2004－ICS－138　(5)
2004／12／4

Association Rule Mining From Textual Data using Passages

Kentaro Nagai † and Ho Tu Bao†

Discovering knowledge from large amount of textual data is an important problem. Especially, application of association rule mining to textual data has been studied excessively. Many works has successfully found relationships between words that reflects syntactical rules, co-occurences, or phrases. These rules are useful for understanding the liguistic nature, but in real life, the relationships between the topics or contents are important and useful, such as what kind of topic tends to appear in same paper or books. Our objective is to find relationships between contexts or topics. In this paper, we propose an approach to use passages to take in some level of semantics in rule mining. We show some preliminary results to show its potential and give discussions on the problem for further improvement.

1. Introduction

Text mining is a process of finding previously unknown and potentially useful information from large amount of unstructured, natural-language texts. Since the major part of the electronical available data is said to be in unstructured or semi-structured data[1)], that is texts and webs, mining from textual data is an attracting increasing attention. Many standard data mining methods has been extended to deal with unstructured data[2)] and one of them is association rule mining, first proposed by 3).

Association rule mining is task to extract associations from transactional databases. When we apply association rule mining to textual data, we must first consider to convert textual data into transactional format. There are several approaches for different purposes. Most of them used association rule mining to find relationships between single word or phrases. Relationships between words or phrases are useful for revealing the linguistic nature, but in some cases relationships between topics and contents is more important like in the task of emerging trend detection. Another problem of using single words is that the words are often ambiguous without contexts.

To overcome this problems, we propose an approach to passages as items in association rule mining. One objective is to find relation between topics, another is to increase the understandability of rules.

In section 2, we analyze the related works to see the features and limitations. In section 3, we propose our approach. In section 4, we show the result of preliminary experiments. In section 5, we give discussion the result of the experiments. Last section is the conclusion.

2. Background

2.1 Mining Association Rules from Texts

There has been several works of applying traditional association rule mining methods to discover relationships from textual data. The major problem we face in applying traditional association rule mining to textual data is the conversion of texts into transactional format. Transactional format consists of transactions and items within transactions. Selection of transactions and items depends on their purpose of application and, in fact, many researcher took different linguistic units as trasactions and/or items with different purposes. In this section, we will look into the related works to see how the selection of items and transactions affect the result rules.

2.2 Selection of Transactions

The approach can be classified to 3 main approaches: documents[4)5)], windows[6)], and passages[7)]. In traditional market basket analysis, one transaction is equivalent to one event. When dealing with texts, selection of transaction is equal to selecting the scope of co-occurence. For example, if you choose documents as transactions and words as items, you were to regard the occurence of word in the

† School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa, 923-1292,
email:{k-nag,bao}@jaist.ac.jp

opening and the closing of document as co-occurence as well as the words appear next to each other. In many cases, this assumption is too strong. 7) proposed of using passages as transactions to bring some level of locality in counting the co-occurences. In this case, the result rules shows the co-occurences within the same or similar contexts.

2.3 Selection of Items

We classified the works into 3 major groups depending on the selection of items: Word-based approach, Term-based approach, Entity-based approach. We will take closer look into each approach in the following subsections.

2.3.1 Word-based Approach

This approach uses all or most of occurrences of words as items. Initial work is done by 4). They used documents as transactions and words as items. Their experiments showed many relationships between words, but some problems become clear. This approach often produces huge number of redandunt rules, and some of the rules are un-interpretable. Another work is lead by 6), using window as a transaction and applied frequent sequence mining to find frequent phrases. Resulted rules show the frequent phrase like "Knowledge Discovery in → Databases". Since this approches' result show the relations between single words, the result tends to be syntactic relations or co-occurence of words.

2.3.2 Term-based Approach

Second group is the ones which use terms as items introduced by 5). The word "term", in 5)'s sense, is a sequence of words that conform single meaningful unit. For example, "net income" and "joint venture". First they parse the documents to extract terms based on the co-occurence, and then apply association rule mining to find relations between terms. They successfully mined relation of joint ventures like "(america online inc)(bertsmann inc) → (joint venture)". The results are more interpretable and meaningful compared to the word-based approaches. This could be easily understood by the fact that each item is more rich in meaning. A word alone is often ambiguous. For example, rule like "net → joint" will not be understandable and/or useful as previous example.

2.3.3 Entity-based Approach

The last approach is to use entities as items[8)9)10)]. Entities are the words or terms extracted by information extraction method. Information extraction is a method to extract entities like person name, location, company names from free texts, often guided by some kind of prior knowledge about the entities to extract. For example, to extract job type, required skills, ages, and locations from job listings. This approach uses information extraction as a first step and then apply association rule mining or its variants to mine rules to find relationships between entities. The resulted rules are comparable to or better than term-based approaches in the sense of understandability or usefulness. This is partly because the entities are words or terms extracted with some level of intention. So it might be fair to assume that rules provide minimum level of interestingness to users since they already filtered out the non-interesting items. In the term of data mining field this process can be seen as noise reduction and these works are especially successful in reducing huge number of redundant rules by focusing on items in interest.

But there are some limitations for this approach. Most limitation comes from the limitation of information extraction methods. Information extraction often requires templates or extraction rules. These prior knowledge is often differs from domain to domain, and therefore it can be said that this approach has some limitations of domain dependence[†y]. Another problem or question for this approach is that they might be dropping useful items in what they regarded as noises. As mentions above, information extraction are often used with some level of intention. The kinds of the words to be extracted are often required to be predefined. In this way, we could find relationships between known items, but we could not find out relationships between un-awared items. This problem might not be significant in many cases, but we must be awared that we are losing the chances to find relationships between those unpredicted items.

[†y] Recent study of information extraction uses finite state transducers or ontologies-driven approaches to overcome this problem, but here we only think about the traditional information extraction.

3. Proposed Method

In the previous section we showed several approaches for selecting transactions and items, and their differences and limitations. In short, understandability of rules is strongly affected by the quality of meaning of items. Entities-based approach can overcome this problem but has many limitations. The question is whether there is a semantically rich unit without the limitation of entities-based approach.

In this paper we propose the use of passages as items and documents as transactions. In other words, our approach is to view document as bag of topics or contexts and find relationships between topics.

3.1 Passages

Passages, in a general, just means an excerpt of a document. But we redefine the meaning of passage as below.

Definition. Passage is an excerpt of a document, which holds some contexts or topics.

Using of passages has been studied in many fields like information retrieval, document clustering, question and answering, word sense disambiguation. Not all the work use passages in the same definition as ours, but some works' results are good support of our definitions. Especially the work by 11) showed that information retrieval based on passages got better result than the ones based on documents, especially when when the length of the document statements are long. This result implies that users' interest lies not in the whole document but often in the part of the document.

Using of passages is an attracting approach but there are some problems to overcome. One problem is identification of passages, another is representation of passages.

For the identification of passages, the most simple approach is to use the natural paragraphs. This is because paragraph is the boundary of topic or context given by the author. But paragraphs are not always available and sometimes paragraphs are not sufficient in granularity. There are also many works to identify the passages boundary automatically[12)13)]. In this paper we will not go further into this problem but for future works, it is worth while considering an algorithm for identifying passages that suits for our purpose.

Representation of passage is also an important problem. We usually have to represent some data in a format that is easy to process but still retains the original information. As to our knowledge there is few work regarding to representaiton of passage. But the studies for document respresentation is similar or same to passage representation. Some of which are index terms (or keywords), vector space model[14)], bag of words and latent semantic indexing[15)]. These representation can be directly applicable to passages. We chose keywords as for an experiment due to its simplicity. We will discuss more on this problem in the later sections.

3.2 Why Passages?

Before going into system overview. We will give 3 main reasons why we have chosen passages as items.

The first reason is that passages can be seen as meaningful units. As mentioned in previous section, 11) work shows that passage is good unit of users' interest. Question and answering field uses passage retrieval as a first step to find answer to the question. This indicates that we could see passages as representation of single fact or evidence.

The second reason is that using of passages can be semantically less ambiguous compared to words. For example, the word "bank" alone can mean (1) financial organization or (2) the side of river. But passages consists of several words and in many cases we could solve this kind of ambiguity by looking at the words appear near. This is also known as distributional hypothesis: "the meaning of the words that appear in similar contexts tends to be similar." We can expect to get rules that are less ambiguous than the rules consists of words alone.

The third reason is that using of passage will reduce the risk of generating rules about words that appear in different contexts. This will be remarkable when the document's content is long and contains several topics.

4. Evaluation

We carried out a small experiments to see its effectiveness.

4.1 Data Sets

We used the data sets of news articles of Los Angeles Times from TREC5(Text REtrieval

#6526: (juice-lemon-tablespoon)
→(fat-plain-yogurt) [support:7]
#6617: (manufacturing-semiconductor)
→(hiring-reducind-trend) [support:63]
#13731: (family-leave-sick)
∧(handgun-possession-purchase)
→(abortions-afford-woman) [support:5]

Fig. 1 Examples of mined rules

Conference)[†y] collection. This collection has been tagged with SGML.

We used the paragraph tags to identify the boundaries of passages. Headings are also regarded as one paragraph. After identifying the boundary, top 3 words are selected from the passages. The score similar to TFIDF[14)] is used to select the keywords. The only difference between TFIDF and our scoring is that we used inversed paragraph frequency (IPF) instead of inversed document frequency (IDF). Let D be the data set, d as a document in D, w as a word in d then, the score of IPF is given in the next formula.

$$IPF(d,w) = \frac{\#\text{ of passages which include } w}{\#\text{ of passages in } d}$$

The passages are notated as (keyword1 - keyword2 - keyword3), where one bracket represents one passage. Keywords are sorted alphabetically to avoid redundancy. Stopwords are filtered with SMART system stopwords[16)]. We did not apply stemming. We used the Apriori[?)] implementation by 17) for association rule mining. We did not use confidence threshold.

4.2 Results

Results are given in tables and figures. Table.1 shows the number of rules with different support threshold. **Fig. 1** shows some examples of rules mined from this experiment.

The rules obtained are mostly from periodical articles like recipes, public statements, space for rent, etc. But we could see that the result rules catch some of the nature in data sets.

Here we explain the rules in Fig. 1. The rule #6526 arise from articles of recipe. This rule can be interpreted as "low-fat (or fat-free) plain yogurt is used with tablespoon(s) of lemon juice." The rule #6617 come from business trend forecasts. The antecedent has only 2 words. This is because the "semiconductor manufacturing" is the heading. The rule #13731 is from questionnaire for senators. This rule can be interpreted as "Women abortions problem are also asked in questionnaire with family sick leave, and purchasing and possession of handguns."

Minimumsupport	Numberofrules
0.01	321,070
0.05	1,094
0.10	634

Table 1 Number of rules generated on different threshold

[†y] http://trec.nist.gov/

5. Discussion

In this section we discuss some of the problems that became clear after the experiment. We also give some possible improvement or refinement.

5.1 Matching Passages

In the experiment we selected set of keywords from the paragraphs and tried to find the passages that has exactly the same keywords. But, in nature, texts have wide variety of words to express similar or even the same meanings.

For example, (father-mother-child) and (father-mother-infant) can be seen a similar passage since "child" and "infant" are similar in sense and other terms are the same. Same kind of problem happens when there is passages like (tea-pot-price) and (coffee-pot-price). We failed to catch the possible topic (pot-price). Another similar problem happens when there is a passage that contains several topics. Currently our system cannot handle passages with multiple topics. For example, the passage about "military hospital" cannot be recognized as topics of "military" and "hospital."

To solve these problems, we are considering to apply soft matching association rule mining. The work of extending traditional association rule mining to support soft matching is done by several researchers. One of the work can be seen in 8). Although we could adopt their methods directly, 8)'s method are oriented to the words extracted by information extraction, and may need some modification for our purpose. For example, they proposed string edit distance based similarity, but this is effective when we want to group similar spelling words like "Netscape 4.5" and "Netscape 5". We may need some extention to support semantic soft matching. One

way is to use vector space model[14] as in 8). Another way is to generate the tolerent class of words to cluster the words with similar meaning and regard those words as replacable to each other[18]. Tolerance class is the sets that satisfies only the reflexive and symetric property. 18) used tolerence class for document clustering and showed its effectiveness in dealing with textual data. Finally, the way to utilize the lexical knowledge is also seeked. We are considering to use WordNet[19] to calculate the semantic distance of words in passages, and furthermore the similarity of passages.

5.2 Representation of Passage

Three words representation is poor in many senses. Passages might contain few words to several hundred words, but it is always represented in three words. Since passages have variable length, fixed size keyword set is not appropriate in many cases. We are now planning to use all the keywords with score higher than a threshold or to normalize the size of keyword set in proportion to the passage length.

Passages have few words compared to documents. We used TFIPF to select keywords of passages. But IPF alone could be better. This is because we want to distinguish the topic represented in a passage. IPF is the score that put high score on the words that appear only in a certain passage. Term frequency is useful information for knowing the document but it might well to assume that document's keyword appear in most of the passage with some frequency. It might not be appropriate to choose document keywords for every passage's keywords.

5.3 Redundant Rules

As shown in Table.1, the system produces enormous number of rules so it is hard for users to check all the rules. Although this is a common problem in association rule mining, we would need to improve this situation. Obviously one way is to set higher support threshold, but we might lose the chance of finding rare but highly confident event. Solving the hard matching problem is promissing in that similar rules will flock together with soft matching. Another approach is to reduce or rank the rules according to additional rule measure of rule interestingness. Our system has no rule evaluation module yet, but there have been many works on developing measures for rule interestingness[20]. Some works try to find rules using background knowledge to show only the unusual or unexpected rules. Especially 21) proposed a measure for text-mined rules, in which they utilized the lexical knowledge to evaluate the text-mined rules' semantic surprisingness. Similar thing can be thought in our work.

Another way to overcome this problem is to extend the successful works from standard data mining community. For example, mining of frequent closed item sets has been studied[22], and is successful in removing redundant rules.

6. Conclusions

In this paper we proposed an approach to apply association rule mining to textual data. Our work is different from previous works in that our aim is to acquire relationships in semantic level like between topics or contexts, while most of the past works focus on the syntactic units. We carried out an experiment to show its effectiveness and reveals its potential of using passages. We also shown several problems that become clear after the experiment. We gave discussions on the problems and proposed possible improvement for the future works. We will continue to improve this framework to make the system more flexible and reliable.

References

1) Treloar, N.: Text Mining: Tools, Techniques, and Applications, *Knowledge Technologies Conference* (2002).

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

3) Agrawal, R., Imielinski, T. and Swami, A.: Mining Associations between Sets of Items in Massive Databases, *Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data*, pp. 207–216 (1993).

4) Rajman, M.: Text Mining, knowledge extraction from unstructured textual data, *Proceedings of EUROSTAT Conference* (1997).

5) Feldman, R., Fresko, M., Kinar, Y., YehudaLindell, O. L., Rajman, M., Schler, Y. and Zamir, O.: Text Mining at the Term Level, *PAKDD*, pp. 65–73 (1998).

6) Ahonen, H., Heinonen, O., Klemettinen, M. and Verkamo, A.I.: Applying data mining techniques in text analysis, Technical Report Report C-1997-23, Department of Computer Sci-

ence, University of Helsinki (1997).
7) Theeramunkong, T.: Applying passage in Web text mining, *International Journal of Intelligent Systems*, Vol. 19, No. 1-2, pp. 149–158 (2004).
8) Nahm, U. Y. and Mooney, R. J.: Mining Soft-Matching Rules from Textual Data, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, pp. 979–984 (2001).
9) Feldman, R., Aumann, Y., Fresko, M., OrlyLipshtat, B.R. and Schler, Y.: Text Mining via Information Extraction, *PKDD*, pp. 165–173 (1999).
10) Clifton, C. and Cooley, R.: TopCat: data mining for topic identification in a text corpus, *In Proceedings of the 3rd European Conference of Principles and Practice of Knowledge Discovery in Databases* (1999).
11) Salton, G., Allan, J. and Buckley, C.: Approaches to Passage Retrieval in Full Text Information Systems, *SIGIR*, pp. 49–58 (1993).
12) Hearst, M.: TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, Vol. 23, No. 1, pp. 33–64 (1997).
13) G.Hirst and St-Onge, D.: *Lexical Chains as representation of context for the detection and correction malapropisms*, The MIT Press, chapter 13 (1997).
14) Salton, G.: *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley (1989).
15) Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T.K. and Harshman, R.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407 (1990).
16) Cornell University SMART System: ftp://ftp.cs.cornell.edu/pub/smart/.
17) Goethals, B.: Apriori frequent itemset mining implementation, http://www.cs.helsinki.fi/u/goethals/.
18) Ho, T. B. and Nguyen, N. B.: Nonhierarchical Document Clustering by a Tolerance Rough Set Model, *International Journal of Intelligent Systems*, Vol. 17, No. 2, pp. 199–212 (2002).
19) Fellbaum, C.(ed.): *WordNet, an electronic lexical database*, MIT Press (1998).
20) Tan, P. and Kumar, V.: Interestingness Measures for Association Patterns: A Perspective, Technical Report TR00-036, Department of Computer Science, University of Minnesota (2000).
21) Basu, S., Mooney, R. J. and Krupakar V. Pasupuleti, J. G.: Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge, *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, pp. 233–238 (2001).
22) Pei, J., Han, J. and Mao, R.: CLOSET: An efficient algorithm for mining frequent closed itemsets, *DMKD 2000*, pp. 11–20 (2000).