3C1-4

内包による外延評価に基づくTop-N形式概念文書クラスタの抽出

A Proposal for Extracting Top N Formal Concepts Based on Quantification

李　愛香[*1]　　原口　誠[*2]
Aixiang Li　　Makoto Haraguchi

北海道大学大学院情報科学研究科
Hokkaido University

In this paper, we propos a method for extracting Top N formal concepts dynamically. In order to search out the aimed concepts more efficiently by a branch and bound algorithm, we try to adjust the order of selecting candidates for extending a branch in the formal concept space. The order is decided by an evaluation of intents. The evaluation is from quantification of features which constitute intents of formal concepts. Quantification value is from coordinates of the eigenvector corresponding to the smallest positive eigenvalue of Laplacian matrix of all features.

1. Background

Currently we are living in the information explosion era. For example, when we search a specific content on the web, millions of associated documents and web pages will be listed up. It is not easy to find out the exact ones for the need from the large scale data set. To solve the problem, document clustering is considered as a standard method. There have been many advanced proposals provided by formers for clustering. In this paper, given a document -term co-occurrence table, instead of clustering all the documents in the space, we only extract top N well -evaluated groups of documents.

On the other hand, the semantic meaning of document cluster is also necessary. The common terms possessed by the documents in one cluster can express the meaning of clusters. Thus a document cluster becomes a dual- cluster (or co-cluster) from both documents and terms. A formal concept (FC) consists of extent and intent. If we consider dual cluster as formal concept, the set of documents can be considered as extent, the set of common terms as intent correspondingly. Thus it is much easier to grasp the meaning of document clusters.

In the formal concept space of the above co-occurrence table, more general FCs (more documents, fewer terms) and more specific ones (fewer documents, more terms)are positioned in two end parts respectively. For those FCs, because their number is relatively small, it is some easy to search out them by using top-down and/or bottom-up mining algorithm. To the contrary, the FCs in middle part are numerous, efficient methods for searching out them are not yet fully developed. For the purpose of extracting Top N FCs in that part, even though an efficient approach has been invented based on clique search in static order in [1], in this paper, we try to design more a new algorithm based on closure calculation in dynamic order.

2. Strategy

2.1 preliminary

Given an document-term co-occurrence table $\mathcal{C} =< \mathcal{D}, \mathcal{T} >$, $\mathcal{T}$ is a set of feature terms and $\mathcal{D}$ a set of documents. Regarding $\mathcal{C}$ as a formal context in FC analysis, a formal concept (D, T) under $\mathcal{C}$ can be viewed as a cluster of documents. Each document in D shares the set of feature terms T and any other document never contains T ; each term in T occurs in all the documents in D and no other term does so. D and T are closed to each other and correspond to the extent and the intent respectively in a FC.

Quality Control: constraint on intents: The fewer terms in intent will result in a general FC, so it is necessary to give a constraint on intent quality. For example, the size of intent must be no less than a given parameter δ , that is $\parallel intent \parallel \geq \delta$. In our algorithm, the quality control will be used to filter candidates of a closure of extent.

Preference in Extents: Among the δ -valid FC-clusters, we prefer ones with higher evaluation values of their extents. Under the constraint of intents, in order to avoid a specific FC clusters, we try to optimize the evaluation of extent. From the view point, a function which behaves monotonically according to expansion of extents is preferred. Simply the function is defined as the size of extent– $\parallel extent \parallel$.

2.2 Problem definition

Input: A document-term co-occurrence table C; δ, a constraint threshold on intent; N (for Top N). Output: Top N δ-valid formal concepts (T, D) in evaluation on extent.

2.3 Algorithm

The main procedure is closure (extent and intent) calculation based on *branch-and-bound depth-first algorithm* with some pruning rules. It is showed in Figure 1. where X_n is a closure of extent in nth stage. Initially we begin it from an empty set of documents with the intent of all the terms. $R(X)$ is the δ- valid candidates set of X. It is showed in formula: $R(X) = \{x \mid \varphi(X \cup x) \geq \delta\}$.

連絡先: 李　愛香，北海道大学大学院情報科学研究科，札幌市北区北１４条西９丁目，TEL:011-706-7575，aixiang@kb.ist.hokudai.ac.jp

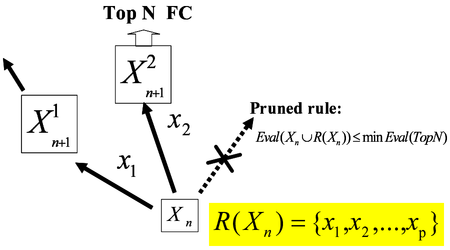
Top N FC
X^2_{n+1}
X^1_{n+1}
x_2
x_1
Pruned rule:
$Eval(X_n \cup R(X_n)) \le \min Eval(TopN)$
X_n
$R(X_n) = \{x_1, x_2, \ldots, x_p\}$

図 1: A branch bound algorithm with pruning rules

Pruning rule1: Any document which has fewer common terms than δ will not be added to $R(X)$ for extending branch. Pruning rule2: During our searching procedure, a Top-N FC list is maintained tentatively according to the evaluation values of extents. When $eval(X) + eval(R(X)$ is less than the minimum evaluation value of FC in Top-N list, the branches of $R(X)$ will be pruned.

From figure 1, it is found that selecting candidate x_2 is more rapidly than selecting x_1 to obtain Top N FCs. Thus there exists a problem: which candidate should be selected firstly?

2.4 Quantification

In order to optimize the evaluation of extent more efficiently, we hope that the corresponding intent has more similar terms. That is we need a indicator number (here, called *score*) to measure the distance of terms and the more similar terms should have more closer scores. In the similarity matrix ($W = C^t \times C$)) of terms, s_{ij} represents the similarity between t_i and t_j, and x_i is the score of t_i. Under the above need, we are aiming to minimize the value $Q = \Sigma_{i,j} s_{ij}(x_i - x_j)^2$ under the condition $\Sigma x_j = 0$ and $\Sigma x_j^2 = n$ (n is the number of all terms in C). In [5], it is proved that x_i is just the coordinate of the eigenvector corresponding to the smallest positive eigenvalue of Laplacian matrix (L=D-W, D: degree matrix of terms) of terms.

After terms are assigned scores, the candidates $x \in R(X)$ are sorted in ascending order by the following evaluation values at every stage before calculating the next closure. We call it dynamic order.

$$e(x) = \sum_{f \in \varphi X \cap \varphi x} (score(f) - \overline{score(\varphi X \cap \varphi x)})^2 \qquad (1)$$

The evaluation reveals the closeness among the next intent when x is selected. At last, the above algorithm will be improved from the originally fixed order to the dynamic order for extending branch. As a result, the higher quality FCs are obtained more efficiently.

3. Experiment

The experiment data is a collection of 1000 short files and 500 words in them. The co-occurence table is represented in boolean cell values. This time we only give a comparation of computation time among the original degree static order, document-score dynamic order and term-score dynamic order in the same search engine we created. Experiment environments : OS: windows XP CPU: pentium4 2.8G RAM: 1.0G Language: java. Experiment results are showed in figure 2 with $N = 2$, $\delta = 25, 20, 15, 10$.

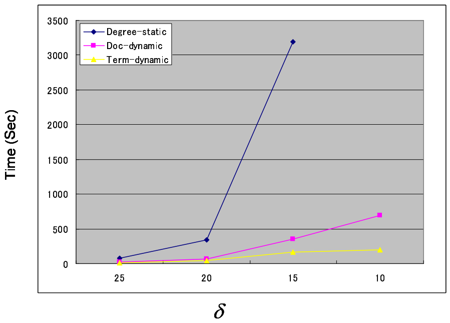

図 2: the comparison of dynamic orders and static order

4. Discussion and conclusion

From the results, it is showed that the dynamic order based on term scores performs better than the static order on degrees, even than the dynamic order based on document scores. Especially when δ becomes small, it means that the aimed FCs are positioned in more hard part in the FC lattice, the dynamic order based on term scores keeps steady state in ascending trend, while the static order even breaks down.

This paper provides a heuristic method for searching Top N FCs from a co-occurrence data which is not limited in a document-term table. And there is duality between two dimensions in the method. Moreover it is indicated that the similarity among terms of intent is an important factor for clustering documents. This time we only use it as a measure for selecting candidates of closure, in the next time, we will analyze it in detail before searching FCs. Even though we have not experimented more kinds of data, it provides a valuable way to apply quantification to descrete objects for a data analysis

参考文献

[1] M.Haraguchi and Y.Okubo, An Extended Branch and Bound Search Algorithm for Finding Top-N Formal Concepts of Documents, Springer-LNCS 4384, pp.276-288, 2007.

[2] Y.Okubo and M.Haraguchi, Finding Conceptual Document Clusters with Improved Top-N Formal Concept Search, Proceeding of IEEE/WIC/ACM WI-2006, pp.347-351, 2006.

[3] Ulrike von Luxburg, A Tutorial on Spectral Clustering, Technical Report No. TR-149, Max Planck Institute for Biological Cybernetics, 2006.

[4] B.Ganter and R.Wille, Formal Concept Analysis - Methematical Foundations, Springer, 1999.

[5] 水野　欽司, 多変量データ解析講義, 朝倉書店, 1996.