# Association Rule Generation and Mining Approach to Concept Space for Collective Documents

Ken Nittono
Hosei University; Faculty of Business Administration,
2-17-1 Fujimi, Chiyoda, Tokyo 102-8160, JAPAN

## Abstracts

An approach to extracting essential terms and mining intrinsic contexts for a large amount of documents is studied. An application based on the apriori algorithm is proposed for the extraction of terms and indexing methods for abstraction of a set of documents by the use of latent semantic indexing on a conceptualized document space are discussed.

Keywords: big data, collective intelligence, latent knowledge, text mining

## 1. Introduction

In recent years, the amount of exchanged online documents has been growing with the diversification of computer network services and terminal device technology. And many kinds of publications which have been published traditionally in paper-based style such as books, proceedings and transcripts become to have more opportunities to be addressed in digitized text data. Based on those backgrounds, attempts to analyze such large amounts of text data and trial to obtain useful knowledge have been attracted attention, and various methods are proposed in many areas. In those circumstances, this study aims to mine collective documents for significant terms or contexts and extract particular information. Finding desired information from the documents in such cases as books or journals that are systematically edited by the author or editor is relatively straightforward, however, it is rather difficult to find it with later reading for the whole data which is recorded in accordance with passage of time such as communication log in some network service or transcript of interview because they often contain redundant expression, incomplete sentence or sometimes irrelevant context. The latter case of text data is dealt with in this study and a method for extracting essential part from large text data is proposed.

In the approach, the documents are formulated as a term-document matrix. And some characterizing terms or phrases are mined throughout the documents by the application method based on association rules (Agrawal et al. 1993; Agrawal and Srikant 1994). On the occasion of the generation of association rules, new meanings for the traditional formulation of support, confidence and lift are given. And the amount of information to be extracted from original documents as a result depends on the new meanings and adjustment of several parameter values for the generation of the rules.

In the next stage, essential contexts are extracted from original documents by the use of latent semantic analysis (LSA). The whole documents are translated in a concept space by singular value decomposition (SVD) and the space enhances latent meanings contained by the documents. Mining the concept space with terms or phrases obtained by the above method enables to extract essential contexts or paragraphs. Specifically, an applied method based on latent semantic indexing (LSI; Deerwester et al. 1990) is proposed and some approaches to indexing documents with the generated association rules are discussed. And some relevant problems and application areas are also discussed in the latter part.

## 2. Generation of Association Rule

Let $t_i$ and $T = \{t_1, t_2, ..., t_M\}$ denote a term and a set of all terms gathered throughout all of target documents, respectively. Then association rule for terms in the documents is represented by $X \Rightarrow Y$, where *X* and *Y* are subsets of a term set *T*. And quality measures support, confidence and lift are defined as follows,

$$supp(X \Rightarrow Y) \triangleq P(X,Y),$$
$$conf(X \Rightarrow Y) \triangleq P(Y \mid X),$$
$$lift(X \Rightarrow Y) \triangleq \frac{P(Y \mid X)}{P(Y)},$$

where $P(A) = \sigma(A) / N,\ \sigma(A) = \mid \{d_j \mid A \subset d_j, d_j \in D\} \mid$, $\mid \cdot \mid$ is the number of elements and *D* is a set of all documents which has *N* elements.

For the above formulation obtained by the traditional manner, applied meanings are given, here, based on some conditions that the rules should have their own roles in the context of this study. More precisely, we define particular principles for each quality measure, that is, the support principle (P1) means that terms in the rule imply common knowledge, the confidence principle (P2) means that terms in the rule imply valued information for some sort of knowledge or interest and the lift principle (P3) means that terms imply rare information.

And also define a set of terms *K* by following formula,

$$K \triangleq \bigcup_r (X_r \cup Y_r),$$

where $X_r$ and $Y_r$ are sets of terms generated by one of the above principle and *r* is the number of rules to be adopted. Thus, *K* becomes a set of key-terms which has particular meaning along with the principle and it enables to extract some features from documents, that is to say that the terms implies the source information which has connection to latent knowledge in documents.

In general, calculation burden for searching the rules throughout documents becomes enormously large, thus some devised methods, such as the apriori algorithm, are applied in most practical situations.

## 3. Conceptualization of Collective Documents

For the fundamental analysis of a set of documents or texts, a matrix representation for the documents, that is term-document matrix, is in wide use to involve the whole structure and information.

Let *A* be a term-document matrix on the target documents, then singular value decomposition (SVD) for *A* is obtained by $A = USV^T$ and its *k*-dimensionality reduction formula is represented as follows,

$$\hat{A} = U_k S_k V_k^T$$

By the reduction of dimensionality, an enhanced concept space in the sense of latent semantics is generated. Especially in this case, it indicates exclusion of influence of some fluctuations in the use of words or exaggerated expression of phrases from the original context. The conceptualized document space enables information retrieval methods to search and find involved essential context throughout documents effectively rather than by the normal keyword search for the original documents.

## 4. Mining Concept Space

The aim of the approach is firstly, to measure distance between the extracted term set and each document and secondly, to make ranking of documents with respect to the distance. As a result, the selected documents along with the ranking become an essential or abstracted subset of all documents.

Here, we introduce a modified document retrieval method which is based on latent semantic indexing (LSI; Deerwester et al. 1990). Let *q* denote a query vector composed of the number of terms on the key-term set as the way of term-document matrix. Then a concept space representation of the vector is formulated as follows,

$$\hat{q} = q^T U_k S_k^{-1}.$$

There are many kinds of method for measuring distance between two vectors, however; in this case, cosine is applied to calculate the distance between query vector and each document as the similarity among them, and the formulation is as follows,

$$\cos(\hat{q}, d^j) = \frac{\sum_i \hat{q}_i d_i^j}{\|\hat{q}\| \cdot \|d^j\|},$$

where $d^j$ is a row vector, that is representation of a document, in a right singular matrix $V_k$ of the above LSA representation and $\|\cdot\|$ is a norm of a vector. Thus a similarity ranking is generated with respect to the cosine values for each document.

Table 1 shows an example of similarity calculation (document numbers and their cosine values) for a document set via the principles described above. The result implies that each principle has the ability of capturing particular feature of the target documents.

Table 1. Similarity via P1 and P2

| | | | | | |
|---|---|---|---|---|---|
| Support principle | 118 | 86 | 80 | 119 | 12 |
| | 0.585 | 0.549 | 0.536 | 0.527 | 0.496 |
| Confidence principle | 20 | 10 | 85 | 92 | 34 |
| | 0.671 | 0.535 | 0.481 | 0.416 | 0.378 |

Obviously the results of the measuring depend on the aspects of the singular matrix and its dimensionality reduction. Table 2 shows the definition of 4 types of the query vector $\hat{q}$ and right singular matrix $V_k$.

Table 2. Criteria and formulae

| criterion type | query vector $\hat{q}$ | right singular matrix $V_k$ |
|---|---|---|
| $C_1$ | $q^T U_n S^{-1}$ | $V_n$ |
| $C_2$ | $q^T U_k S_k^{-1}$ | $V_k$ |
| $C_3$ | $S\hat{q}_1^T$ | $(SV_n^T)^T$ |
| $C_4$ | $S\hat{q}_2^T$ | $(SV_k^T)^T$ |

And the above definitions are also regarded as criteria of conceptualization of documents. Criterion $C_1$ is original SVD without dimensionality reduction, that is no enhancement. $C_2$ is based on typical dimensionality reduction by $k$ and the number of $k$ is, for example, determined by a percentage of shared number against singular values of $S$. $C_3$ is a modification of $C_1$ which is emphasized by the product of singular value. And $C_4$ is a modification of $C_2$ which is also emphasized by singular value.

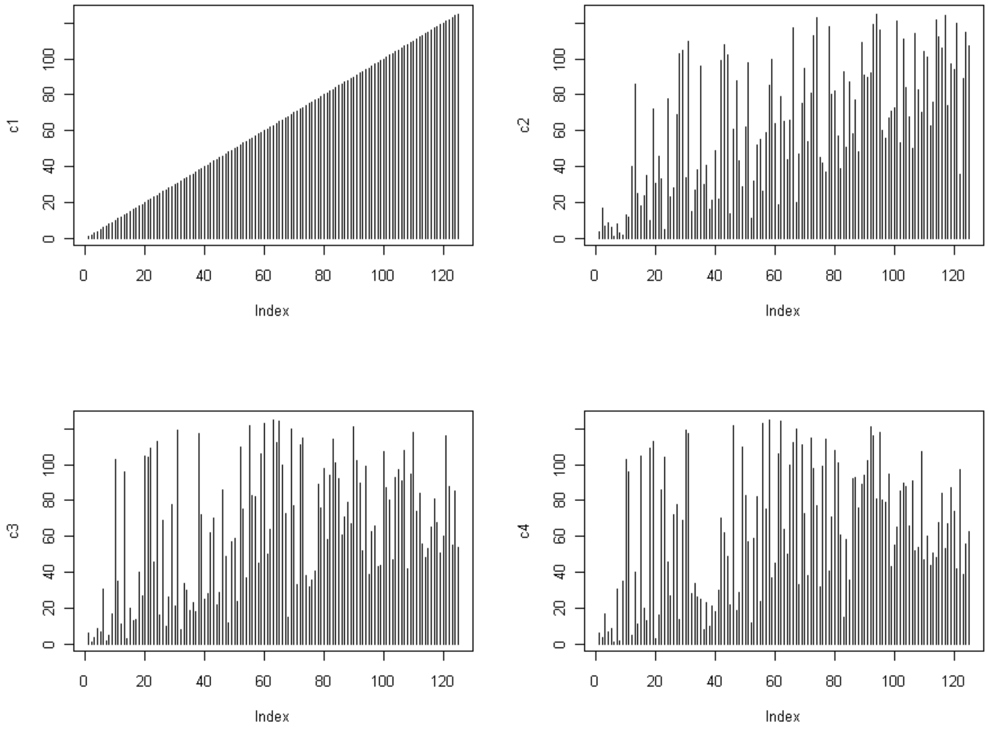

Fig. 1. Indices of documents by similarity (case 1, $f$ =3)

Table 3. Rank order of documents (case 1, $f$ = 3)

| $C_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $C_2$ | 4 | 17 | 7 | 9 | 6 | 1 | 8 | 3 | 2 | 13 |
| $C_3$ | 6 | 1 | 4 | 9 | 7 | 31 | 2 | 5 | 17 | 103 |
| $C_4$ | 6 | 4 | 17 | 7 | 9 | 1 | 31 | 2 | 35 | 103 |

Fig. 1 illustrates the results of the calculation of the similarity for experimental over 100 documents. A key-term set, which has 3 terms ($f$ = 3) in this case, is obtained by P1 principle and similarity is measured by each criterion. The result is sorted by the ranking index of $C_1$. And Table 3 shows the comparison of rank order for top 10 documents by $C_1$ and other criteria. The vague upward tendency is viewed in $C_2$, $C_3$ and $C_4$ criteria, however, at the same time, locally intense variation also can be seen.

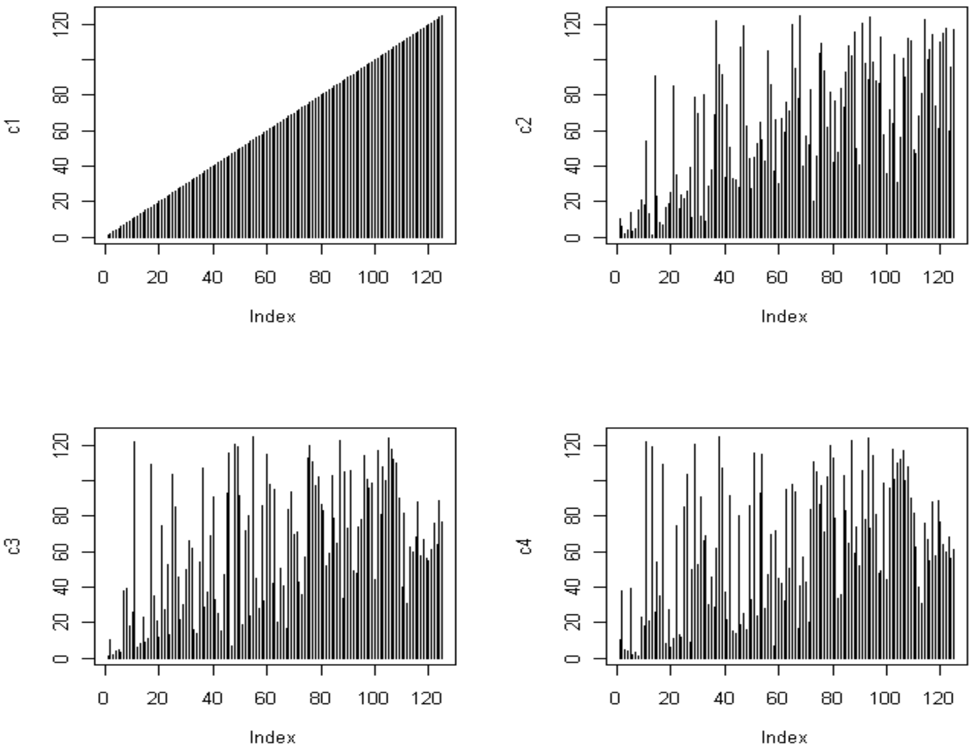

Fig. 2. Indices of documents by similarity (case 2, *f* = 6)

Table 4. Rank order of documents (case 2, *f* = 6)

| $C_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $C_2$ | 10 | 6 | 2 | 4 | 14 | 3 | 5 | 15 | 21 | 18 |
| $C_3$ | 1 | 10 | 2 | 4 | 5 | 3 | 38 | 39 | 18 | 26 |
| $C_4$ | 10 | 38 | 5 | 4 | 39 | 2 | 3 | 1 | 23 | 18 |

Fig. 2 and Table 4 also show the result by P1 and, in this case, the key-term set is composed of 6 terms. These imply that the tendency of results does not so differ if the size of key-term set becomes moderately large.

## 5. Discussion

In addition to the simple approach of term frequency count as a component of term-document matrix, other studies of using weighted values such as TF-IDF or entropy should be investigated further also for the cases described above.

On the SVD and measuring method, the optimal number of the dimension *k* is a common problem. In general, the optimal dimension depends on given data and some ad-hoc approaches have been adopted for each situation. Thus, the suitable number for *k* and also relevant similarity measure for the cases above should be examined with further numerical experiments.

And also a suitable number of elements of the key-term set and number of rules *r* should be explicitly formulated. This problem is concerned with the concept of the principles on association rules in our cases and definition of effectiveness of the rank of documents based on the principles. Thus, it is an open problem to be discussed widely.

As the target documents, in this study a large amount of documents such as exchanged messages in online services or recorded text along with live stream are assumed. They are more concretely, for example, bunch of messages on BBS or SNS, dialog messages in mailing lists, responses of questionnaire survey, transcript of lecture, question and answers of interview and so on. Especially, in recent years, approaches to exhaustive and enormously large data, that is to say big data, have been attracted interest and became active in many other fields. The target documents dealt with in the approach above can be regard as a part of such a comprehensive big data, thus, further investigation for improvements should be done with a view to including relevant methods in the other fields.

And as is the case in the traditional text mining analysis, this kind of approach needs some pre-processing procedures such as stemming and stop-words filtering. The way to define the stop-words, in many cases, differs depends on the given text data or situations of the use of results. In the point of view of dealing such increasing and a large amount of online text data, some automatic mechanism for the pre-production of the appropriate dictionary are desired. Thus, some other application methods such as machine learning based on statistical methodology should be developed to approach the problems. Note that the difficulty of the problems also highly depends on the kind of language notated in the text. For example, in Japanese language, dividing text into each word itself needs some devised process, because of no explicit spacing between words.

The implementation of the approach described above is also an important issue. Extracting features from bunch of text and processing them for helpful information, for example as collective intelligence, has direct connection to the recent network services (Alag 2008, Segaran 2007).

**References**

Agrawal, R, Imielinski, T. and Swami, A. (1993) “Mining association rules between sets of items in large databases,” *Proceedings of the ACM SIGMOD Washington, D.C*, 207-216.

Agrawal, R. and Srikant, R. (1994) “Fast algorithms for mining association rules in large databases,” *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, 487-499.

Alag, S. (2008) *Collective Intelligence in Action*, Manning Publications Co., Greenwich.

Baldi, P., Frasconi, P. and Smyth, P. (2003) *Modeling the Internet and the Web*, JohnWiley & Sons Ltd.

Berry, M. (1992) “Large-Scale Sparse Singular Value Computations.” *International Journal of Supercomputer Applications* 6 (1), 13-49.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990) “Indexing by Latent Semantic Analysis”, *Journal of the American Society for Information Science* 41 (6), 391-407.

Segaran, T. (2007) *Programming Collective Intelligence: Building Smart Web 2.0 Applications*, O’Reilly.

Silverstein, C., Brin, S. and Motwami, R. (1999) “Beyond Market Baskets: Generalizing Association Rules to Dependence Rules.” *Data Mining and Knowledge Discovery* 2 (1), 39-68.