Association Rule Based Characterization of Collective Text Data

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

Abstract

A wide variety of methods using data mining approaches for web data has been proposed and achieved success. Especially in the field of analyzing collective data of text, the approaches and systems which are accumulating documents as collective intelligence along with the extracted features of documents have attracted interest and have been energetically studied. Based on such a background, the aim of this study is to extract essential features from documents and suggest certain documents for contribution to knowledge accumulation.

1. Association Rule

$$A = \begin{array}{c|ccc} & d_1 & d_2 & d_3 \\ \hline \textit{term 1} & 1 & 0 & 2 \\ \textit{term 2} & 0 & 1 & 0 \\ \textit{term 3} & 1 & 2 & 0 \\ \textit{term 4} & 1 & 0 & 3 \end{array}$$

In this study, the collective documents are formulated as a term-document matrix and an association rule based approach is applied in order to extract some characteristic terms throughout the documents. For example, a term-document matrix *A* is composed of the number of terms in each document and each row and column in the matrix represents term and document d_n.

Association rule is a well-know analysis method to find helpful combination of items throughout a large combinatorial space of item sets. In the analysis method, support and confidence are used as the quality measures on a transaction set *T*, like follows,

$$supp(X \Rightarrow Y) \triangleq P(A,B)$$

$$P(A) = \sigma(A)\,/\,N,\ \sigma(A) =\mid \{t_i \mid A \subset t_i, t_i \in T\} \mid,\ \mid\cdot\mid\!: \textit{the number of elements}$$

$$conf(X \Rightarrow Y) \triangleq P(B \mid A)$$

2. Characterization of Documents

The support measure is defined to find frequently appeared combination of items and the confidence measure is defined to find scarcity but strong combination of items. In the context of this study, we consider the support as an approach to obtain common knowledge and also consider the confidence as an approach to obtain precious information. And an essential term set is composed of terms which are extracted by the rule based on the support or confidence principle, as follows,

$$\textit{Essential term set } E \triangleq X \cup Y$$

And the characterization scheme is based on the relationship measured by distance between the essential term set and each documents.

3. Document retrieval

To find particular document which has possibility of contribution to knowledge, a document retrieval method, which is based on latent semantic analysis (LSA), is introduced. A singular value decomposition for a term-document matrix A and its k-dimensionality reduction formula $\hat{A}$ are represented as follows,

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

In the method, a distance between the extracted essential term set and each document is measured by some prescribed criteria, for example cosine values, as follows,

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

where d^j is a row vector (a document) in a right singular matrix V_k, and $\hat{q}$ is a concept space generated by the LSA, as follows,

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

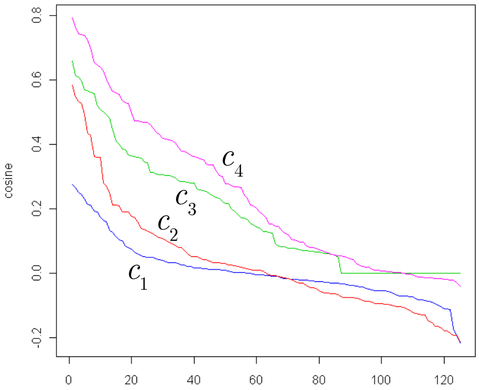

Fig. 1. Change of cosine values

Fig. 1 illustrates the change of cosine values for experimental 120 documents by following similarity measures,

$$c_1 : \hat{q}_1 = q^T U_n S^{-1}, \ V_k = V_n$$
$$c_2 : \hat{q}_2 = q^T U_k S_k^{-1}, \ V_k$$
$$c_3 : \hat{q}_3 = S\hat{q}_1^T, \qquad V_k = (SV_n^T)^T$$
$$c_4 : \hat{q}_4 = S\hat{q}_2^T, \qquad V_k = (SV_k^T)^T$$

and Fig. 2 shows correlation of the cosine values.

In this way, the documents which have higher similarity for the essential terms are selected as a result and the documents which have higher similarity contribute to knowledge accumulation and collective intelligence, eventually.

4. Discussion

In the implementation of systems such as working in corporation with web based text data, computational procedure for the introduced methods should be studied further and especially for the large size data, other statistical approaches also should be taken into account.

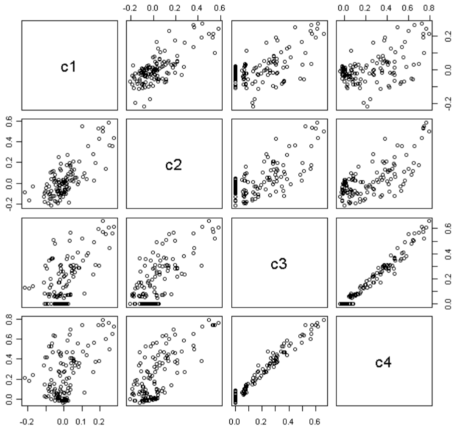

Fig. 2. Correlation of the similarities

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.