A New Measure for Query Disambiguation using Term Co-occurrences

Hiromi Wakaki[1], Tomonari Masada[2], Atsuhiro Takasu[2], and Jun Adachi[2]

[1] The University of Tokyo, Graduate School of Information Science and Technology,
Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan 113-0033
hiromi@nii.ac.jp
[2] The National Institute of Informatics,
Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo, Japan 101-8430
{masada, takasu, adachi}@nii.ac.jp

Abstract. This paper explores techniques that discover terms to replace given query terms from a selected subset of documents. The Internet allows access to large numbers of documents archived in digital format. However, no user can be an expert in every field, and they trouble finding the documents that suit their purposes experts when they cannot formulate queries that narrow the search to the context they have in mind. Accordingly, we propose a method for extracting terms from searched documents to replace user-provided query terms. Our results show that our method is successful in discovering terms that can be used to narrow the search.

1 Introduction

The Internet allows us to access large numbers of documents archived in digital format. However, when a search presents many documents, we usually look at only a few top-ranked documents. Further, the search results often contain a large number of unrelated documents. Therefore, we need results to be more compact and relevant to our intentions. This is why we should use query terms that indicate our needs exactly when we search in the Internet. However, queries consisting of two or three terms are often not reliable enough to gather appropriate Web pages, because such queries are intrinsically ambiguous. Moreover, we cannot be experts in every field, so we often cannot formulate queries that narrow the search to the context we have in mind. In many cases, we hit upon only a few terms to refine the initial query, and end up with unsatisfactory search results even after refining the query. Therefore, the research challenge of query processing has been to find more appropriate query terms and to provide search results that can meet the needs of various users[1]. To meet this challenge, we propose a method for extracting terms from a given set of documents to replace or add to the original query terms. Our results show that our method is successful in discovering terms that can be used to narrow the search.

Our method analyzes the co-occurrence of terms in the top-ranked documents of the initial search result and extracts terms having a special feature called

Tangibility. When a term refers to a specific concept or denotes a particular thing, we say the term has Tangibility. A proper noun is a typical example of a term that has Tangibility. Our method is based on the following observation: we can easily disambiguate a short query by adding just one term that has Tangibility. Our method works regardless of the retrieval method we used. Moreover, the method can extract terms to expand queries without using additional data such as word networks or structural directories of concepts. Our approach is unique because we propose a new term-weighting formula that can extract terms that imply a distinct topic. Our experiments have shown that the terms extracted by our method are qualitatively different from those extracted by other existing measures.

The rest of the paper is organized as follows: Section 2 introduces our new concept, Tangibility; Section 3 reports the details and results of our experiments; Section 4 discusses prior work; and Section 5 concludes with a summary of the paper and an indication of future work.

2 Tangibility

2.1 Hypothesis of Tangibility

To cope with ambiguities in queries, we propose a new method for selecting terms. The aim of our method is to find more specific terms than the query terms the user has given; these specific terms can match more easily with distinct topics and resolve query ambiguity. Here we are introducing a new concept called *Tangibility*. We say that a term has Tangibility when it keeps a fairly close relationship with the given query and, at the same time, is strongly related to a distinct topic, regardless of whether or not the topic is principal in the retrieved document set. We measure the Tangibility of a term t by focusing on the variety of terms frequently co-occurring with t. To obtain numerical estimates of Tangibility, we formulate our hypothesis as follows:

> *A term co-occurring frequently with a limited number of terms in a retrieved document set can establish a distinct topic in the document set.*

We call this the *hypothesis of Tangibility*. We say two terms *co-occur* when they appear in the same document. Each term is counted only once, even if it appears many times within the document. In the following subsection, we propose two numerical estimates for term Tangibility: TNG1 and TNG2.

2.2 TNG1: First Formulation of Tangibility

Suppose we have a document corpus U. Let $S \subset U$ be the set of the top l ranked documents retrieved with a query. $U(t_i)$ (resp. $S(t_i)$) denotes the set of documents from U (resp. S), in which the term t_i appears. $V(d)$ denotes the number of terms in document d. Our first numerical estimate of Tangibility, denoted by TNG1, is based on the hypothesis described in Section 2.1. To obtain

TNG1, we introduce the average number of terms that appear in documents that include term t_i, and denote it by $F(t_i)$. More formally, $F(t_i)$ is defined as follows:

$$F(t_i) = \frac{\sum_{d \in S(t_i)} (V(d) - 1)}{|S(t_i)|}. \tag{1}$$

$F(t_i)$ shows how many terms appear with t_i in S. Therefore, we can regard a term with small $F(t_i)$ as a term representing a distinct topic. However, a term having small $|S(t_i)|$ intrinsically has small $F(t_i)$. We therefore introduce an additional component into our formula so that terms of small $|S(t_i)|$ should not always be regarded as having Tangibility. Consequently, we obtain the following formula as TNG1, which expresses the Tangibility of a term t_i:

$$TNG1(t_i) = \frac{|S(t_i)|^2}{|U(t_i)|} \cdot \frac{1}{F(t_i)}, \tag{2}$$

where the first half is obtained by multiplying $|S(t_i)|$ by $|S(t_i)|/|U(t_i)|$. $|S(t_i)|$ is simply the document frequency of t_i in S and indicates how strongly t_i is *unconditionally* related to S. In contrast, $|S(t_i)|/|U(t_i)|$ indicates how strongly t_i is related to S *in comparison with* U.

2.3 TNG2: Second Formulation of Tangibility

To provide the second formulation TNG2, we rewrite Equation (1) as follows:

$$F(t_i) = \sum_{t_j \neq t_i} \frac{|S(t_i \wedge t_j)|}{|S(t_i)|}, \tag{3}$$

where $|S(t_i \wedge t_j)|$ is defined to $|S(t_i) \cap S(t_j)|$. Since we can interpret $|S(t_i \wedge t_j)|/|S(t_i)|$ as the probability of the occurrence of t_j among the documents including t_i, we denote it by $P(t_j|t_i)$. According to TNG1, t_i has Tangibility when $\sum_{t_j \neq t_i} P(t_j|t_i)$ is small. In contrast, we devise the second formulation, TNG2, by requiring $P(t_j|t_i)$ to be *smaller than* $P(t_j)$ for a large number of t_js ($j \neq i$). TNG1 and TNG2 share the same intuition. However, we introduce an elaboration into TNG2, i.e., the comparison of $P(t_j|t_i)$ with $P(t_j)$. For the discrepancy evaluation of the two probability distributions, Kullback–Leibler Divergence (KLD) is often used. In our case, $P(t_j|t_i)$ and $P(t_j)$ are to be compared. Moreover, the event complementary to the occurrence of t_j is the non-occurrence of t_j, denoted by $\neg t_j$. Therefore, the KLD for our evaluation can be written as:

$$KL(t_j; t_i) = P(t_j|t_i) \log \frac{P(t_j|t_i)}{P(t_j)} + P(\neg t_j|t_i) \log \frac{P(\neg t_j|t_i)}{P(\neg t_j)}.$$

However, $\sum_{t_j \neq t_i} KL(t_j; t_i)$ cannot evaluate the Tangibility of t_i, because this sum is large when any of the following two conditions holds for many t_js:

(a) $P(t_j|t_i) \log \frac{P(t_j|t_i)}{P(t_j)} < 0$ $\left(\text{and thus } P(\neg t_j|t_i) \log \frac{P(\neg t_j|t_i)}{P(\neg t_j)} > 0 \text{ also holds}\right)$

(b) $P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} > 0$ $\left(\text{and thus } P(\neg t_j|t_i)\log\frac{P(\neg t_j|t_i)}{P(\neg t_j)} < 0 \text{ also holds}\right)$

Only (a) is important for the Tangibility of t_i. Therefore, we propose a new measure, called Signed Kullback–Leibler (SKL), as follows:

$$SKL(t_j;t_i) = -P(t_j|t_i)\log\frac{P(t_j|t_i)}{P(t_j)} + P(\neg t_j|t_i)\log\frac{P(\neg t_j|t_i)}{P(\neg t_j)}. \quad (4)$$

SKL is derived from the KLD by changing the sign of the first term. Consequently, we propose the following as the second formula for Tangibility:

$$TNG2(t_i) = \frac{|S(t_i)|^2}{|U(t_i)|}\cdot\sum_{t_j\neq t_i} SKL(t_j;t_i).$$

3 Experiment

3.1 Metrics for Term Selection

Before describing our experiment in detail, we present the various term-weighting schemes that we tested. Term weight $W(t_i)$ for term t_i is computed by multiplying two weights: $|S(t_i)|^2/|U(t_i)|$ and $CW(t_i)^\sigma$. The former weight was introduced in Section 2.2. As for $CW(t_i)^\sigma$, we obtain $CW(t_i)$ by summing $cw(t_i,t_j)$ for all t_js ($j \neq i$), i.e., $CW(t_i) = \sum_{t_j\neq t_i} cw(t_i,t_j)$, where each summand $cw(t_i,t_j)$ is computed based on the co-occurrence of t_i and t_j; σ takes a value of 1 or -1. When we want an increase (resp. decrease) in $CW(t_i)$ to contribute to an increase of $W(t_i)$, σ is set to 1 (resp. -1).

Many recent studies have proposed various term-weighting methods for term extraction. Some of them use term co-occurrence frequencies as in our formulations of Tangibility. We compared eight term-weighting methods[13] (see Table 1). *UnitWeight* is so called because $CW(t_i) = 1$ for any t_i. This method ignores the effect of term co-occurrence. That is, *UnitWeight* is intended to reveal how the difference of $CW(t_i)$ works in each of the other term-weighting methods. Co-occurrence Frequency (CF) is prepared for ranking terms based on the intuition contrary to that of TNG1. Mutual Information (MI) [14], KLD, and χ^2 measure the discrepancy between $P(t_j|t_i)$ and $P(t_j)$. Thus, they are all based on nearly the same intuition about term importance. With these measures, however, we cannot distinguish the two cases (a) and (b) shown in Section 2.3. Since these methods and our two methods use term co-occurrence frequencies, we next discuss the computational cost to obtain co-occurrence frequencies for all pairs of terms. Let m be the number of terms that appear in S. In the worst case, the computational cost is proportional to m^2. However, for most t_i, $|\{t_j : |S(t_i \wedge t_j)| > 0\}| \ll m$ holds. Therefore, the actual computational cost can be reduced by choosing an appropriate data structure. Robertson's selection value (RSV) [10] is a term weight used for query expansion in information retrieval[11]; it does not use co-occurrence information. The exact formulation is

as follows:

$$RSV = \left(\frac{|S(t_i)|}{|S|} - \frac{|U(t_i)|}{|U|}\right) \cdot \left\{ \alpha \cdot \log \frac{|U|}{|U(t_i)|} + (1-\alpha) \cdot \log \frac{\frac{|S(t_i)|+0.5}{|S|-|S(t_i)|+0.5}}{\frac{|U(t_i)|-|S(t_i)|+0.5}{|U|-|U(t_i)|-|S|+|S(t_i)|+0.5}} \right\},$$

where α is a parameter. We set $\alpha = 1/2$ in our experiment.

Table 1. Formulations used in our experiments. We replace $P(t_j|t_i)$, $P(t_j|\neg t_i)$, $P(\neg t_j|t_i)$, $P(\neg t_j|\neg t_i)$, and $|S(t_i)|^2/|U(t_i)|$ with A, B, C, D, and U, respectively.

method	$cw(t_i, t_j)$	$W(t_i) = U \cdot \left\{\sum_{t_j \neq t_i} cw(t_i, t_j)\right\}^{\sigma}$
TNG1	$\lvert S(t_i \wedge t_j)\rvert / \lvert S(t_i)\rvert$	$U \cdot \left\{\sum_{t_j \neq t_i} cw(t_i, t_j)\right\}^{-1}$
TNG2	$-A \log \frac{A}{P(t_j)} + C \log \frac{C}{P(\neg t_j)}$	$U \cdot \sum_{t_j \neq t_i} cw(t_i, t_j)$
UnitWeight	—	$U \cdot 1$
CF	$\lvert S(t_i \wedge t_j)\rvert / \lvert S(t_i)\rvert$	$U \cdot \sum_{t_j \neq t_i} cw(t_i, t_j)$
MI	$P(t_i)\left\{A \log \frac{A}{P(t_j)} + C \log \frac{C}{P(\neg t_j)}\right\}$ $+P(\neg t_i)\left\{B \log \frac{B}{P(t_j)} + D \log \frac{D}{P(\neg t_j)}\right\}$	$U \cdot \sum_{t_j \neq t_i} cw(t_i, t_j)$
KLD	$A \log \frac{A}{P(t_j)} + C \log \frac{C}{P(\neg t_j)}$	$U \cdot \sum_{t_j \neq t_i} cw(t_i, t_j)$
χ^2	$\frac{\{A-P(t_j)\}^2}{P(t_j)} + \frac{\{C-P(\neg t_j)\}^2}{P(\neg t_j)}$ $+\frac{\{B-P(t_j)\}^2}{P(t_j)} + \frac{\{D-P(\neg t_j)\}^2}{P(\neg t_j)}$	$U \cdot \sum_{t_j \neq t_i} cw(t_i, t_j)$

3.2 Experimental Procedure and Results

We used a document set prepared for the NTCIR3 Web Retrieval Task[3]. This set includes about ten million Web pages written in Japanese. We denote this Web page set by U. The Web pages in U are decomposed into terms by using a morphological analyzer MeCab[8] equipped with a Japanese dictionary ipadic-2.5.1[6]. There are 47 queries prepared for the NTCIR3 Web task, and each query includes two or three query terms. First, we issued the queries and obtained the top 1000 Web pages for each query. Although our experiment adopted an Okapi-type term-weighting scheme for Web page retrieval[4], our method can be applied to the search results obtained with other term-weighting schemes. From the top 1000 pages of each of the 47 retrieval results, we gathered terms appearing in five or more pages. We obtained about 10,000 terms for each query. We did not delete stop words. Next, we computed the eight term weights described in Section 3.1. As a result, we obtained eight term rankings by sorting the terms with respect to their eight kinds of weights. For every term ranking, we added each of the top five terms (a, b, c, d, and e) separately to the original query term set $\{A, B, C\}$

Table 2. Overall precisions and their improvements compared to the baseline. The baseline overall precision is 0.1606.

method	overall precision	improvement(%)
UnitWeight	0.1765	9.9
TNG1	**0.1847**	15.0
TNG2	**0.1899**	18.2
CF	0.1801	12.1
MI	0.1829	13.9
KLD	0.1733	7.9
χ^2	0.1751	9.0
RSV	**0.1867**	16.3

and made five expanded sets of query terms $\{A, B, C, a\}$, $\{A, B, C, b\}$, ..., $\{A, B, C, e\}$. Finally, we retrieved the Web pages with these expanded query term sets. Consequently, we obtained five search results for each query. We computed the average precisions of these five results by using `trec_eval`[12]. Of these five average precisions, we kept only the best one, because this average precision can be taken as the performance measure of the information retrieval most desirable for users who are supposed to issue the corresponding query. Finally, we regarded the mean of the best average precisions of the 47 queries as the overall precision for each term-weighting method.

For the original 47 queries, we obtained 0.1606 as the overall precision and regarded it as the baseline. Among the eight term-weighting formulae, TNG1, TNG2, and RSV significantly increased overall precision (Table 2). TNG2 achieved the best overall average precision with an 18.2% improvement. The most important point is that TNG1 and TNG2 showed qualitative differences from RSV. While RSV tends to extract terms having general meanings, TNG1 and TNG2 can extract many technical terms used in specific domains relative to the query. For a query including "loudspeaker", "comparison", and "evaluation", our method extracted such technical terms as "woofer" and "bass reflex" (Table 3). For a query involving "the World Tree", "Norse mythology", and "name", our method extracted "Yggdrasill", which is the name of a mythological tree in Norse mythology and is a synonym of "the World Tree" (Table 4).

4 Related Work and Discussion

Numerous studies have been done on keyword extraction. Most of them report methods for extracting topic-centric terms, such as technical terms and proper nouns[2][9]. Rennie and Jaakkola [9] introduced a new informativeness measure, the Mixture score, which focuses on the difference in log-likelihood between a mixture model and a simple unigram model, to identify informative words. They compare it against a number of other informativeness criteria, including the Inverse Document Frequency (IDF) and Residual IDF (RIDF)[2]. While the results show their measure works well when compared with existing methods,

Table 3. Terms extracted for the query consisting of "loudspeaker", "comparison", and "evaluation". (The baseline average precision is 0.0596.)

Query terms are"スピーカー (loudspeaker)", "比較 (comparison)", and " 評価 (evaluation)"

method	rank	top five terms	average precision
RSV	1	"アンプ (amplifier)"	0.0067
	2	"可能 (possible)"	0.0279
	3	"結果 (result)"	0.0341
	4	"システム (system)"	0.0246
	5	"音 (sound)"	**0.0593**
TNG1	1	"アンプ (amplifier)"	0.0067
	2	"ウーファー (woofer)"	0.0660
	3	"ソフトドームツィーター(soft dome tweeter)"	0.0154
	4	"スーパーウーファー (super-woofer)"	0.0177
	5	"バスレフ (bass reflex)"	**0.0661**
TNG2	1	"アンプ (amplifier)"	0.0067
	2	"ウーファー (woofer)"	0.0660
	3	"バスレフ (bass reflex)"	**0.0661**
	4	"サブウーファー (sub-woofer)"	0.0640
	5	"低音 (bass sound)"	0.0434

Table 4. Terms extracted for the query consisting of "the World Tree", "Norse mythology", and "name". (The baseline average precision is 0.0675.)

Query terms are"世界樹 (the World Tree)", "北欧神話 (Norse mythology)", and " 名前 (name)"

method	rank	top five terms	average precision
RSV	1	"神 (God)"	0.0253
	2	"たち (they)" *1	0.0280
	3	"それ (it)" *2	**0.0377**
	4	"歴史 (history)"	0.0266
	5	"物語 (tale)"	0.0198
TNG1	1	"Pandaemonium"	0.0586
	2	"イグドラシル (Yggdrasill)"	**0.3767**
	3	"ソグネフィヨルド(Sognefjorden)"	0.0523
	4	"エッダ (Edda)"	0.0525
	5	"シルマリル (Silmaril)"	0.0533
TNG2	1	"イグドラシル (Yggdrasill)"	**0.3767**
	2	"古事記 (Kojiki)" *3	0.0227
	3	"ギリシア (Greece)"	0.0160
	4	"ノルウェー (Norway)"	0.0203
	5	"フィヨルド (fjord)"	0.0287

*1 A suffix used to make Japanese nouns plural.
*2 A kind of Japanese pronoun.
*3 The oldest known historical book about the ancient history of Japan.

the documents they used are all posts to a restaurant discussion bulletin board, so these results cannot be seen as conclusive.

The method proposed by Hisamitsu et al. [5] compares the term frequency distributions in an entire document set with those in the set of documents containing a specific term t. When a large discrepancy exists between them, t is said to have *representativeness*. This method estimates a discrepancy similar to ours. However, our concern lies in the *direction* of the discrepancy. We ask whether the frequency of terms other than t in an entire set is higher or lower than that in a set of documents including t. When the latter is less than the former with respect to a large number of terms, we say that t has Tangibility. Therefore, we believe Tangibility is novel. Matsuo et al. [7] also proposed a term extraction method based on term co-occurrences. Their method combines term ranking by the χ^2 measure with term clustering. However, this method is designed for application to a single document. In contrast, our aim is to disambiguate a query by finding the terms corresponding to distinct topics latent in a set of hundreds of retrieved documents. Therefore, we have proposed a new measure for term extraction.

5 Conclusion and Future Work

We proposed a co-occurrence-based measure, called Tangibility, for term extraction to disambiguate queries. Our experiments obtained very interesting results worthy of further investigation. Both of our numerical estimates for term tangibility, TNG1 and TNG2, realized good average precisions. In addition, many of the extracted terms were related to more specific topics than that implied by the original ambiguous query terms. Our method may be used as a key component of a system that helps users to discover specific topics from a given corpus simply by using fairly general terms as search keywords. As future work, we plan to propose a method of clustering the terms that have Tangibility; we will test to determine whether the term clusters correspond to distinct topics implied by the initial query terms.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
2. K. Church and W. Gale. Inverse document frequency (IDF): A measure of deviations from poisson. In *Proc. of 3rd Workshop on Very Large Corpora*, pages 121–130, 1995.
3. K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Overview of the Web retrieval task at the third NTCIR workshop. In *Proc. of NTCIR-3*, pages 1–24, 2003.
4. H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proc. of SIGIR 2004*, pages 49–56, 2004.
5. T. Hisamitsu, Y. Niwa, S. Nishioka, H. Sakurai, O. Imaichi, M. Iwayama, and A. Takano. Extracting terms by a combination of term frequency and a measure of term representativeness. *Terminology*, 6(2):211–232, 2001.
6. ipadic-2.5.1. http://chasen.naist.jp/stable/ipadic/.
7. Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-ocurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13:157–169, 2004.
8. MeCab. http://mecab.sourceforge.jp/.
9. J. Rennie and T. Jaakkola. Using term informativeness for named entity detection. In *Proc. of SIGIR'05*, pages 353–360, 2005.
10. S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.
11. M. Toyoda, M. Kitsuregawa, H. Mano, H. Itoh, and Y. Ogawa. University of Tokyo/RICOH at NTCIR-3 Web retrieval task. In *Proc. of NTCIR-3*, pages 31–38, 2003.
12. TREC. trec_eval, http://trec.nist.gov/trec_eval/.
13. Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of ICML-97*, pages 412–420, 1997.
14. M. Yoshioka and M. Haraguchi. Study on the combination of probabilistic and boolean ir models for www documents retrieval. In *Working Notes of NTCIR-4 (Supplement Volume)*, pages 9–16, 2004.