A Clustering Method for News Articles Retrieval System

Hiroyuki Toda
NTT Cyber Solutions Laboratories, NTT Corporation
1-1 Hikarinooka Yokosuka-Shi
Kanagawa, 239-0847 Japan
toda.hiroyuki@lab.ntt.co.jp

Ryoji Kataoka
NTT Cyber Solutions Laboratories, NTT Corporation
1-1 Hikarinooka Yokosuka-Shi
Kanagawa, 239-0847 Japan
kataoka.ryoji@lab.ntt.co.jp

ABSTRACT

Organizing the results of a search facilitates the user in overviewing the information returned. We regard the clustering task as the tasks of making labels for a list of items and we focus on news articles and propose a clustering method that uses named entity extraction.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process, Clustering*

General Terms

Algolithms, Experimentation

Keywords

Search result organization, Document clustering, Named entitiy

1. INTRODUCTION

Services that provide news articles collected from various publishers have become popular. In these services, the latest articles are shown on the top page and allow the user to search for news stored for some period. However, since these services collect the articles from many publishers and since so new news articles are being continuously published, the amount of these contents is becoming so vast that the user can not efficiently find the desired articles.

To solve this problem, Google News[1], groups the news articles that discuss the same news event as indicated by the similarity of the news titles. This can reduce the number of news articles shown to the user. However, this approach presents a flat list of the search results to the user, which is not the most efficient tactic.

Clusty[2] extracts the terms (words or phrases) that seem to be important in the document lists (search results or latest news lists), as cluster label candidates. The documents are assigned to the appropriate labels to form clusters. The label to which a document is assigned is the label present in the document, As a result, the user can easily overview the list and find the topics that exist there. However, this method has some problems with label creation. One is label quality. Because the appropriate positions may not be extracted, inappropriate terms are selected as labels. Besides, the bases of the labels are unclear and various kinds of labels are often mixed. This makes it awkward for the user to overview a list. Many researchers have recently tackled search result clustering, but the fundamental problem of label readability has not be well solved[1][2].

In this paper, we focus on news articles and propose a search result(or latest news) clustering method that is based on named entity extraction.

[1] http://news.google.com/
[2] http://www.clusty.com/

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

2. PROPOSED METHOD

2.1 Named entity extraction

The named entity(NE) extraction task was born in MUC[3] of the 1990's. It is the task of extracting the information units important to recognition like names, including people, organizations, and location names, and numeric expressions including time, date, money and percent expressions[5]. Accordingly, we consider that NE extraction is the best way to identify label candidates in news articles. In this paper, we use the proper noun expressions among the NE as label candidates. We use Isozaki's NE extraction tool[4].

2.2 Algorithm

Our algorithm of making labels for documents list is shown here. We consider that all news articles are first registered with our system. Next, the terms (named entities) are extracted in pre-processing. When our system accepts a query, the system uses the following algorithm.

1. Fetching documents list
2. Listing the terms in the documents list
3. Selecting the labels from listed terms
4. Organizing the labels by NE category

At first, we fetch the documents list, for example, search results or latest news articles. Second, we list the terms that are extracted from the documents in the documents list. Each term has its category information.

In the third process, we first calculate the score of each term using label selecting criterion that based on two ideas "terms that are useful in overviewing the results and efficiently locating the desired documents will not be too rare or too common" and "terms that are related to the query are useful as labels". The criterion is represented by the following equation.

$$Criterion = DF_{R,i} \times \log(\frac{|R|}{DF_{R,i}}) \times \frac{DF_{R,i}/|R|}{DF_{D,i}/|D|}$$

$DF_{R,i}$ is document frequency of term i in the search result R. $DF_{D,i}$ is document frequency of term i in document collection D. The terms that have high score are selected

Figure 1: Examples of labels (Query: "President Bush")

as labels. More precisely, the labels that construct similar clusters are combined using the similarity of clusters and labels but we will discuss this point in another paper. In organizing the labels process, the labels are organized by the categories given by NE extraction, which allows the user to easily overview the search results. Furthermore, to ensure that the index provides not only easy overviewing but also efficient document location, we define a category-ranking criterion. The criterion is based on three considerations.

- *Clearness of category:* Paucity of overlapping documents between each label. It is calculated as the ratio between distinct number and total number of the documents related to the labels.
- *Equality of category size:* Equality of the number of the documents related to each label. It is calculated by average entropy.
- *Exhaustiveness of category:* The percentage of the documents in the search result that can be labeled. It is calculated from the ratio between the number of search results and the number of the labeled documents.

In this paper, we integrate these criteria to realize category ranking. Examples are shown in Figure 1.

3. EVALUATION

In this section, we evaluate the above label selecting criterion and the efficiency of label organization.

In the evaluation of label selecting criterion, we used the Japanese newspaper collection of IREX[6]. The collection covers 2 years(1994 and 1995) and holds about 200,000 articles. We used 30 search topics and the relevance judgment data for the topics as defined by IREX. Queries to the system were constructed using DESCRIPTION[3]. We also created the "Useful Keywords List", a collection of useful keywords to overview the relevant documents for each topic[4].

We consider that the label selecting criterion should be estimated from two viewpoints. One is whether users select the labels. This is related to label readability and meaning.

[3]DESCRIPTION is defined for each topic; each expression consists of 2 or 3 nouns. We extract the meaningful terms from DESCRIPTION and link them by the "or" operator.

[4]To make this list, we showed each set of relevant documents and a topic to 5 subjects, and each subject selected the most suitable keywords. Keywords that were selected by more than 3 subjects were added to the keyword list. The result was a "Useful Keywords List" for each topic.

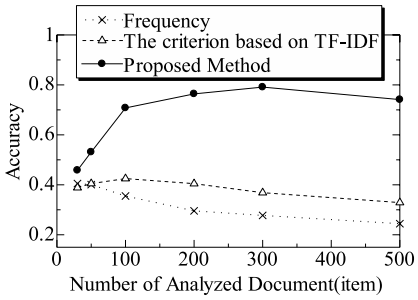

Figure 2: Evaluation Result

The other is whether the labels are related to the relevant documents. Accordingly, our evaluation basically examined the accuracy of the search results using the labels that the user could be assumed to select. The labels that also existed in "Useful Keywords List" are assumed to be selected. The accuracy is calculated by the average precision of the top 10 results when the labels are selected.

The results are shown in Figure 2. The proposed method has much higher accuracy than the ordinary method. We can also see that if we analyze about 100 documents and use the proposed criterion to select the labels, the resulting labels yield high accuracy (more than 70%).

On the other hand, we also evaluated the efficiency of the label organization method. To evaluate this, we constructed a news search system for Japanese latest news that can switch between different label generation methods. One is the proposed method, which organizes the labels by category; the other is the ordinary method which simply lists the labels. Subjects were given a simple questionnaire that asked "which method is more useful?" The result of this questionnaire was that 85.6%(113/132) said that the proposed method was more useful than the ordinary method.

4. CONCLUSION

In this paper, we regard the clustering task as the tasks of making labels for a list of items and we focus on news articles and propose a clustering method that uses NE extraction. We proposed a label selecting criterion and a label organization method. Evaluations indicated that the proposed methods are more useful than the current methods. Our evaluation used only Japanese newspaper articles, but our method is not language specific and so it could be used to handle other languages.

5. REFERENCES

[1] Zeng, H. J., He, Q. C., Chen, Z., Ma, W. Y. and Ma, J.: "Learning to Cluster Web Search Results." Proceedings of SIGIR'04, pp.210-217, 2004.

[2] Kummamuru, K., Lotlikar, R., Roy, S., Signal, K. and Krishnapuram, R.: "A hierarchical monothetic document clustering algorithm for summarization and browsing search results." Proceedings of WWW'04, pp.658-665, 2004.

[3] Grishman, R. and Sundheim B.: "Message Understanding Conference - 6: A Brief History." Proceedings of COLING'96, pp.466-471, 1996.

[4] Isozaki, H. and Kazawa, H.: "Efficient Support Vector Classifiers for Named Entity Recognition." Proceedings of COLING'02, pp390-396 , 2002.

[5] Sekine, S.: "Named Entity: History and Future." http://cs.nyu.edu/ sekine/papers/NEsurvey200402.pdf, 2004.

[6] Sekine, S. and Isahara, H.: "IREX Project Overview." Proceedings of the IREX Workshop, 1999.