Query Refinement based on Topical Term Clustering

Hiromi Wakaki[1] , Tomonari Masada[2] , Atsuhiro Takasu[2] , and Jun Adachi[2]
[1] The University of Tokyo [2] The National Institute of Informatics
Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo, Japan 101-8430
{hiromi, masada, takasu, adachi}@nii.ac.jp

Abstract

We propose a method for supporting query refinement using topical term clusters. First, we propose a new term weighting method that can extract terms strongly related to a specific topic, because a document set retrieved with an ambiguous query may include divergent topics. Our formulation of term weighting is based on the statistics of term co-occurrence. Then, we generate term clusters using extracted terms, and rerank the documents in the search results by using each term cluster as a query. This clustering procedure is intended to isolate each topic as a set of related terms. In our experiments, we evaluated our term weighting method by checking: 1) whether each of the top-ranked document sets corresponds to one topic; and 2) whether some of the top-ranked document sets cover all the topics included in the synthesized document set. The results of our experiment show our method outperforms the existing term weighting methods MI, KLD, CHI-square and RSV.

1. Introduction

When we use existing search engines such as Google, Yahoo, and MSN, we enter only a few terms to form a query (Jansen et al., 1998). The search engines often then return a long list of search results. Even if we use effective query terms, e.g., proper nouns and technical terms, various topics related to the query are contained in the search results retrieved by such a short query. Therefore, we must select the documents we are interested in from the list by examining the titles and snippets. This is a time-consuming task because the list is unstructured, and it is not easy for web users to understand the multiple topics contained in the search results.

In this paper, we propose a method for supporting query refinement by using clusters of topical terms extracted from a retrieved set of documents. We assume that a topic is implied by a specific set of terms that frequently co-occur in the same documents. Therefore, we introduce a new measure of term importance called *tangibility*. A term is said to have tangibility when it frequently co-occurs exclusively with a specific set of terms. Our method aims to extract terms exclusively related to one of those topics. Then, we divide the extracted terms into clusters using a distributional clustering algorithm, which leads to agglomerates of terms frequently co-occurring with each other. In our experiments, we examined the quality of term clusters by checking the effectiveness of the clusters for query refinement.

2. Related Work

2.1 Organizing Search Results

Result categorization and query refinement are well-known techniques for the further improving the search results. For the purpose of result categorization, clustering and classification are traditionally used in the categorization algorithms. Scatter/Gather (Cutting et al., 1992) was one of the systems where the clustering approach was tested. On the other hand, DynaCat (Pratt & Fagan, 2000) was a prototype system that used the classification approach. Both of them aimed to categorize documents retrieved as the search results. In this paper, we aim to make term clusters corresponding to each topic contained in the search results. We do not make document clusters but term clusters.

There is another stream to handle the search results. The Findex system (Maki, 2005) does not classify the retrieved documents but extracts terms that represent the features of the documents

included in the search results. When users select one of the sets of terms, the system shows the document containing the selected terms. The Findex system aims to show terms and phrases corresponding to the document categories regardless of the topics. In contrast, we aim to make term clusters, each of which corresponds to each topic contained in the search results.

After these streams, a faceted search is the next approach to organizing the search results (Hearst, 2006). Facets denote attributes in various orthogonal sets of categories (Adkisson, 2005; Hearst et al., 2006). We believe that our method provides good candidates to generate facets because our method aims to extract terms exclusively related to one of topics contained in the search results and to make term clusters, each of which corresponds to each topic.

2.2 Keyword Extraction

Sanderson et al. (Sanderson & Croft, 1999) extracted terms and made concept hierarchies from the search results. They used term co-occurrences to find strong relationships between terms. In this paper, we aim to improve the search results with term clustering using term co-occurrences.

Specific terms such as proper nouns and technical terms, can succinctly represent a specific topic. Therefore, we aim to make term clusters that can easily understand the topics included in a document set. Many studies have been conducted on well-known term weighting methods (Sebastiani, 2002; Yang & Pedersen, 1997). We compare those methods with our proposed method in this paper. Other studies have concentrated on term extraction, such as the measurement of term representativeness (Hisamitsu et al., 2000), keyword extraction from one document by using term clustering (Matsuo & Ishizuka, 2004), and DualNAVI (Takano et al., 2000).

There are various methods to look for additional terms used for query expansion, e.g., Robertson's Selection Value (RSV) (Robertson, 1990). However, query expansion methods do not take into consideration that the search results may contain multiple topics. On the other hand, we think it is easier to browse term clusters divided into topics than to browse many documents showing various topics. Moreover, term clusters can also be used as additional query terms to refine the original query.

3. New Method for Topical Term Extraction and Term Clustering

3.1 New Term Weighting Method

In this paper, we say two terms co-occur when they appear in the same document. Let $P(t_i)$ be the occurrence probability of a term t_i. $P(t_i)$ is estimated as the number of documents in which t_i appears divided by the total number of documents. Let $P(t_j|t_i)$ be the occurrence probability of t_j among the documents including t_i. $P(t_j|t_i)$ is estimated as the number of documents in which t_i and t_j co-occur divided by the number of documents where t_i appears. In the same way, let $P(\neg t_j|t_i)$ be the nonoccurrence probability of t_j among the documents including t_i. Let U be a document set, and let $U(t_i)$ be a document set in which t_i appears. Let S be a document subset, and let $S(t_i)$ be a document subset in which t_i appears. For example, U is a corpus for retrieval, and S is a set of retrieved documents.

In this work, we aim to find topical term clusters for query refinement. For this purpose, we extract useful terms from the retrieved documents to discriminate between the topics included in those documents. We call this feature of a term, *tangibility*. Here, we introduce a hypothesis that a term co-occurring frequently with a group of terms has a high tangibility, that is, such a term is useful for topic discrimination. To calculate how high a term's tangibility is, we will introduce a formula, called TNG. First, the following equation measures how much the probability of t_j's appearance increases by adding the condition that t_i appears.

$$\Delta_{t_i}(t_j) = P(t_j \mid t_i) \cdot \log \frac{P(t_j \mid t_i)}{P(t_j)} \qquad (1)$$

By setting $F_i = \{ t_j \mid \Delta_{ti}(t_j) > 0 \}$, we obtain the formula for TNG, as follows.

$$TNG(t_i) = \frac{\sum_{t_k \in F_i} \{\Delta_{t_i}(t_k)\}}{|F_i|} \quad (2)$$

Equation (2) is a revised version of our previous formulations in (Wakaki et al., 2006). Equation (2) weights terms, and we can rank terms by using those weights. If a term has a low frequency, we must avoid the problem of data sparseness. Therefore, we use Dirichlet smoothing (Huo, Liu, & Feng, 2005).

3.2 Term Clustering using Terms Extracted by Proposed Method

It is important to define an appropriate similarity between the terms for term clustering. In this paper, let the similarity between terms t_i and term t_j be as follows.

$$Sim(t_i, t_j) = \frac{|S(t_i) \cap S(t_j)|}{|S(t_i) \cup S(t\ )|} \quad (3)$$

We set $Sim(t_i, t_j)=0$ if $|S(t_i) \cap S(t_j)|<5$ in our experiment. Next, the distance between clusters is defined as follows.

$$Sim(C_1, C_2) = \frac{s(C_1, C_2)}{s(C_1, C_1) \times s(C_2, C_2)} \quad (4)$$

Here, we defined $s(C_1, C_2)$ as follows.

$$s(C_1, C_2) = \sum_{t_i \in C_1} \sum_{t_j \in C_2} Sim(t_i, t_j) \quad (5)$$

We use the distributional clustering (Baker & McCallum, 1998; Mallela et. al., 2003) proposed by Baker et al. This clustering algorithm uses the ranks of terms when it makes clusters, so the generated clusters are different for different methods of ranking terms. Baker et al. ranked the terms by using mutual information (MI). However, in this paper, we exchange MI for other term weighting methods, such as TNG.

4. Experiments

4.1 Comparison Methods

We compared our proposed method TNG with four other term weighting methods: MI (Yang & Pedersen, 1997; Yoshioka & Haraguchi, 2004), KLD, χ-square (Sebastiani, 2002), and RSV (Robertson, 1990). MI, KLD, and χ-square use term co-occurrence, just like TNG does, and these methods can measure how t_j's occurrence probability changes by adding the condition that t_i occurs. Then, the weight of term t_i for these methods is calculated as follows.

$$W(t_i) = \sum_j X(t_j; t_i) \quad (6)$$

Here, $X(t_j;t_i)$ is replaced by $MI(t_j;t_i)$, $KLD(t_j;t_i)$, and $\chi^2(t_j;t_i)$, respectively. We also used the same smoothing for the three other methods as we did for TNG.

4.2 Data Set for Experiments

Our experiments require a document set that includes multiple topics and in which each document has labels indicating a topic. A data set for document classification surely has these labels. Moreover, a test collection for the evaluation of information retrieval is accompanied with relevant document sets for each test query, so we can generate a mixture data set containing multiple topics that are indicated by each relevant document set. Therefore, we used both types of data. Furthermore, we used data sets in Japanese and English. We used seven data sets, described in Table 1. Each data set has categories, and three of the largest categories were mixed for the experiment as pseudo-data including multiple topics. The names of the categories or query terms we used are also shown as A, B, and C in Table 1.

Data set	L	Type	No. of documents (A+B+C)	A	B	C
NTCIR3 web	J	IR	*1108*	0032	0013	0028
NTCIR4 web	J	IR	*2113*	0006	0058	**0082**
Dmoz	E	DC	*21089*	Math	Chemistry	Astronomy
Reuters	E	DC	*6615*	Earn	acq	Crude
Sankei sports news	J	DC	*3519*	Japanese baseball	MLB	Soccer
Newsgroup20	E	DC	*3000*	talk.politics . Guns	talk.politics . mideast	talk.politics . Misc
NTCIR-CLIR	E	IR	*209*	0036	0023	0018

Table 1: Languages, data types, and numbers of documents used for our experiment. (L:Language; E:English; J:Japanese; DC: document classification; IR: information retrieval). A, B, and C are category names or query terms used. Where the original task is IR, query IDs are shown.

4.3 Experimental Procedure

First, we weighted each term using the five methods TNG, MI, KLD, χ-square, and RSV for each data set, as indicated in Section 4.1. Next, we made term clusters by using the top 100 terms ranked by each method. Finally, we used every term cluster as a query, and ranked the documents included in the heterogeneous data by using the Okapi probabilistic model (Robertson & Walker, 1999). Let $Prec(C_i,L_j)$ be the precision when we retrieve documents with a term cluster C_i as a query and use the documents in the category L_j as relevant documents. $Prec(C_i, L_j)$ is defined as the number of retrieved relevant documents in the top x documents divided by x. We define the category $L(C_i)$ corresponding to C_i, as follows.

$$L(C_i) = \arg\max_{L_j} Prec(C_i, L_j) \qquad (7)$$

Further, we define the precision $Prec(C_i)$ of a term cluster C_i, as follows.

$$\Pr ec(C_i) = \max_{L_j} Prec(C_i, L_j) \qquad (8)$$

Additionally, we define the precision $Prec(L_j)$, which is the maximum precision of $Prec(C_i)$ corresponding to L_j, as follows.

$$\Pr ec(L_j) = \max_{\{C_i \quad s.t. \quad L(C_i)=L_j\}} Prec(C_i) \qquad (9)$$

First, we examined the relevance of the documents retrieved with the term clusters generated by each method as a query. For this purpose, we compared the average of all $Prec(C_i)$s. Second, we compared the completeness of the categories of the term clusters by each method, by examining the average of $Prec(L_j)$s for all L_j.

4.4 Experimental Results

We compared the average of $Prec(C_i)$ for all C_i for the top 5, 10, and 100 ranked documents (See Fig. 1). TNG had the best average precision for the top 5 ranked documents in all the data sets. TNG also had the best average precision for the top 10 ranked documents in NTCIR3, NTCIR4, Dmoz, Sankei Sports News, Newsgroup20, and NTCIR-CLIR. TNG showed stable performance for all data sets and for the top 5, 10, and 100 ranked documents. On the other hand, RSV had the second-best average precision for the top 5 and 10 ranked documents in NTCIR, NTCIR4, Dmoz, and NRCIR-CLIR. As a result, we found that TNG outperformed the other comparison methods in the top-ranked documents. Furthermore, TNG outperformed the other comparison methods for a wide variety of data sets. That is, TNG extracts terms that are strongly related to one of the three topics.

We compared the completeness of categories for the top 5, 10, and 100 ranked documents (See Fig.2). We evaluated the completeness by the average of $Prec(L_j)$ for all L_j. TNG had the best completeness of categories in NTCIR3, NTCIR4, Dmoz, Sankei Sports News, and Reuters. Although χ-square is the best in Newsgroup20, and RSV is the best in NTCIR-CLIR, these methods did not outperform TNG in the other data sets. Therefore, TNG outperformed the other methods with respect to overall performance.

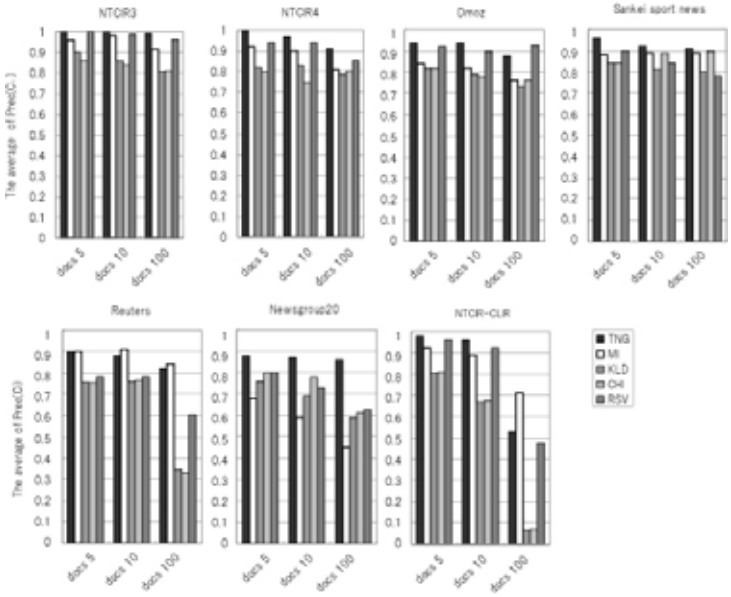

Figure 1: Average of $Prec(C_i)$ for all C_i for top 5, 10, and 100 ranked documents of NTCIR3, NTCIR4, Dmoz, Sankei Sports News, Newsgroup20, and NTCIR-CLIR.

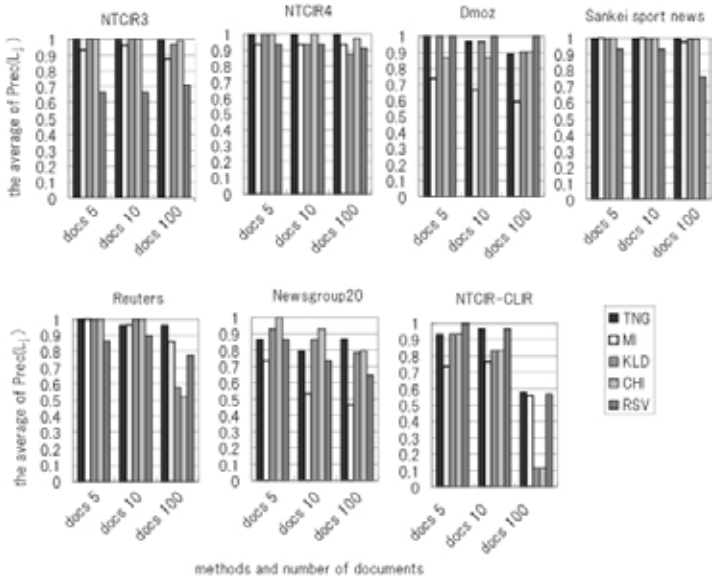

Figure 2: Average of $Prec(L_j)$ for all L_j for top 5, 10, and 100 ranked documents of NTCIR3, NTCIR4, Dmoz, Sankei Sports News, Newsgroup20, and NTCIR-CLIR.

5. Conclusion

We have proposed a method to support query refinement by using clusters of topical terms extracted from a retrieved set of documents. We introduced the hypothesis that a term co-occurring frequently with a specific group of terms is useful in discriminating the topics included in a document set. This hypothesis is reflected in the formulation of our new term weighting method, called TNG.

In our experiments, we compared the performance of TNG with those of the term weighting methods MI, KLD, χ-square, and RSV. First, we extracted terms using each method and generated term clusters by using these terms. Next, we retrieved documents with the term clusters as a query from a heterogeneous set of documents. With respect to the average precision of documents retrieved by the clusters, TNG outperformed the other methods. Furthermore, TNG had a good completeness of categories from the documents retrieved by the term clusters. We can conclude that TNG is an efficient term weighting method for detection of topics included in a heterogeneous set of documents. We think that TNG can be used for query expansion that takes into consideration that various topics are contained in the first-retrieved documents and that users can select one of those topics efficiently with our method.

References

Adkisson, H. P. (2005). *Use of Faceted Classification*, http://www.webdesignpractices.com/navigation/facets.html

Baker, L. D. & McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings. of SIGIR-98* (pp. 96–103).

Cutting, D., Karger, D., Pedersen, J., & Tukey, J. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings. of SIGIR 92* (pp. 318–329).

Hearst, M. A. (2006). Clustering versus faceted categories for information exploration, *Communications of the ACM*, 49(4):59-61.

Hearst, M. A., Smalley, P., & Chandler, C. (2006). Faceted Metadata for Information Architecture and Search, *CHI 2006 Course*.

Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M., & Takano, A. (2000). Extracting terms by a combination of term frequency and a measure of term representativeness. *International journal of theoretical and applied aissues in specialized communication*, 6(2):211–232.

Huo, H., Liu, J., & Feng, B. (2005). Multinomial approach and multiple-bernoulli approach for information retrieval based on language modeling. In *Proceedings. of Fuzzy Systems and Knowledge Discovery* (pp. 580–583).

Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Searchers, the subjects they search, and sufficiency: A study of a large sample of excite searches. In *1998 World Conference on the WWW and Internet*.

Maki, M. (2005). Findex: Search result categories help users when document ranking fails. In *Proceedings. of the SIGCHI conference on Human factors in computing systems* (pp. 131–140).

Mallela, S., Dhillon, I. S., & Kumar, R. (2003). A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research(JMLR): Special Issue on Variable and Feature Selection* (pp. 1265–1287).

Matsuo, Y. & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13:157–169.

Pratt, W. & Fagan, L. (2000). The usefulness of dynamically categorizing search results. *Journal of the American Medical Informatics Association*, 7(6):605–617.

Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.

Robertson, S. E. & Walker, S. (1999). Okapi/keenbow at trec-8. In *TREC*.

Sanderson, M. & Croft, B. (1999). Deriving concept hierarchies from text. In *Proceedings. of SIGIR 99* (pp. 206–213).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Takano, A., Niwa, Y., Nishioka, S., Iwayama, M., Hisamitsu, T., Imaichi, O., & Sakurai, H. (2000). Associative information access using dualnavi. In *Proceedings. of ICDL'00* (pp. 285–289).

Wakaki, H., Masada, T., Takasu, A., & Adachi, J. (2006). A New Measure for Query Disambiguation using Term Co-occurrences, In *Proceedings. of 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)* (pp. 904-911).

Yoshioka, M. & Haraguchi, M. (2004). Study on the combination of probabilistic and Boolean IR models for WWW documents retrieval. In *Working Notes of NTCIR-4 (Supplement Volume)* (pp. 9–16).

Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings. of ICML-97* (pp. 412–420).