REVEALING TOPIC-BASED RELATIONSHIP AMONG DOCUMENTS USING ASSOCIATION RULE MINING

Kritsada Sriphaew and Thanaruk Theeramunkong
Information Technology Program
Sirindhorn International Institute of Technology
131 Moo 5, Tiwanont Rd.
Bangkadi, Muang, Pathumthani, Thailand
email: kong@siit.tu.ac.th, thanaruk@siit.tu.ac.th

ABSTRACT

With a large volume of electronic documents, finding documents the contents of which are same or similar in their topics has recently become a crucial aspect in textual data mining. Towards revealing so-called topic-based relationship among the documents, this paper proposes a method to exploit co-occurring unigrams and bigrams among documents to extract a set of topically similar documents with association rule mining techniques. To evaluate effectiveness of the method, a collection of well-organized scientific research publications is employed. The experimental result indicates that any two documents with referential links can be found with the accuracy of 60-80% in the case of unigrams, and 80-90% in the case of bigrams. An analysis of discovered association rules is also given.

KEY WORDS

association rule mining, data mining, knowledge discovery, document, topic-based, relationship.

1 Introduction

Recently the volume of electronic textual content has grown dramatically and continuously in the form of numerous digital libraries, web pages and publicized text databases. They are varied in their formats ranging from absolutely unstructured documents (e.g. plain text), semi-structured documents (e.g. HTML) to fully-structured documents (e.g. XML). Due to an immense quantity of these electronic documents, it is costly and often unrealistic for users to examine such textual data in detail, resulting in overlooking a great deal of useful information. To utilize the textual data efficiently and effectively, three main approaches are information retrieval (IR), text categorization (TC) and text mining (TM).

By indexing terms of documents, a keyword-based IR system efficiently finds a set of documents from a collection, related to input keywords. Given one or two keywords, the system may return thousands of probably relevant documents. Ranking these documents according to only one or two keywords, may not be so meaningful. In addition, browsing along a long list of those documents is not practical. As another approach, TC assigns a class to a document given a set of predefined classes and a training set of labeled documents. In the past, a variety of classification models were developed in different schemes, including k-NN-based classification, Bayesian approaches, support vector machines and so on. Although the categorization result makes us possible to find some needed documents via its assigned category, managing into a set of classes is quite rough and may not be enough to efficiently support the way to access documents related to each individual document. As a more recent approach, some TM tasks including information extraction, text summarization and automatic link generation, have come into prominence with the forthcoming high-performance computing environment. Among these tasks, automatic link generation is a task to find relationship among documents, a kind of information source for knowledge management, knowledge visualization, and knowledge retrieval. The basic problem is to link a document with its similar documents, given a collection of documents.

In the past, many works [1],[2] utilized link information to find the similarity between documents. This method can be applied when there exist links among documents, i.e., hypertexts. Moreover, some works applied a clustering algorithm on link information to group similar documents [1],[3]. In contrast with the link-based methods, there are still relatively few works [2],[4] based on a so-called term-based approach that applies common terms shared among documents to find similar documents. Assume that there is no existing link information, the approach can adopted for both hypertext and non-hypertext documents. Since it exploits terms for finding similarity, similar documents found are mostly topically related, in contrast with logical or structural relationship. Towards revealing topic-based relationship among the documents, this paper proposes a method to exploit co-occurring unigrams and bigrams among documents to extract a set of topically similar documents. The method applies the FP-tree algorithm which is a well-known technique to mine association rules efficiently. The effectiveness of the method is evaluated using a collection of well-organized scientific research publications. The rest of the paper is organized as follows. Sec-

Document	w_1	w_2	w_3	...	w_m
d_1	1	0	1	...	0
d_2	0	1	1	...	0
d_3	1	0	0	...	1
...	...	...	...	...	...
d_n	1	1	1	...	0

Figure 1. The original view of textual database

tion 2 formulates association rule mining for textual data. In section 3, the definition of topic-based relationship and document representation are described. The implementation detail is shown in section 4. In section 5, the evaluation method is described. Section 6 illustrates experimental results and their analysis. Finally, a conclusion is given in section 7.

2 Association Rule Mining for Textual Data

A document can be viewed as a set of words (terms), and a set of documents can be united to form a textual database where each row represents one document. Figure 1 shows the original view of a textual database. Assume that i, j are integers, d_i is the i^{th} document and w_j is the j^{th} word (term). This database has n documents and m distinct word terms. The values in the column of each row represent the existing and non-existing of particular words appearing in that document. Any documents can be transformed to a relational table.

The original concept of association rule mining is to find all subsets of items (called itemsets) that frequently occur in the database, and then to extract the rules indicating how a subset of items influences the presence of another subset. As done in several works [5],[6], it is possible to mine a set of words that frequently co-occur. However, since we focus to mine the rules that associate a set of documents containing the topic-based relationship, the textual database has to be pivoted by swapping rows and columns as shown in Figure 2. That is, documents are treated as items while words (terms) form transactions. This database presents an alternative view of which documents contain the particular words. By the concept of association rule mining, we can find all itemsets which are the subsets of documents and extract the rules among a set documents.

The support of an itemset is defined by a percentage of the number of words (terms) in which that itemset occurs as a subset to the total number of all distinct words (terms) in a corpus. The confidence of a rule is the conditional probability that a word (term) contains all the document in the rule, given that it contains the documents appearing in the antecedent of a rule. There are two user-specified thresholds, i.e. minimum confidence and minimum support, concerning with two tasks of mining. The first task is to discover all itemsets the supports of which is at least *minimum support* (called *frequent itemsets*) and the second task is to discover all rules the confidences of which is at least *minimum confidence* (called *association rules*). The minimum support and minimum confidence are used to limit the lower bound of statistical significance in the aspect of support and confidence, respectively.

Word (Term)	Documents			
	d_1	d_2	d_3	d_4
computer	1	1	1	1
image	1	0	1	1
artificial	0	0	1	1
intelligence	0	1	1	0
neural	0	1	1	0

frequent itemsets	%support
d_2	60
d_3	100
d_4	60
d_2d_3	60
d_3d_4	60

association rules	%confidence
$d_2 \rightarrow d_3$	100
$d_3 \rightarrow d_2$	60
$d_3 \rightarrow d_4$	60
$d_4 \rightarrow d_3$	100

Figure 2. An example of the pivoted textual database, frequent itemsets and association rules using 50% minimum support and 50% minimum confidence

Given the textual database, frequent itemsets and association rules using 50% minimum support and 50% minimum confidence are listed as shown in Figure 2.

3 Topic-Based Relationships and Document Representations

3.1 Topic-Based Relationships

The topic-based relationship can be defined as basis relationships of any set of documents that are similar in their semantic contents. The topic of any document can be revealed by the content of that document. There has been much interest in identifying the topic of the documents such as those in [7],[8]. The approach to correctly identify the topic-based relationships among documents is not completely trivial since we may face with many problems when unstructured documents are considered. Furthermore, even we can obtain the potential approach to extract the topic-based relationships, there are no benchmark data to evaluate the results with human judgement. However, scientific literature is a kind of documents which obviously present a sort of topic-based relationship in its citations. The citations contained in scientific research publications can be considered as a link between the citing and cited publications. Most of citations are used for giving the credit to previous works related to that article. Anyway, in some cases they may be used for common citations and have little or no relations to the content of the document.

The co-occurring words in a set of documents can be used as clues to detect topic-based relationship. Unfortunately, simply representing a document by a bag of single words (single terms) may not be useful when applying in the large variety of domain because of the ambiguity of words and uncertain meaning of words in the context. Towards these problems, we focus on finding suitable document representation that is sufficient to extract the topic-based relationships.

3.2 Document Representations

In this section, we discuss the details of selected document representations. We describe the task of constructing a textual database from our corpus. We model that the association rules will reveal the topic-based similarity among documents. The documents are represented as a list of bags of words (BOWs), one bag for each column of textual database. After removing punctuation and capitalization from the document, words were defined as unique strings separated by spaces. We currently eliminate 524 commonly-occurring stop-words (e.g. "a", "is", "his", "the"). Two types of corpus representations are used: stemming and non-stemming. With Porter's stemming algorithm [9], any word in a document will be stemmed to the root word. In this work, we also focus to study the characteristic of *n*-gram term representations [10] which can be used to reveal the topic-based relationship. We use unigrams and bigrams as term representations. Unigrams are single isolated word, and bigrams are any two successive words occurring in the text. The concept of bigrams in our approach is as follows. The bigram representation must contain two actual contiguous words in the sentence before the exclusion of stop-words. For example, two sentences "Mary has a red lamp. Her godfather, Ben Johnson, gives it to her" has six unigrams, i.e. "mary", "red", "lamp", "godfather", "ben", and "johnson", and two bigrams, i.e. "red lamp" and "ben johnson", when the stop-words and punctuation (i.e., "has", "a", "gives", "it", "to", "her", "." and ",") are eliminated. In order to avoid the excessive distinct terms for mining process, we used only the terms that have term frequency more than one as the representation of documents.

4 Implementation

4.1 Document Acquisition

We use research articles retrieved from the Web as a corpus for discovering topic-based relationships among these research publications. (A corpus is a collection of research publications retrieved from the Web.) We collected the research publications from ACM Digital Library [1]. Three classes of CCS; B:Hardware, E:Data and J:Computer, are selected as three search keywords. In each class, top 200 articles in the PDF format and their information pages in HTML format are collected as seeds. We then explored to collect the publications in the PDF format and their information pages which are referred by seeds, and added them to the set of seeds. After three iterations, totally 10865 research articles are collected and used as the dataset for our experiments. We then use the *pdftotxt* Plug-In of Adobe Acrobat [2] to convert all publications in the PDF format to the ASCII text format.

4.2 Term Extraction

The term extraction process plays an important role to preprocess a corpus of documents in order to pass the documents to the mining process. Using the BOW library [11] for the bag-of-words text processing, bigrams of each articles will be generated. The term which contains stopwords and its term frequency lower than two will be pruned. Every term and its term frequency are generated to form textual database of a whole collection of research articles.

4.3 Mining Process

The mining process is considered as a key component for the performance of the overall process. In order to test our idea, it is necessary to use a practical algorithm that performs an efficient mining process. The ICDM'03 workshop had concentrated on the performance of various association rule mining algorithms based on the process of mining frequent itemsets in different parameters and datasets [12]. In different algorithms, the same set of frequent itemsets is obtained with different performances, i.e. computational time and memory utilization. The reports showed that the FP-tree algorithm performs the best or second best in many datasets and variations. In this work, the FP-tree algorithm is used as the association rule mining algorithm.

After the term extraction process, a textual database can be constructed from a corpus. Traditional association rule mining can then be applied to this database to discover topic-based relationships. As shown in Figure 2, samples of association rules are presented with their confidences. The mining task is to discover the rules that include the documents which have the topic-based relationship with each other.

5 Accuracy Evaluation

To evaluate the discovered association rules, we formed a benchmark of evaluation. In this approach, discovered association rules are only useful and informative if it shows the topic-based relationship among documents. Therefore it is important to measure the accuracy of discovered rules on the dimension of topic-based. A primary question we address in this section is whether discovered rules are relatively reliable compared to knowledge provided by the authors of the documents. In our insights, the author information of research publication is usually a good clue for

[1] http://www.portal.acm.org

[2] http://www.adobe.com

finding the topic-based related documents in the reference section of that document. However, we can extend the concept of citation by the transitivity function. The citation transitivity function is defined as follows.

Definition 1: i^{th} *Order Citation*. Y is i^{th}-order citation of X if there is at least one citation path from X to Y via i-1 document(s). X is called the i^{th}-order citation of Y.

For example, document A is referred by document B, document B is referred by document C, and document D is also referred by document C. A is the 1^{st}-order citation of B, A is the 2^{nd}-order citation of C, A is the 3^{rd}-order citation of D, and so on.

Although the citation information can be extracted using the NLP techniques as described in [2], we used the information pages which are previously defined in the citations of each article provided by the ACM Digital Library. Nevertheless, we construct a matrix for each i^{th}-order citations of every document in order to evaluate the rules.

Definition 2: i^{th} *Order Citation Matrix*. The i^{th}-order citation matrix is the matrix size $n \times n$ where n is the number of distinct documents. Assume that X and Y are documents where $X \neq Y$, δ is defined on the relation between X and Y, δ = 1 when X is the j^{th}-order citation of Y where $j \leq i$ otherwise δ = 0.

To show effectiveness of discovered rules in revealing topic-based relationship, a set of discovered rules will be evaluated by these i^{th}-order citation matrices. The accuracy of each rule is defined by the ratio of the maximum number of documents in the rule which is referred by each other to the total number of documents in the rule. Therefore, the accuracy value of each rule becomes one when all documents in the rule is referred by each other. For example with a rule $A, B \rightarrow C, D$, if A, B, C are found as a largest group citing/cited located in a same line of citation matrix, then the accuracy of this rule is 3/4 = 0.75. An accuracy result in the next section is shown in the form of percentage value which is the sum of the accuracies of all discovered rules divided by the number of rules.

6 Experiments

In all experiments, we evaluate the performance of discovered rules by using accuracy evaluation method described in section 5.

6.1 Environments and Parameters

The experiments were performed on a Pentium IV 2.4GHz Hyper-Threading with 1GB physical memory and 2GB virtual memory running Linux TLE 5.0 as an operating system. The data are stored on SCSI RAID5 70GB. All algorithms and processes are implemented with C/C++ language based on UNIX library.

6.2 Experimental Results

6.2.1 Citation Matrix

Intuitively, different matrices provide different kinds of information. The accuracy in the 1^{st}-order citation matrix is quite low in every dimension of the result table. With the association rule mining, each rule provides a set of documents which contain a number of co-occurred word terms. Most of the 1^{st} citations of any document do not contain the co-occurred terms with that document. This results evidence the lack of topic-based relationships. They may be referred since the specific part of the literature needed the citation without concerning in its topic or they are unsuitable citations. However, if the documents refer to the same citations or even the documents are referred from the same citations, those sets of documents highly tend to contain the co-occurred word terms and have topic-based relationships with each others. This assumption can be proved by the results as shown in Table 1. In each line, the accuracy of the 1^{st} citation matrix is extremely lower than the other matrix, and the accuracy of 2^{nd}- and 3^{rd}-order citation matrix are in the vicinity. We conclude that the citation matrix with an order of more than one represents the topic-based relationship.

6.2.2 Document Representations

The results in Table 1 show the mining performance of different document representations. Three kinds of document representations, i.e. unigram, bigram and unigram+bigram, are explored. In order not to introduce any bias, the number of frequent itemsets and the number of rules are fixed for each representation. We do not directly control the minimum support and minimum confidence because their ranges of percentage to specify equality number of frequent itemsets and number of rules are inconsistent when different document representations are applied. The percentage of co-occurred unigrams is likely more than unigram+bigram and bigram, respectively. As the result, bigram representation outperformed the unigram+bigram and unigram, respectively. This means that bigram seems to be better than unigram and unigram+bigram.

6.2.3 Number of rules

The rows 1, 7 and 14 show the accuracy when we evaluate all frequent itemsets rather than the association rules. To limit the number of rules, we specified the minimum confidence. A smaller number of rules is gained with higher minimum confidence, and a larger number of rules is obtained with lower minimum confidence. The minimum confidence helps to filter out the rules that have low confidence for revealing the topic-based relationships. However, too high minimum confidence may not provide high accuracy in the case of bigram. Moreover, we investigate

Table 1. Accuracy results of discovered association rules

row#	#freq	#rule	unigram			bigram			unigram+bigram		
			%accuracy of matrix			%accuracy of matrix			%accuracy of matrix		
			1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}
1	5000	-	3.59	22.41	57.29	25.72	79.50	95.91	5.00	27.30	62.01
2		100	0.00	67.00	86.50	26.50	93.50	97.00	4.33	76.33	88.67
3		500	15.43	45.77	71.53	37.57	96.60	98.90	18.03	55.27	79.37
4		1000	12.40	41.58	72.38	43.42	97.80	99.30	12.33	47.68	78.92
5		5000	7.04	33.22	69.12	42.83	92.97	98.53	9.50	40.03	74.24
6		10000	5.31	27.47	63.35	29.90	84.19	97.14	6.62	32.01	67.22
7	10000	-	3.31	21.68	57.22	21.21	73.00	94.11	4.27	25.42	60.57
8		100	2.33	63.33	83.00	27.83	93.33	98.50	3.33	66.00	84.33
9		500	12.17	51.30	76.20	34.60	94.30	98.70	16.13	59.67	80.83
10		1000	12.37	43.97	71.07	36.60	94.50	99.20	16.52	50.70	75.73
11		5000	6.61	34.79	75.27	47.23	95.09	99.03	7.50	37.75	77.99
12		10000	6.05	32.85	71.22	37.41	89.64	98.07	8.68	38.03	74.46
13		20000	4.84	27.28	64.10	26.15	79.22	96.03	6.00	31.04	67.61
14	20000	-	2.99	21.64	58.21	17.00	66.95	91.80	3.77	24.63	60.87
15		100	4.00	62.50	88.00	18.00	84.00	95.50	3.50	64.00	87.00
16		500	5.72	60.55	82.25	23.93	83.18	96.93	8.63	65.42	85.70
17		1000	10.46	50.72	77.84	24.83	82.00	97.00	13.79	56.56	81.93
18		5000	9.05	38.80	72.30	28.90	84.64	97.67	9.88	41.37	75.90
19		10000	6.43	35.60	74.42	31.28	87.91	98.48	6.82	36.50	76.18
20		50000	4.50	27.77	66.24	21.75	74.37	94.44	5.25	30.30	68.63
21		75000	4.20	26.56	65.28	20.98	72.86	93.65	4.92	28.92	67.10

on the characteristic of these rules and found those documents, which are incorrectly predicted, are the same documents appearing in different publications or quite similar content documents published in the same year. They tend not to refer to each others. That is, these rules should be assumed to contain topic-based relationship even they have no citation to each others.

6.2.4 Number of frequent itemsets

With different minimum supports, we investigate three different numbers of frequent itemsets. We found that when we increase the number of frequent itemsets (lower the minimum support), the accuracy decreases trivially. In practical, large data may need a high minimum support to filter out most of document combination and avoid expensive computational time. This makes the process feasible because the informative rules can occur when a high minimum support is set.

6.2.5 Stemming

To investigate more about document representation, words (terms) with/without stemming are compared. The previous results show the accuracy of the case that we make use of the stemming. We then select the cases that gain the highest accuracy in all cases of document representation, i.e. bigram, for further exploration of the stemming approach. The result is shown in Table 2. The performance of the discovered rules using stemming is higher than that of the discovered rules using non-stemming.

6.3 Error Analysis

In the previous experiments, we found that some rules are incorrectly predicted as topic-based relationships. To check their validity, we investigate the reason why they are incorrect. We found that the set of documents in these rules have the following characteristics.

- They are the same documents which appear in various version of publications or they are the minor change articles. They do not directly refer to each others. By evaluation, these rules are wrongly predicted when using the 1^{st}-order citation matrix but they will be correctly predicted when evaluating by the 2^{nd}-order citation matrix and succeeding order citation matrices.
- Those documents have topic-based relationships but they do not link to each other or even share the same citing or cited articles, since they are published in the same year or same proceedings. We know that these documents should have the topic-based relationships because they appears in the same title of proceedings.
- There are some mistakes in the information pages downloaded from ACM Digital Library. In the information pages, some citations appearing in the publications are not given. Since we extract the citation matrix using the links from the citations in information page of each paper, therefore the mistake triggered by this missing information. Additionally, if the papers are not located in the ACM database, then there is no link between them. This result is not true when those documents in a rule sharing the same citation. There are 5-10 rules of this type.

Table 2. Accuracy results on bigram: stemming vs non-stemming

#freqs		5000			10000			20000		
#rules		100	1000	10000	100	1000	10000	100	1000	10000
non-stemming	%acc 1st	15.17	22.90	31.53	18.83	2.43	1.74	15.00	3.75	0.84
	%acc 2nd	75.50	74.95	84.52	69.50	38.44	43.89	52.50	33.78	37.37
	%acc 3rd	94.50	94.40	96.59	84.00	69.35	74.85	64.50	60.97	72.84
stemming	%acc 1st	26.50	43.42	29.90	27.83	36.60	37.41	18.00	24.83	31.28
	%acc 2nd	93.50	97.80	84.19	93.33	94.50	89.64	84.00	82.00	87.91
	%acc 3rd	97.00	99.30	97.14	98.50	99.20	98.07	95.50	97.00	98.48

7 Conclusion and Future Work

This paper shows the way to apply association rule mining to reveal the topic-based relationship among the documents with several document representations. In this work, bigrams with stemming excluding stopwords are the best document representations for revealing the topic-based relationship. In practical, the mining algorithm is appropriate to use as a tool for learning a large corpus of documents, and the informative rules are generated, even setting high minimum support and high confidence. We also propose a benchmark citation matrix to evaluate the discovered association rules. The results in various order of citation matrices provide different kinds of information. The topic-based relationship may not be detected by the citation matrix with an order of more than one.

Another direction for future work is to apply the information of term frequency, which is well-known in text categorization, for our mining process. Later, we will also evaluate the discovered association rules with human judgement and investigate the incorrect predicted rules in depth. Furthermore, the vocabulary problems will be taken into accounts. Since different terms may convey the same meaning, we can reduce the large dimension of terms to be a small dimension of semantic and concept. Finally, this approach is currently applied to the scientific research publications, however various kinds of documents are challenged for textual learning.

8 Acknowledgements

This paper has been supported by Thailand Research Fund (TRF) and NECTEC under project number NT-B-06-4C-13-508.

References

[1] D. Gibson, J. Kleinberg, and P. Raghavan, Inferring web communities from link topology, *Proc. of the 9th ACM Conf. on Hypertext and Hypermedia*, Pittsburgh, Pennsylvania, USA, 1998, 225–234.

[2] C. L. Giles, K. Bollacker, and S. Lawrence, CiteSeer: An automatic citation indexing system, *Digital Libraries 98 - The Third ACM Conf. on Digital Libraries* (I. Witten, R. Akscyn, and F. M. Shipman III, eds.), Pittsburgh, PA, 1998, 89–98.

[3] R. Weiss, B. Velez, M. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, and D. Gifford, Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering, *Proc. of the 7th ACM Conf. on Hypertext*, Washington, USA, 1996, 180–193.

[4] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, Syntactic clustering of the web, *Proc. of the 6th Int'l Conf. on World Wide Web (WWW-6)*, Santa Clara, CA, 1997, 391–404.

[5] R. Agrawal, T. Imielinski, and A. N. Swami, Mining association rules between sets of items in large databases, *Proc. of the 1993 ACM SIGMOD Int'l Conf. on Management of Data* (P. Buneman and S. Jajodia, eds.), Washington, D.C., 1993, 207–216.

[6] C. Zhang and S. Zhang, *Association rule mining: models and algorithms*. Springer-Verlag New York, Inc.,2002) .

[7] C. Clifton and R. Cooley, Topcat: Data mining for topic identification in a text corpus, *Principles of Data Mining and Knowledge Discovery*, 1999, 174–183.

[8] X. He, H. Zha, C. Ding, and H. Simon, Web document clustering using hyperlink structures, 2001.

[9] M. Porter, “An algorithm for suffix stripping.” Progam, vol.14, no. 3, 1980.

[10] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (. Boston, MA: Addison-Wesley,1989) .

[11] A. K. McCallum, Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996.

[12] B. Goethals and M. J. Zaki, Advances in frequent itemset mining implementations: Report on fimi'03, 2003.