# Topic-Oriented Query Expansion for Web Search

Shao-Chi Wang * and Yuzuru Tanaka
Meme Media Laboratory, Graduate School of
Information Science and Technology, Hokkaido University
Kita 13, Nishi 8, Kita-Ku Sapporo, Hokkaido, Japan 060-8628
{wang, tanaka}@meme.hokudai.ac.jp

## ABSTRACT

The contribution of this paper includes three folders: (1) To introduce a topic-oriented query expansion model based on the Information Bottleneck theory that classify terms into distinct topical clusters in order to find out candidate terms for the query expansion. (2) To define a term-term similarity matrix that is available to improve the term ambiguous problem. (3) To propose two measures, intracluster and intercluster similarities, that are based on proximity between the topics represented by two clusters in order to evaluate the retrieval effectiveness. Results of several evaluation experiments in Web search exhibit the average intracluster similarity was improved for the gain of 79.1% while the average intercluster similarity was decreased for the loss of 36.0%.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval—*clustering, query formulation, retrieval models*

## General Terms

Theory, Experimentation, Measurement

## Keywords

query expansion, topic-oriented, information bottleneck, term-term similarity matrix, intracluster similarity, intercluster similarity

## 1. INTRODUCTION

One of the problems in selecting candidate terms for the query expansion strategy is the ambiguity of the original query term[2]. In the current research, we suppose that each query term can be separated into multiple concepts, and the identical concepts of a term shall be grouped into an identical cluster. In each cluster, a bundle of terms will keep a unique topic that only relates to one concept of the original query. In the cluster to which the query term concepts of our concern belong, other terms can be selected as candidates of the query expansion. We call this strategy "topic-oriented query expansion". We employ the Information Bottleneck (IB) method[5] for clustering purposes. Before performing the clustering technique, we first propose a new definition for the construction of a term-term similarity matrix. The current work is based on the problem of word sense disambiguation[3][6] and its probabilistic solutions [7][8]. Furthermore, we propose intracluster and intercluster similarity measures to evaluate the relevance of the topic of the cluster associated with the candidate terms respectively to the retrieved documents in the cluster itself, and to the documents in the other clusters. The result shows that the clusters obtained by our method have higher intracluster cohesiveness and lower intercluster cohesiveness.

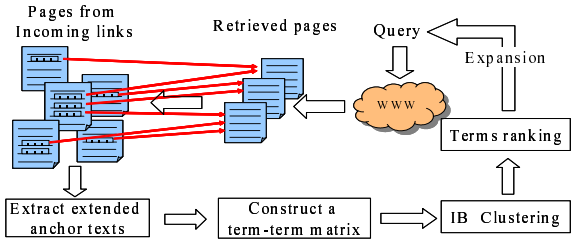


**Figure 1: The topic-oriented query expansion model**

## 2. TOPIC-ORIENTED QUERY EXPANSION MODEL

Figure 1 describes the outline of the topic-oriented query expansion model which consists of five steps. They are,

**(1)** to collect pages that come from incoming links.
**(2)** to extract extended anchor texts.
**(3)** to construct a term-term similarity matrix.
**(4)** to establish a topic-oriented cluster.
**(5)** to rank terms.

The first two steps can be readily achieved. In the following sections, we will introduce the three remaining steps.

### 2.1 Constructing a term-term similarity matrix

The conventional term-term similarity matrix of reference[1] was redefined as follows:

DEFINITION 1. *The frequency of a term $t_i$ ($t_i \neq q_k$, $q_k$ denotes a term of the query $q$) in a mini-document $d_j$, is referred to as $f_{t_i,j}$. At the same time, we suppose that a term $q_k$ of the query $q$ is of divergent concepts when it appears in different mini-documents. Hence, a query term $q_k$ in a mini-document $d_l$ is substituted by its concept $c_{k,l}(l \in j)$ and its frequency is referred to as $f_{c_{k,l},j}$. A term-document matrix is represented as $\vec{m} = (m_{rj})$, where $m_{rj} = f_{t_i,j} or f_{c_{k,l},j}$. Here, the number of rows $r$ is equal to $|T_L| + \sum_{k \in |q|} |C_{k,L}| (t_i \in T_L, c_{k,l} \in C_{k,L})$ and the number of columns is $|D_L|$, where $T_L$ denotes a term set that exclude the query terms, $C_{k,L}$ the concept set of query $k$, and $D_L$ the document collection.*

*The first author is now in Yahoo Japan Corporation
E-mail: swang@yahoo-corp.jp


*WWW 2006*, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

*Let $\vec{m}^t$ be the transpose of $\vec{m}$. The matrix $\vec{s} = \vec{m}\vec{m}^t$ defines the local term-term similarity matrix.*

In this new matrix, we suppose that if a term appears in different documents, it will have different meanings. We then count the frequency of each concept rather than the term itself. After performing clustering, the identical mutually equivalent meanings of a term will be aggregated into the same cluster because this term is used in the same sense to describe a common topic. Note that we use an extended anchor text as a mini-document.

## 2.2 Establishing a topic-oriented cluster

We employ sIB (sequential Information Bottleneck)[4] to cluster the extracted terms in the proposed model. The input information structure is the obtained term-term similarity matrix in Section 2.1.

## 2.3 Ranking terms

Based on the clustering results, we propose a conditional entropy to rank terms as follows.

$$e(y|t) = -\sum_{t} p(t)p(y|t)log_2 p(y|t) \quad (1)$$

The variables $X(x \in X)$ and $Y(y \in Y)$ denote the same term collection. The formula (1) represents the probability distribution of each term being yielded in the obtained cluster.

**Table 1: Term clustering(the conventional definition)**

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| business | offer | committee | people | mining |
| knowledge | clustering | chairman | discovery | data |
| research | development | thesis | tip | storm |
| team | localization | member | workshop | web |
| directory | interface | advisory | project | link |
| acquistion | risk | management | proceeding | model |
| mine | marketing | student | expert | suite |
| mission | goverment | conference | library | intelligence |
| ma | tree | consulting | technique | software |
| record | decision | job | these | traffic |

**Table 2: Term clustering(the proposed definition)**

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| mining(8) | mining(9) | mining(8) | mining(6) | mining(10) |
| Web(8) | Web(5) | Web(5) | Web(2) | Web(6) |
| data | offer | business | perspective | ma |
| software | clustering | stand | suite | track |
| research | people | technology | find | information |
| text | conjunction | knowledge | workshop | link |
| intelligence | tip | chairman | ole | retrieval |
| server | database | committee | tree | review |
| expert | end | advisory | storm | support |
| system | localization | member | risk | format |

# 3. EXPERIMENT AND EVALUATION

## 3.1 Description of experiments

Here, we report an example of searching Web documents with a query *"Web mining"*. We picked up top 10 valid pages and set the maximum number of the incoming links that point to one core document to 100. Tables 1 and 2 show the top 10 terms corresponding to the conventional and the proposed definitions. Note that the number in the bracket after each query term denotes the number of its occurrences in each cluster.

## 3.2 Description of evaluation

The intracluster and intercluster similarities which are based on proximity between topics are described as follows.
A term vector $V_q^i$ $(i \in C$, C is the set of clusters) associated with the expanded query is defined by the equation (2) and a document vector $V_{d_k}^j$ $(j \in C)$ is defined by the equation (3). $w_{q_l}$ and $w_{d_{kl}}$ denote the weights of the terms associated with the expanded query and the document, respectively. $N$ the set of the documents.

$$Term\ vectors: \quad V_q^i = \{w_{q_1}, w_{q_2}, ..., w_{q_n}\}^i \ \ (i \in C) \quad (2)$$

$$Document\ vectors: \quad V_{d_k}^j = \{w_{d_{k1}}, w_{d_{k2}}, ..., w_{d_{kn}}\}^j \ \ (j \in C, k \in N) \quad (3)$$

The average intracluster and intercluster similarities are defined as follows:

*Average intracluster similarity* :

$$\frac{1}{|C|} \sum_{i=j \in C} \sum_{k \in N} cosine(V_q^i, V_{d_k}^j) \quad (4)$$

*Average intercluster similarity* :

$$\frac{1}{|C| \cdot (|C|-1)} \sum_{i,j \in C, i \neq j} \sum_{k \in N} cosine(V_q^i, V_{d_k}^j) \quad (5)$$

The larger intracluster similarity and the smaller intercluster similarity correspond to the higher cohesiveness of the cluster, namely the higher topic proximity of the retrieved documents in the cluster. The results are exhibited in Table 3. In comparison with the conventional definition, the average intracluster similarity of the proposed definition is improved for the gain of 79.1% whereas the average intercluster similarity is decreased for the loss of 36.0%. This shows that the new definition of the term-term similarity matrix results in significantly better performance than the conventional one in the topic-oriented query expansion model.

**Table 3: Average intracluster and intercluster similarities**

| | Intracluster similarity | Intercluster similarity |
|---|---|---|
| Old Definition | 1.101 | 1.038 |
| New definition | 1.972 | 0.763 |

# 4. CONCLUSIONS

In order to develop a topic-oriented query expansion strategy, we have proposed a new definition of the term-term similarity matrix to employ the IB clustering technique, and two measures to evaluate the relevance. Experimental results and evaluations have verified a significant improvement obtained by the proposed query expansion model.

# 5. ACKNOWLEDGMENTS

The work presented in this paper has been supported by the 21st Century COE (Center of Excellence) Program of Japan Society of the Promotion of Science (JSPS).

# 6. REFERENCES

[1] B.-Y. Ricardo and R.-N. Berthier. *Modern Information Retrieval*, 1999.
[2] E.-N. Efthimiadis. Query expansion, 1996.
[3] M. Sanderson. Word sense disambiguation and information retrieval, 1994.
[4] N. Slonim et al.. Unsupervised document classification using sequential information maximization, 2002.
[5] N. Tishby et al. The information bottleneck method, 1999.
[6] S. Hinrich. Automatic word sense discrimination, 1998.
[7] T. Hofmann. Probabilistic latent semantic indexing, 1999.
[8] V. Lavrenko and W. B. Croft. Relevance-based language models, 2001.