# Search Result Clustering Method at NTCIR-5 WEB Query Term Expansion Subtask

Hiroyuki Toda Ryoji Kataoka
NTT Cyber Solutions Laboratories
NTT Corporation
1-1 Hikarinooka, Yokosuka-shi, Kanagawa 239-0847, Japan
{toda.hiroyuki, kataoka.ryoji}@lab.ntt.co.jp

## Abstract

*We use a retrieval system with search result clustering to tackle the NTCIR-5 WEB Query Term Expansion Subtask. The system clusters the search results in such a way as to make it easier for the user to select relevant documents as feedback documents. In addition, we select phrase words or named entities(NE) as query-expansion keywords from the feedback documents because these words tend to represent the characteristics of feedback documents and can retrieve relevant documents that were not retrieved by the initial keywords. Based on our evaluations, we report the efficiency of keyword expansion and the number of relevant documents in the feedback documents.*
**Keywords:** *query expansion, search result clustering, named entity recognition*

## 1 Introduction

This paper describes a retrieval system with search result clustering that we created to tackle the NTCIR-5 WEB Query Term Expansion Subtask. This system clusters the search results and the clusters make it easier for the user to select relevant documents for feedback. From the feedback documents we select phrase words or NEs as query-expansion keywords because they tend to well represent the characteristics of feedback documents and can retrieve relevant documents that may be overlooked by the initial keywords.

We describe our evaluation of the retrieval system. The evaluation elucidated the efficiency of keyword expansion, as well as the relevancy of documents in the feedback documents. We show the evaluation result of not only feedback-type="user" but also feedback-type="auto". We also consider cluster based feedback by examining the evaluation results.

The paper is organized as follows. The next section introduces our search result clustering method. In Section 3, we describe the system created to tackle the NTCIR-5 Query Term Expansion Subtask. The evaluation of the proposed method is shown in Section 4. Finally, we conclude the paper in Section 5.

## 2 Our search result clustering method

In this section, we introduce the search result clustering method used.

### 2.1 variation of search result clustering

Many approaches are being researched for organizing the search results to improve efficiency of searching. There are two main approaches: the document-based approach and the label-based approach.

Instances of the document-based approach include the many methods that employ document clustering. Such methods cluster the documents using the similarity of features such as the keyword vectors of documents. After that, they extract representative term(s) or sentence(s) as labels from each cluster, which are then presented to the user with the search result. Scatter/Gather[1] and the method of Leuski[2] adopt this approach.

These document-based approaches usually yield non - overlapped clusters and label quality is influenced by the accuracy of clustering. Though the number of clusters or similarity threshold generally controls clustering, it is difficult to select the value that suits the user's intention. As a result, labels are often unreadable, so it is difficult to adopt this approach in search engine applications.

The label-based approach, on the other hand, first extracts informative terms(words or phrases) from the search result as labels using a statistical analysis of components such as the frequency of appearance or the properties of clusters formalized by the terms(The cluster is formalized by the documents that include the term). The labels are presented to the user with the search result.

Zeng et al.[3] proposed a label based approach that uses phrases scored by a combination of some properties of labels and document clusters formalized by the labels. Kummamuru et al.[4] regard the label list as a taxonomy of the search result, and proposed a label selecting criterion based on taxonomy likelihood.

The methods of this approache don't have the problem of threshold and labeling, and so are the main approach used for search result clustering.

### 2.2 characteristics of our search result clustering method

Our method is a label-based approach[5]. We consider that proper nouns are important for characterizing documents, so, we use NEs as labels.

### 2.3 Algorithm

The algorithms that creates label list of the search result is shown here. We consider that all documents are first registered with our system. Next, the terms(NEs) are extracted in pre-processing. When our system accepts a query, the system uses the following algorithm.

1. Fetch search result
2. List the NEs in the search result
3. Select the labels from listed NEs
4. Organize the labels by NE category

At first, we fetch the search result. Second, we list the terms(NEs) that are extracted from the documents in the documents list. Each NE is pretagged with category information.

In the third process, we first calculate the score of each NE using the label selecting criterion. 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 as labels.

The terms that have high score according to this criterion are then selected as labels. More precisely, the labels that construct similar clusters are combined using the similarity of clusters and labels. In organizing the labeling process, the labels are organized by the categories given by NE extraction.

## 3 System

In this section, we explain the system that implements our proposal.

### 3.1 Methods of feedback document

We use the search system provided by the organizer[6] and we cluster the search result by our clustering method. We regard the clustered results as the output of the initial search system, and select the documents for feedback from these results.

The method of selecting feedback documents is as follows. If the top ranked cluster was related to more than $n$ documents, we select the documents ranked lower than the $n$-th ranking document. If the top ranked cluster was related to fewer than $n$ documents, we use the $n$ documents gathered from the ranked clusters in decreasing order.

We discuss here the results gained using both user-indicated (user) and automatic (auto) feedback. When feedback-type is "user", to simulate the user, each cluster score follows the number of relevant documents in each cluster. This is because we assume that the user selects clusters according to their relevancy. This consideration is based on an evaluation of clustered search results at "NTCIR-4 Web D"[7]. Here, we continue to select feedback documents until $n = 20$ or the number of relevant document is $5$. On the other hand, when feedback type is "auto", cluster rank follows the score of the label of the cluster given by our clustering system. Here, we set $n = 20$.

### 3.2 Selection of expansion-keywords

To select the expansion keywords, we first extract keywords from feedback documents as expansion keyword candidates. Next, we score each keyword and the top 10 keywords, highest scores, are selected as expansion keywords.

When feebback-type is "user".

$$Score_j = \sum_{\{X|X=S,A,B,C\}} (w_X \times \sum_{d_i \in D_X} t_{i,j})$$

When feebback-type is "auto".

$$Score_j = \sum_{d_i \in D} t_{i,j}$$

$D$ is a feedback document collection. $D_X$ is a feedback document collection whose relevance judgment is $X$. $w_X$ is weight for the document whose relevance judgment is $X$. Here, weights are set to $w_S = 1.5, w_A = 1.0, w_B = 0.5, w_C = -1.0$. $d_i$ is a document in feedback document collection $D$. $t_{i,j}$ is tf-idf weight of word $w_j$ in document $d_i$. This is calculated by the following equation.

$$t_{i,j} = \log(1 + tf_{i,j}) \times \frac{N}{df_j}$$

$N$ is the document collection registered in the search system. $tf_{i,j}$ is term frequency of term $w_j$ in

**Table 2. Comparison of proposed system to baseline system(feedback-type="auto")**

| system | Rel_ret | Average Precision | P@5 | P@20 | R-Precision |
|---|---|---|---|---|---|
| baseline | 2256 | 0.1511 | 0.2914 | 0.2529 | 0.1991 |
| auto-ph | 2432 | 0.1708 | 0.3543 | 0.2829 | 0.2107 |
| auto-ne | 2322 | 0.1536 | 0.3314 | 0.2514 | 0.1962 |

the document $d_i$. $df_j$ is document frequency of term $w_j$ in $N$.

In the search process using expansion keywords, initial keywords and expansion keywords are weighted by bm25 using the feedback documents. In the evaluation, the search results generated by organizer's system with weighted initial keyword are regarded as baseline results, and the search results generated by organizer's system with weighted initial keywords and weighted expansion keywords are regarded as our final results.

We try to use two kinds of keyword(phrase word or NE) as expansion keywords. Accordingly, our final results are as follows.[1]

- user-ph
  feedback-type="user", keyword="noun phrase"
- user-ne
  feedback-type="user", keyword="NE"
- auto-ph
  feedback-type="auto", keyword="noun phrase"
- auto-ne
  feedback-type="auto", keyword="NE"

Note that the feedback documents are the same when the feedback-type is the same.

## 4 Evaluation

### 4.1 Benefits of expansion keywords

The benefits of keyword expansion are confirmed by the values of each search result.

**Table 1. Comparison of proposed system to baseline system(feedback-type="user")**

| system | P@100 | P@50 | P@20 | P@5 |
|---|---|---|---|---|
| baseline | 0.2886 | 0.3121 | 0.34 | 0.36 |
| user-ph | 0.3384 | 0.3846 | 0.4271 | 0.5371 |
| user-ne | 0.3138 | 0.3624 | 0.4487 | 0.5349 |

[1] Our submitted results of feedback-type="user" have a mistake. Therefore, the revised results were evaluated by organizer after the Formal Run Evaluation Results Release.

First, we show some search results gained when feedback-type is "user" in Table 1. These values are the averages gained from of 30 search topics. These values are the precision achieved when we select the top $n$ ranked documents; the feed back documents are eliminated from the search result. We consider that when the user manually selects the feedback documents, we should evaluate the accuracy after the user selected documents are eliminated. This result shows that both feedback systems yield higher precision than the baseline system. Additionally, the precision of user-NE decreases in P@50 and P@100. The reason for this is considered to be that the NEs are too specific to retrieve many relevant documents.

Next, we show some of the results gained when the feedback-type is "auto" in Table 2. The system with expanded keywords generally yields higher scores than the baseline system, and the system that uses noun phrases as expanded keywords yields higher values than the system with NEs. This is the same tendency as shown with user-NE. This tendency is confirmed by the observation that P@5(Precision at top 5 results) of the NE expanded system is a little higher than the baseline system but other values, which required the retrieval of many relevant documents to achieve high score, are not much higher than the baseline equivalents.

### 4.2 Correlation between quality of feedback documents and accuracy of search result

Table 3 shows correlation coefficient between quality of feedback documents and accuracy of search result. According to this result, the correlation between recall and quality of feedback documents is low. On the other hand, the correlation between precision and quality of feedback documents is high. Furthermore, The correlation between R-precision and quality of feedback documents is higher than correlation with averaged precision. This indicates that the quality of feedback documents affects the precision of higher ranked documents than that of lower one. In addition, the correlation between search accuracy and ratio of S and A documents in feedback documents are higher than correlation with S,A and C documents.

**Table 3. Correlation coefficient between quality of feedback documents and accuracy of search result**

| feedback type | expansion keyword type | relevance document type | Recall | Average Precision | R-Precision |
|---|---|---|---|---|---|
| auto | phrase | S,A | 0.3557 | 0.7924 | 0.8195 |
| auto | NE | S,A | 0.2955 | 0.7972 | 0.7849 |
| auto | phrase | S,A,B | 0.0777 | 0.6385 | 0.6768 |
| auto | NE | S,A,B | 0.0017 | 0.6486 | 0.6617 |
| user | phrase | S,A | 0.2189 | 0.6733 | 0.7312 |
| user | NE | S,A | 0.2300 | 0.5584 | 0.7061 |
| user | phrase | S,A,B | -0.1775 | 0.4842 | 0.6140 |
| user | NE | S,A,B | -0.1820 | 0.4452 | 0.7411 |

## 4.3 Analyzing result of feedback documents

Because we considered that search result clustering can efficiently select relevant documents as feedback documents when the user manually selects feedback documents, we created a search result clustering system. To verify the assumption made, we evaluated the relevancy of the feedback documents between the system with search result clustering (proposed system) and that without search result clustering(ordinary system). We separately evaluated the result of feedback-type="user" and that of feedback-type="auto". When we evaluated the results of feedback-type="auto", we also considered cluster based feedback.

### 4.3.1 In the case of feedback-type="user"

Table 4 shows the relevance judgments of feedback documents. These values are the averages gained from 30 search topics. These results indicate that the proposed method selects more relevant documents than the ordinary system. Furthermore, the proposed method has fewer irrelevant feedback documents than the ordinary system. This suggests that the proposed method reduces the cost of feedback-document selection.

**Table 4. Number of relevant documents in feedback documents(feedback-type="user")**

| system name | S+A | S+A+B | C |
|---|---|---|---|
| ordinary | 104 | 132 | 387 |
| proposed(user) | 108 | 146 | 292 |

### 4.3.2 In case of feedback-type="auto"

Table 5 and 6 show the relevance documents ratios of feedback documents of the ordinary system and the proposed system, respectively. These values are the averages gained from 30 search topics. Though the results shown were gained through the use of 20 feedback documents, we show the relevance judgment result for feedback document quantities of 1,3,5,10, and 20 for reference.

**Table 5. Ratio of relevant documents in feedback documents of ordinary system (feedbak-type="auto"; without search result clustering)**

| feedback documents | ratio of S+A | ratio of S+A+B |
|---|---|---|
| 1 | 0.4286 | 0.4857 |
| 3 | 0.3810 | 0.4381 |
| 5 | 0.3429 | 0.4 |
| 10 | 0.3086 | 0.3943 |
| 20 | 0.2629 | 0.3486 |

**Table 6. Ratio of relevant documents in feedback documents of proposed system (feedbak-type="auto"; with search result clustering)**

| feedback documents | ratio of S+A | ratio of S+A+B |
|---|---|---|
| 1 | 0.4 | 0.5429 |
| 3 | 0.3429 | 0.4667 |
| 5 | 0.3257 | 0.4286 |
| 10 | 0.2829 | 0.3971 |
| 20 | 0.2414 | 0.3429 |

When we regard the documents judged to be S, A, B as relevant documents, the proposed system outputs more relevant documents than the ordinary system. However, when we regard the documents judged to be S, A as relevant documents, the reverse is true. The reason for this difference is not clear. These results indicate that our cluster scoring does not give adequate priority to relevant documents.

**Table 7. Ratio of relevant documents in feedback documents of the system that simulates the user cluster selection behavior(with search result clustering)**

| feedback documents | ratio of S+A | ratio of S+A+B |
|---|---|---|
| 1 | 0.5429 | 0.6286 |
| 3 | 0.4857 | 0.5905 |
| 5 | 0.48 | 0.5771 |
| 10 | 0.3639 | 0.4957 |
| 20 | 0.3164 | 0.4282 |

On the other hand, Table 7 shows the result gained when we simulated cluster selection by the user. In this case, the relevance ratio of feedback documents was about 10 points higher than that of the ordinary system. This result indicates that the relevance of feedback documents is improved, when the user simply selects the clusters without detail evaluation of each document.

Some examples of cluster labels are shown in Table 8. These clusters seem to be easy for the user to select. This indicates the desirability of the user using only cluster labels to select clusters since this approach yields many more relevant documents as feedback documents.

**Table 8. Example of cluster labels which have large number of relevance documents**

| initial keywords | cluster label |
|---|---|
| オフサイド, サッカー, ルール | オフサイドトラップ |
| クリムト, 美術館 | ウィーン |
| マイケル・ジョーダン, バスケットボール | NBA |
| 点数制度, 運転免許 | 道路交通法 |
| 江戸時代, 建築物, 東京 | 江戸東京博物館 |

## 5 Conclusion

We reported the results gained in tackling the NTCIR-5 WEB Query Term Expansion Subtask. We used a search result clustering system. A comparison with a baseline method indicated that our query expansion approach is useful with both types of feedback: user-driven and automatic. However, query expansion using NE is less useful than that using noun phrases. We also show that the quality of feedback documents affects the precision of higher ranked documents in our method. An evaluation of feedback documents showed that our clustering method selects many more relevant documents as feedback documents when the feedback type is "user". Furthermore, we showed that the proposed method reduces the cost of feedback-document selection when the feedback type is "user". We also show that desirability of the user using only cluster labels to select feedback documents.

Though our clustering system selects similar number of relevant documents as feedback documents with the system without clustering, the precision after feedback is not better than the organizer's baseline. One reason seems to be quality of feedback documents. In the feedback document selection process of organizer system, there are some heuristics to eliminate useless relevant documents. These heuristics might be important to yield higher search accuracy.

My future works are to examine the above problem and evaluate the interrelation between feedback keyword type(word or phrase or named entity) and search accuracy.

## References

[1] Hearst, M., and Pedersen, J.: "Reexamining the cluster hypothesis: scatter/gather on retrieval results." *Proceedings of SIGIR'96, pp.76-84,* 1996.

[2] Leuski, A.: "Evaluating Document Clustering for Interactive Information Retrieval." *Proceedings of CIKM'01, pp.33-40,* 2001.

[3] 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.

[4] 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.

[5] Toda, H. and Kataoka, R.: "A Search Result Clustering Method using Informatively Named Entities." *Proceedings of WIDM'05, pp.81-86,* 2005.

[6] Yoshioka, M. and Haraguchi, M.: "On a combination of probabilistic and boolean ir models for www document retrieval." *ACM Transactions on Asian Language Information Processing(TALIP), 4* 2005.

[7] Eguchi, K.: "Overview of the Topical Classification Task at NTCIR-4 WEB." *Working Notes of NTCIR-4 Vol. Supl. 1, pp.ov-48-ov-55,* 2004.