents in the database. Compared to this the transitive non-metadata based method has a distinct advantage. The "meta-data" is generated automatically, using a vocabulary generator. This saves effort and also eliminates errors. Also having meta-data to find associations is not necessary and we can mine for multilevel transitive associations against the single level transitive association of the MeSH method.

As future work it will be interesting to see how we can combine the results from both MeSH and the unstructured text method to get more accurate results than what would be possible with each individual method. It would be also interesting to map the generated vocabulary in the non-metadata method to translate to the MeSH terms and use the generated MeSH terms to find transitive associations. . We would also like to test our method on different domains to prove that our method is domain independent.

5. REFERENCES

[1] Sehgal, A., Qiu, X. Y., and Srinivasan, P. Mining MEDLINE Metadata to Explore Genes and their Connections. *Proceedings of the SIGIR 2003 Workshop on Text* Analysis and Search for Bioinformatics. July 2003.

[2] Srinivasan Srinivasan, P and Sehgal, A.K Mining MEDLINE for Similar Genes and Similar Drugs, *Techincal Report, Department of Computer Science, The University of Iowa, July 2003, TR# 03-02*

[3] Hearst M. Untangling text data mining. *Proceedings of the 37th ACL Conference, 1999.*

[4] Ronen Feldman and Ido Dagan, Mining Text Using Keyword Distributions, *Journal of Intelligent Information Systems 10, 281–300 (1998)*

[5] Fayyad U.M. and Uthurusamy R. Data mining and knowledge discovery in databases *(Introduction to the special section). CACM, 39(11): 24-26, 1996.*

[6] Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, *Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342*

[7] Rindflesch T.C. Weinstein, J.N, Tanabe .L, Hunter L. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature, Nature Genetics, 2000 March, 24(3):227-234

[8] V. Narayanasamy, S. Mukhopadhyay, M. Palakal and D. Potter. TransMiner: Mining Transitive Associations among Biological Objects from Text. Accepted (to appear) Journal of Biomedical Sciences, 2004.

[9] G. Salton. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983

[10] Welcsh PL, Lee MK, Gonzalez-Hernandez RM, Black DJ, Mahadevappa M, Swisher EM, Warrington JA, King MC. BRCA1 transcriptionally regulates genes involved in breast tumorigenesis. Proc Natl Acad Sci U S A. 2002 May 28;99(11):7560-5.

[11] Ouchi T, Lee SW, Ouchi M, Aaronson SA, Horvath CM. *Collaboration of signal transducer and activator of transcription 1 (STAT1) and BRCA1 in differential regulation of IFN-gamma target genes.* Proc Natl Acad Sci U S A. 2000 May 9;97(10):5208-13.

[12] J. Hipp, U. Guntzer, and G. Nakaeizadeh. *Algorithms for Association Rule Mining - A General Survey and Comparison.* In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000

[13] R. Agarwal and R. Srikant. *Fast algorithms for mining association rules*. In Proc. Of the 20th Int'l Conf. On Very Large Databases (VLDB '94), Santiago, Chile, June 1994.

[14] Christianne Fellbaum, et al.. *WordNet: An Electronic Lexical Database,* May 1998 ISBN 0-262-06197-X MIT Press).

[15] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. *Mining association rules between sets of items in large databases.* In Proc. of the ACM SIGMOD Conference on Management of Data, pages 207--216, Washington, D.C., May 1993.

[16] Y. Fu, J. Mostafa, and K. Seki. *Protein Association Discovery in Biomedical Literature*. ACM/IEEE Joint Conference on Digital Libraries (JCDL 2003), Houston, TX, May 2003.

[17] C. D. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. *The MIT Press, Cambridge, Massachusetts, 1999.*