International Journal of Intelligent Information Systems
2014; 3(1): 1-7
Published online February 20, 2014 (http://www.sciencepublishinggroup.com/j/ijiis)
doi: 10.11648/j.ijiis.20140301.11

An effective cluster-aware labeling method for web search results using concordant document frequencies

Masafumi Matsuhara[1], Toshihiro Yoshida[2]

[1]Department of Software and Information Science, Iwate Prefectural University, Iwate, Japan
[2]NTT Advanced Technology Corporation, Kanagawa, Japan

Email address:
masafumi@iwate-pu.ac.jp (M. Matsuhara)

To cite this article:
Masafumi Matsuhara, Toshihiro Yoshida. An Effective Cluster-Aware Labeling Method for Web Search Results Using Concordant Document Frequencies. *International Journal of Intelligent Information Systems.* Vol. 3, No. 1, 2014, pp. 1-7.
doi: 10.11648/j.ijiis.20140301.11

Abstract: In recent years, the amount of information on World Wide Web has exploded. Search engines are generally used for web searching; however, robot-type search engines have a few problems. One such problem is that it is difficult for a user to come up with an appropriate query for obtaining the search results she/he intends. Moreover, it is difficult for users to understand the contents of search results because a robot-type search engine outputs many search results in a long list format. To solve these problems, many methods have been proposed that classify the results of a robot-type search engine into clusters that are labeled and then shown to the user. To be effective, the cluster label needs to consist of appropriate words to describe the web sites within the cluster. In this study, we propose a labeling method using concordant document frequencies where the web search results of a query are classified into clusters and we use our techniques to assign the proper labels to those clusters. We then find the set of web sites that result from an AND-query using an original query word and the cluster label. If this set and the members of the cluster are common, we say that the concordant document frequency is high, and the cluster label is assigned a high weight. Thus, it is possible to assign an appropriate label using our proposed cluster-aware method. We demonstrate the effectiveness of our proposed method by simulation experiments.

Keywords: Labeling, Clustering; Web Search

1. Introduction

The amount of information on the World Wide Web is increasing rapidly. For web searching, most users use robot-type search engines, e.g., Google[1], Yahoo![2], and so on. However, the search results are often not appropriate and the list is too long for a user to manually evaluate all results. To solve these problems, many methods have been proposed.

Some of the proposed methods classify the results from a robot-type search engine into clusters that are labeled. The cluster labels are then presented to users. There are systems like Carrot[1][2] and Yippy[3] and so on. Users are easily able to understand the contents of web search results by browsing through cluster labels. If users are not able to think of an appropriate query word for obtaining the results they need, they can also use a vague query and evaluate the resulting cluster labels to decide on the proper keyword for their search.

In general, a significantly representative word should be used as a cluster label. If cluster labels do not truly represent the contents of its members, then users cannot access the contents even though they may be relevant. However, if the accuracy of the labels is improved, it is possible for users to more easily find information they need. Many proposed systems are based on the Term Frequency Inverse Document Frequency(TFIDF),which is used for calculating word significance.TFIDF gives a high weight to a word in a document if that word occurs frequently in only a few documents, indicating that the word is very significant for those documents. However, a word that is common in the documents of a cluster may be appropriate for the label of that cluster, even though the TFIDF of that word is of low weight. Therefore, we believe that the cluster label should not be determined using TFIDF alone.

The proposed concordant document frequency[4] can be calculated for each cluster. We have proposed a labeling method using concordant document frequencies where the

[1]http://www.google.com/
[2]http://www.yahoo.com/

web search results of a query are classified into clusters. We then find the set of web sites that result from an AND-query using an original query word and the cluster label. If this set and the members of the cluster are common, we say that the concordant document frequency is high, and the cluster label is assigned a high weight. Using this proposed cluster-aware labeling method, it is possible to assign the proper labels for each cluster.

We use a system of Japanese morphological analysis to divide Japanese sentences into words in our proposed method; however, it cannot always properly divide compound words. In Japanese, many long compound words are formed by combining similar words with completely different meanings, so a splitting compound word would change its meaning altogether. Compound words are important representatives of a document, so we use both words and compound words as cluster labels.

In the following, all the examples use Japanese words and the corresponding English words are shown in brackets. We performed all experiments using Japanese words, and if similar experiments were to be done using the corresponding English words, similar results would not be obtained, hence, to clearly show our specific results, we included the Japanese words as well.

The rest of the paper is organized as follows. Related works are introduced in Section 2and our proposed method is shown in Section 3.Experiments and results are discussed in Section 4and Section 5 concludes the paper.

2. Related Works and our Approach

2.1. Related Works

Many methods have been proposed for clustering web search results. For instance, Scatter/Gather[5] was one of the first web-clustering applications, and clustering method was proposed based on the Clustering By Committees (CBC) algorithm[6].A clustering and exploring method was also proposed in [7]that uses the temporal information associated with documents, clustering and presenting documents along timelines.

Various search interfaces have also been proposed, such as a visual search interface[8], a dynamic clustering interface[9]-[11], and so on. The dynamic clustering interface uses the Suffix Tree Clustering (STC) algorithm, a linear time clustering algorithm based on identifying phrases that are common to groups of documents.

The cluster labeling methods that have been proposed includea cluster labeling method using Wikipedia[12].In this method, label candidates are extracted from Wikipedia in addition to important terms that are extracted directly from text. A hierarchical clustering method has also been proposed by [13]that consists of hierarchical clustering, labeling, and personalized search results. Lingo[14] is another method based on singular value decomposition. In this method, over 75[%] cluster labels were marked as useful.

2.2. Our Approach

The labeling methods are generally based on the clustering methods. The accuracy of a search interface is affected by the adopted labeling method. Most proposed labeling methods are not for Japanese. In Japanese, compound words are important representatives of a document, so we need to use both words and compound words as cluster labels.

Most robot-type search engines correspond to Japanese although most proposed labeling methods are not for Japanese. Therefore, we use a robot-type search engine for labeling clusters. We use web search results alone in order to use any search engine.

Although many clustering algorithms have been proposed and we are able to use any clustering algorithm, the labeling accuracy is affected by the clustering algorithm. When the accuracy of the clustering algorithm is high, the labeling accuracy of our proposed method is likely to be high as well. However, the effectiveness of our proposed concordant document frequency would not be clear because the concordant document frequencies are highly dependent on the clustering results. For this reason, we use the classical k-means clustering as a baseline algorithm. The situation is similar for TFIDF.

3. Our Proposed Labeling Method

3.1. Outline

The steps of the labeling process are shown in Figure 1.

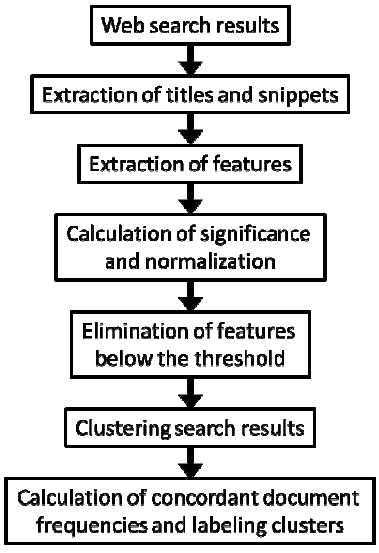

Figure 1. *The steps of the labeling process.*

The search results are obtained using a query word "q" inputted by a user. The titles and snippets of the retrieved documents are obtained from the search results and divided into words by a system of Japanese morphological analysis. Compound words are reconstructed using the results of a morphological analysis system. The system, based on our

proposed method, calculates the word weights that are then used as the feature vectors of the documents. The documents are then clustered based on the word weights and candidate labels are created using the document words. The system calculates the concordant document frequencies of the candidate labels and the most appropriate label is given to each cluster.

3.2. Extraction of Titles and Snippets

Our proposed system obtains web search results for query "q" using a robot-type search engine, and the titles and snippets are extracted from the web search results.

There are two possibilities for document word extraction: to use the full text of document or to use a title and snippet, i.e., a meaningful phrase of the original document. A snippet does not contain much of the "noise" present in the original document that might cause misclassification of the document[10],so, for this reason, we use titles and snippets.

3.3. Extraction of Features

In this process, words are extracted by a morphological analysis system and any compound words are reconstructed from these words. Both words and compound words are used as features in our proposed method.

English and other European languages are written with spaces between words; however, Japanese morphological analysis is difficult because Japanese words are written with no space between words. Therefore, we use ChaSen[3],a Japanese morphological analysis system that is able to divide Japanese text into separate words.

Nouns are extracted from the words divided by ChaSenand used as features for clustering the total set of search results. ChaSen cannot extract the long compound words that can be important features for a document; these are divided as component words by a system of Japanese morphological analysis.

Table 1.*Examples of split of compound words.*

Compound words	The result of morphological analysis
株式会社(Corporation)	株式(Shares)，会社(Company)
心理学(Psychology)	心理(Mentality)，学(Learning)

Examples are shown in Table 1.For instance, the word "心理学(Psychology)" is a compound word that has been split into two words "心理(Mentality)" and "学(Learning)" by morphological analysis. The separated words do not mean the same as the whole word "心理学(Psychology)".In this way, compound words are split into words which often have a different meaning, but it is important to keep them together to retain the original meaning.

Our proposed method uses TermExtract[4], which is able to reconstruct compound words using a result of ChaSen.

3.4. Calculation of Significance and Normalization

Feature weights are calculated for clustering web documents using TFIDF[15].The feature weight $tfidf$ is calculated by

$$tfidf(t,d) = tf(t,d) \cdot idf(t)$$
$$= tf(t,d) \cdot \left(\log\frac{N}{df(t)} + 1\right)$$

where the term frequency $tf(t,d)$ is the number of times the feature t occurs in a document d, and the document frequency $df(t)$ is the number of documents in which the feature t occurs at least once.The inverse document frequency $idf(t)$ can be calculated from the document frequency $df(t)$, and$N$ is the total number of documents.

The inverse document frequency of a feature is low if it occurs in many documents and is high if the feature occurs in a small number of documents. The weight of a feature t in a document d is denoted by$tfidf(t,d)$.Features which occur in many documents are rated less important because their commonality leads to a low inverse document frequency, and they do not carry any characteristic information of the document.

We normalize the feature weights as follows.

$$w(t,d) = \frac{tfidf(t,d) - \min(t,d)}{\max(t,d) - \min(t,d)}$$

Normalization is carried out on each document d and the word weights are used as elements of feature vector.

3.5. Elimination of Features below the Threshold

If a feature weight $w(t,d)$ is below a threshold, the feature is eliminated. An example is shown in Figure 2.

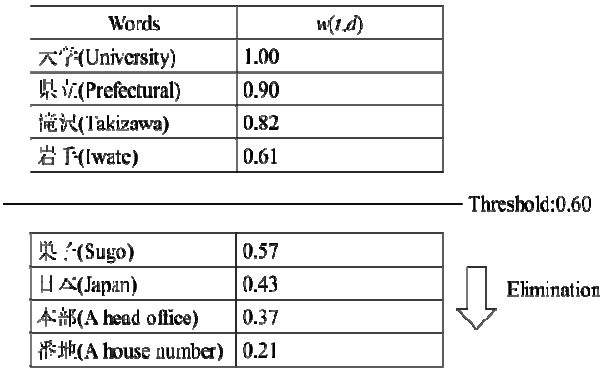

Figure 2. *An example of feature elimination.*

For instance, suppose the threshold is 0.6.The word

[3]http://chasen-legacy.sourceforge.jp/

[4]http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

features that have weights below 0.6 are eliminated in Figure 2, e.g., “巣子(Sugo)”, “日本(Japan)”, “本部(Head office)” and “番地(House number)” are eliminated. In fact, these are common words and do not carry any special information. The features retained for clustering the documents are “大学(University)”, “県立(Prefectural)”, “滝沢(Takizawa)” and “岩手(Iwate)”.

Only the features greater than the threshold are used for clustering documents. The threshold can be set to any value between 0.1~0.9, andit is possible to eliminate less important features of a document by adjusting this threshold.

3.6. Clustering Search Results

For clustering, our proposed method uses Bayon[5], a simple and fast hard-clustering tool. We use the k-means method implemented in Bayon that partitions a data set into k groups by selecting k initial cluster centers and then iteratively refining them as follows:

1. Each document D is assigned to its closest cluster center.
2. Each cluster center C is updated to be the mean of its constituent documents.
3. The algorithm converges when there is no further change in assignment of documents to clusters and can be viewed as the compaction of the cluster.

The word weight vector is used for clustering the documents and the cosine similarity is used to calculate the document similarity of the cluster members. If each cluster has only one document, the document similarity is 100[%];however, the number of clusters will be very large and not helpful for the user. On the other hand, if the number of clusters is small, the document similarity within each cluster is low, and the clusters will have no meaning. We need to reduce the number of clusters as much as possible while retaining a high similarity for the members within a cluster.

We adjust the value of k such that the similarity of all documents within a cluster is over 50[%],indicating a high similarity of documents within the cluster.In this case, if a document within a cluster is moved to another cluster, the similarity will always drop to under 50[%].

We start with $k = 1$.If the document similarity within the cluster is not over 50[%],k is updated to $k + 1$ and then the k-means clustering is run again.If the document similarity is over 50[%], then clustering is stopped.

3.7. Calculation of Concordant Document Frequencies and Labeling Clusters

In a cluster, documents have a word feature t, same for all documents.The sum of the weights of the features is given by $W(t, C(l))$ and is calculated as follows:

$$W(t, C(l)) = \sum_{d \in C(l)} w(t, d)$$

where $C(l)$ is a cluster with a label l (an ordinal number).If many documents in a cluster have the same feature, the feature might be a significant one for the cluster, and thus, the feature is given a higher weight.

The concordant document frequency $CDF(t, C(l))$ of a word feature t in a cluster is calculated by

$$CDF(t, C(l)) = Num(C(l) \cap R(t, q))$$

where q is a query,$R(t, q)$ is a web search result where both the word feature t and the original query q are used, and $Num(C(l) \cap R(t, q))$ is the number of documents which are members of both the cluster $C(l)$ and $R(t, q)$.

An example of the calculation for the concordant document frequencies of each word in a cluster is shown in Figure 3.

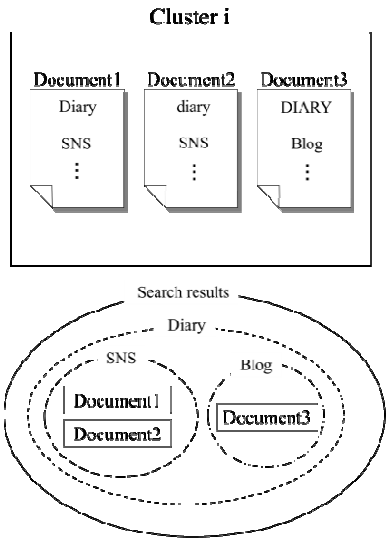

Figure 3. *Example of calculation for the concordant document frequencies of each word feature.*

Suppose a cluster i has three documents.The word features corresponding to each document are shown in the box.The search result for “Diary” includes all documents because the search engine decides “Diary”, “diary” and “DIARY” have the same meaning and finds all documents. The search result for “SNS” includes Document1 and Document2, while the search result for “Blog” includes Document3 only. The values of $CDF(t, C(l))$ for each of the three word features are as follows:

- Diary: 3
- SNS : 2
- Blog : 1

The value of $CDF(t, C(l))$ is different from the word frequency in a search result, for instance,the value of $CDF(t, C(l))$ for “Diary” is three although the frequency of word “Diary” is one in Figure 3.Our proposed method calculates the proper values using $CDF(t, C(l))$, and

[5]http://code.google.com/p/bayon/

thendetermines $CDFW(t, C(l))$ as follows:

$$CDFW(t, C(l)) = CDF(t, C(l)) \cdot W(t, C(l))$$

where $CDFW(t, C(l))$ determines the importance of the word feature t of a cluster $C(l)$, andis calculated by multiplying $CDF(t, C(l))$ and $W(t, C(l))$.

We believe that by using $CDFW(t, C(l))$, our proposed method can appropriately label clusters with high confidence.

4. Experiments

We evaluated the accuracy of the labels given by the system based on our proposed method. We used ranks of labels to evaluate our proposed method. The experiment results were evaluated by human users.

The system used the Yahoo! JAPAN[6] web search engine in the experiment.

4.1. Experimental Data

We use the following Japanese query words: "AKB48"[7], "アマゾン(Amazon)", "地震(Earthquake)", "楽天(Rakuten)"[8] and "価格(Price)" for the experiment. These search queries are the top five queries of the Yahoo! Search Ranking[9] in Japan.We kept the first one hundred results.

The threshold for elimination of features was set at 0.2, determined by preliminary experiments[16].Five university students (not the authors of this paper) took part in the evaluation. The clusters of the search engine results were given labels by our proposed method.

4.2. Evaluation Method

The appropriate label for a cluster was determined by users in order to evaluate accuracy of our proposed method.

First, the titles and snippets of cluster documents were shown to the users. If a user decided that a cluster was inappropriate, the user did not give a label to the cluster and it was excluded from the experimental data. If a user decided a cluster was appropriate, the top ten candidate labels outputted by the system were shown to the user and the user then chose the most appropriate label from among them. If a user decided there was no appropriate label, the user wrote an appropriate label in a text box, and the appropriate label was ranked eleven or more.

Our proposed method then ordered the labels by descending weight. We evaluated our proposed method using the rank of the appropriate label for a cluster, and accuracy was compared using the rank value. We compared our proposed method with algorithms where either $CDF(t, C(l))$ or TermExtract was not used in order to evaluate the importance of the proposed concordant document frequency and TermExtract.

[6]http://www.yahoo.co.jp/

[7]AKB48 is a popular group in Japan.

[8]Rakuten is a popular e-commerce site in Japan.

[9]http://searchranking.yahoo.co.jp/

4.3. Results and Discussion

Table 2 shows the number of clusters for each query. The total number of distinct clusters was 157, and five users evaluated each of the clusters. The number of appropriate clusters evaluated by a user was 432, derived from the number of potential clusters, $157 \times 5(users) = 785$, because inappropriate clusters were decided individually by each users.We evaluated our proposed method for the 432 appropriate cluster labels.

Table 2.*Number of clusters for each query.*

Query q	Number of clusters
AKB48	32
アマゾン(Amazon)	32
地震(Earthquake)	32
価格(Price)	31
楽天(Rakuten)	30
Total	157

Table 3.*The rate[%] of the appropriate labels.*

	Our proposed method	Without $CDF(t, C(l))$	Without TermExtract
1	36.6	36.1	36.6
~2	46.6	44.4	46.6
~3	53.8	51.1	53.8
~4	59.1	55.0	57.6
~5	62.3	57.3	60.6
~6	63.7	58.2	61.8
~7	67.9	61.0	65.7
~8	69.8	62.6	67.8
~9	72.3	63.8	69.9
~10	73.1	64.6	70.6

Table 4. *The recall[%], precision[%] and F-measure of the appropriate labels in our proposed method.*

	Recall	Precision	F-measure
1	36.6	90.3	52.1
~2	46.6	83.3	59.8
~3	53.8	81.4	64.8
~4	59.1	62.4	60.7
~5	62.3	59.9	61.1
~6	63.7	59.7	61.6
~7	67.9	60.3	63.9
~8	69.8	61.2	65.2
~9	72.3	59.4	65.2
~10	73.1	60.4	66.2

Table 3 shows the rate of recall of appropriate labels above each rank, e.g. "~3" indicates the rate of recall of the labels that were ranked 1, 2, or 3.The top ten labels of our proposed method covered 73.1[%] of the labels, nearly equal to the accuracy of Lingo[14].We believe that the rate could be further improved by another, more accurate clustering algorithm.

The rate of the top ten labels of the algorithm without $CDF(t, C(l))$ is 64.6[%] as shown in Table 3, and indicates that the concordant document frequency is effective. The top ten coverage of the algorithm without TermExtract is

70.6[%].We believe that the reason for better labeling is that the compound words reconstructed by TermExtract are ranked in the top ten. Thus, TermExtract is effective for our proposed method.

Table 4 shows the recall, precision and F-measure of the appropriate labels above each rank and Figure 4 shows the changes in recall and precision of our proposed method. The maximum precision is 90.3[%].When the system outputs the labels above rank 3, the F-measure is 64.8, which is a high value.

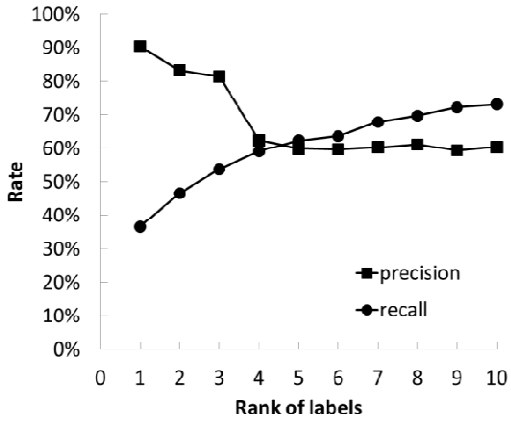

Figure 4. *Changes in recall and precision*

The number of labels for each cluster shown to the users is significant. When clusters are labeled and a few cluster labels for each cluster are presented to users, they are easily able to understand contents of web search results by browsing through the cluster labels. When a lot of labels are presented, they are not able to understand contents. We consider that search is effective when three labels determined by our proposed method are shown to users.

Some compound word labels are not properly extracted by TermExtract. Examples are shown in Table 5.

Table 5.*Examples of compound words not extracted by TermExtract.*

Cluster labels	TermExtract
新物流センター (New distribution center)	物流センター (Distribution center)
楽天イーグルス (Rakuten Eagles)	楽天イーグル (Rakuten Eagle)

For example, “新物流センター(New distribution center)” was divided into “新(New)” and “物流センター(Distribution center)”, while users gave the label “新物流センター(New distribution center)” to the cluster. When the rank is 11 or more, it is not certain whether the contents of the cluster match the query of the user or not.

The results of higher rank labels and lower rank labels given by TermExtract are shown in Table 6.Here, of course, higher rank indicates lower importance. In Table 6, we show seven labels that were assigned rank 10 or less (more important), and two labels that were assigned rank 11 or more (less important).“報告(Report)” and “通販(Mail order)” went up from rank 9 to rank 11.However seven labels were ranked from “11～” to the top ten, and were compound words.

Table 7 shows results of compound words split by ChaSen. For example, the word “避難所(Shelter)” was split into two words “避難(Refuge)” and “所(Place)”.The label “所(Place)” has many meanings, and if the label of the cluster is “所(Place)”, it might be difficult to understand at a glance that the cluster labeled “所(Place)” contains documents about “避難所(Shelter)”.The compound word “避難所(Shelter)” reconstructed by TermExtract is appropriate, and we consider it effective to use compound words as labels.

Table 6.*Results of labels of which ranks change.*

Rank 11 or more	Rank 10 or less
報告(Report) 通販(Mail order)	クラウド(Cloud) データセンター(Data center) 研究所(Laboratory) ＡＭＡＺＯＮ(AMAZON) クックパッド(Cook pad) 楽天カード(Rakuten card) 避難所(Shelter)

Table 7.*Examples of compound words split by ChaSen.*

Compound words	ChaSen
クラウド(Cloud)	クラ(Kura)，ウド(Udo)
データセンター(Data Center)	データ(Data)，センター(Center)
研究所(Laboratory)	研究(Study)，所(Place)
ＡＭＡＺＯＮ(AMAZON)	Ａ(A)，Ｍ(M)，Ａ(A)，Ｚ(Z)，Ｏ(O)，Ｎ(N)
クックパッド(Cookpad)	クック(Cook)，パッド(Pad)
楽天カード(Rakuten Card)	楽天(Rakuten)，カード(Card)
避難所(Shelter)	避難(Refuge)，所(Place)

We showed by simulation experiments that the accuracy of the labels given by our proposed method is high. We also demonstrated that our proposed method is able to represent a few proper labels for each cluster to users. Therefore, our proposed method that uses both the concordant document frequency and TermExtract is effective for labeling clusters.

5. Conclusion

In recent years, the amount of information on the World Wide Web has exploded. A robot-type search engine is often used for searching; however, it can be difficult for users to come up with an appropriate query and to understand the contents of search results. To solve these problems, many methods have been proposed. Many proposed methods classify search engine results into clusters and label them. A significantly representative word should be used as a cluster label, and most proposed methods are based on TFIDF.

We proposed a labeling method using concordant document frequencies for the clusters of web search results. In addition to TFIDF, our proposed method uses the concordant document frequencies in the clusters and is able to assign cluster-aware labels.

Our proposed method uses a system of Japanese morphological analysis, however, the system of Japanese

morphological analysis cannot extract compound words, so the proposed method uses TermExtract for making compound words. Both the words and the compound words are then used as document features and the proposed method classifies web search results into clusters using these features.

We evaluated the accuracy of our proposed method, comparing the accuracy of our proposed method with and without concordant document frequency and TermExtract. We showed by simulation experiments that the use of concordant document frequency and TermExtract are effective. We also demonstrated the recall and precision of our proposed method and proved that our proposed method is able to represent a few proper labels for each cluster to users and that the accuracy is high.

One element of our future work is to evaluate the efficiency of our proposed method, showing the reduction in search time.

We used k-means for clustering documents in our experiments. The clusters of this method do not overlap. In general, however, clusters could overlap because web documents can have content in common. If documents were allowed to overlap among clusters, we believe our proposed method would work better.

Another possibility is that web documents could be classified using labels. If a cluster has a label and other clusters have the same label, then the clusters could be combined using that label. Thus, web documents could be reclassified as a refinement.

References

[1] JerzyStefanowski and DawidWeiss, “Carrot2 and Language Properties in Web Search Results Clustering”, Advances in Web Intelligence, 2003.

[2] Carrot, http://search.carrot2.org/stable/search

[3] Yippy, http://search.yippy.com/

[4] Toshihiro Yoshida, MasafumiMatsuhara, GoutamChakraborty and Hiroshi Mabuchi, “A Novel Ranking Method of Web Search Result Using Clustering and Concordance Count”, Proc. of WCCI 2012 IEEE World Congress on Computational Intelligence, pp.902--907, Brisbane, Australia, June 10-15, 2012.

[5] Marti A. Hearst and Jan O. Pedersen, “Reexamining the Cluster Hypothesis: Scater/Gather on Retrieval Results”, SIGIR'96, ACM, pp.76-84, 1996.

[6] Patrick Pantel and Dekang Lin, “Document clustering with committees”,SIGIR'02, ACM, pp.199-206, 2002.

[7] OmarAlonso, MichaelGertz and RicardoBaeza-Yates, “Clustering and Exploring Search Results using Timeline Constructions”,CIKM'09, pp.97-106, 2009.

[8] Songhua Xu, Tao Jin and Francis C.M. Lau, “A New visual Search Interface for Web Browsing”,Proc. 2nd ACM International Conference on Web Search and Data Mining, ACM, pp.152-161, 2009.

[9] OrenZamir, OrenEtzioni, OmidMadani and RichardM. Karp, “Fast and Intuitive Clustering of Web Documents”,Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997.

[10] OrenZamir and OrenEtzioni, “Web Document Clustering A Feasibility Demonstration”, SIGIR 1998, 46-54.

[11] Oren Zamir and Oren Etzioni, “Grouper: A Dynamic Clustering Interface to Web Search Results”,WWW'99: Proc. 8th international World Wide Web Conference, pp.1361-1374, Elsevier North-Holland, Inc., 1999.

[12] DavidCarmel, HaggaiRoitman and NaamaZwerdling, “Enhancing Cluster Labeling Using Wikipedia”,Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp.139-146, 2009.

[13] Paolo Ferragina and Antonio Gulli, “A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering”,WWW'05: Special interest tracks and posters of the 14th international conference on World Wide Web, ACM, pp.801-810 , 2005.

[14] Stanis law Osinski, Jerzy Stefanowski and Dawid Weiss,“Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition”,Proc. International IIS: IIPWM'04 Conference, pp.359-368, 2004.

[15] ThorstenJoachims, “A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization”,DTIC Document, 1996.

[16] Toshihiro Yoshida, MasafumiMatsuhara, GoutamChakraborty and Hiroshi Mabuchi,“Labeling Method with Threshold in Web Search Results”,Proc. of FIT2011, pp.365--366, Hakodate, Japan, September 7-9, 2011.(*in Japanese*)