TOPIC-ORIENTED QUERY EXPANSION FOR WEB SEARCH

Shao-Chi Wang* and Yuzuru Tanaka**

*: Yahoo Japan Corporation, Japan {swang@yahoo-corp.jp} ; **: Meme Media Laboratory, Hokkaido University {tanaka@meme.hokudai.ac.jp}

ABSTRACT:

Contributions:

(1) Introduce a topic-oriented query expansion model based on the Information Bottleneck theory
(2) Define a term-term similarity matrix
(3) Propose two measures, intracluster and intercluster similarities

Purposes:

- In (1), we classify terms into distinct topical clusters in order to find out candidate terms for the query expansion
- With (2), we are available to improve the term ambiguous problem
- Two measures in (3) are based on proximity between the topics represented by two clusters in order to evaluate the retrieval effectiveness

INTRODUCTION:

Term ambiguous example:

The original text mining algorithm was created by a university student named Taylor. In our interview, we found out that he is currently in his second year as an undergraduate student of the Mining Engineering Department of this university and has many other interests beside of programming.

Remarks:

(1) The first "mining" and the second "Mining" describe two different senses.
(2) Co-occurrence terms are different, i.e., "algorithm" and "Engineering".
(3) However, in the conventional term-term similarity matrix, "mining" is counted twice times (Figure 1) and considered as one concept.

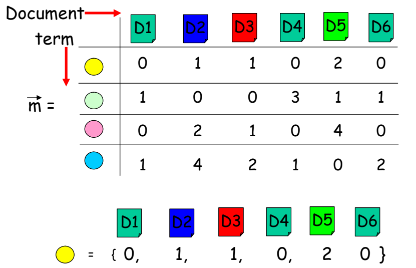

Figure1: Conventional term-term similarity matrix

Remarks:

Term (yellow circle) have different senses in documents D2, D3 and D5. However, in the conventional term-term similarity matrix, we only considered (yellow circle) as one concept and count its numbers of amount in the documents D2, D3 and D5. Using this type of similarity matrix to look for the co-occurrence words of the original query will lead to the problem of term ambiguous.

RELATED WORKS:

Information Bottleneck Theory [5]

A compressed variable is X and its relevant variable is Y.

The purpose is to search a middle variable T (clusters of X) in which the mutual information between T and Y, i.e., $I(T,Y)$ is maximized while the mutual information $I(X,T)$ is minimized.

$$\mathcal{L}[P(T|X)] = I(X,T) - \beta I(T,Y) \quad (1)$$

three self-consistent equations

$$\begin{cases} p(t|x) = \frac{p(t)}{Z(\beta,x)} exp(-\beta D_{KL}(p(y|x)||p(y|t))) \\ p(y|t) = \frac{1}{p(t)} \sum_x p(t|x)p(x)p(y|x) \\ p(t) = \sum_x p(t|x)p(x) \end{cases} \quad (2)$$

TOPIC-ORIENTED QUERY EXPANSION MODEL:

2. Topic-oriented query expansion model:

(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.

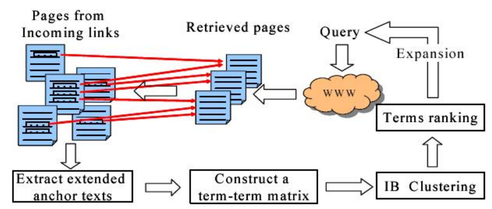

Figure 2. The topic-oriented query expansion model

2.1 Constructing a term-term similarity matrix

Refer to [1], we redefine the term-term similarity matrix

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 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 are 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 are $|D_L|$, 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. 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.*

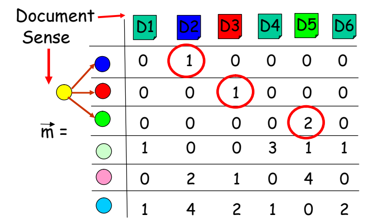

Figure 3: Our proposed term-term similarity matrix

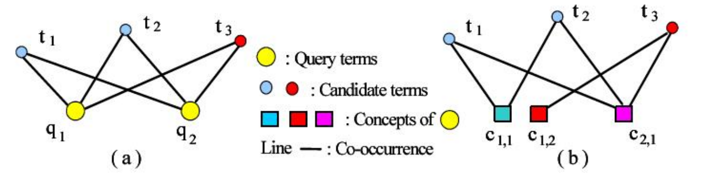

Figure 4: Correlation among candidate terms, query terms, and query concepts

Example:

Query q : $q = \{q_1, q_2\}$
Document d_1 : $d_1 = \{t_1, q_1, t_2\}$
Document d_2 : $d_2 = \{t_1, t_2, q_2, t_3\}$
Document d_3 : $d_3 = \{q_1, t_3\}$
Document d_4 : $d_4 = \{t_2, t_3\}$

Using the conventional similarity matrix

$$\vec{m} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ q_1 \\ q_2 \end{array} \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{array}$$

Using our proposed similarity matrix

$$\vec{m} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ c_{1,1} \\ c_{1,2} \\ c_{2,1} \end{array} \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{array}$$

2.2 Establishing a topic-oriented cluster

We employ sIB (sequential Information Bottleneck)[4] to cluster the extracted terms.

2.3 Ranking terms

$$e(y|t) = -\sum_t p(t)p(y|t)log_2 p(y|t) \quad (3)$$

The variables X(x∈X) and Y(y∈Y) denote the same term collection.

INTRACLUSTER AND INTERCLUSTER SIMILARITY MEASURES:

We expand the candidates together with original query to an n-dimensional vector which are postulated to proximity of a topic.

$$Term\ vectors: \quad V_q^i = \{w_{q1}, w_{q2}, ..., w_{qn}\}^i \quad (i \in C) \quad (4)$$

$$Document\ vectors: \quad V_{d_k}^j = \{w_{dk1}, w_{dk2}, ..., w_{dkn}\}^j \quad (j \in C, k \in N) \quad (5)$$

Average intracluster similarity:

$$\frac{1}{|C|} \sum_{i=j \in C} \sum_{k \in N} cosine(V_q^i, V_{d_k}^j) \quad (6)$$

Average intercluster similarity:

$$\frac{1}{|C| \cdot (|C|-1)} \sum_{i,j \in C, i \neq j} \sum_{k \in N} cosine(V_q^i, V_{d_k}^j) \quad (7)$$

C is the set of clusters, N denotes the collection of the documents that are retrieved by the new expanded query.

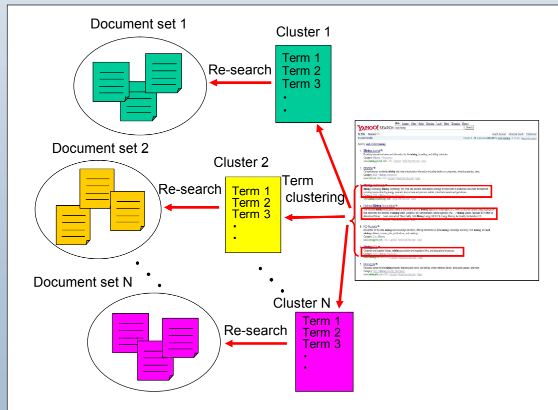

Figure 5: A topic-proximity measure

EXPERIMENT AND EVALUATION:

(1) Experiment in January, 2005

Original query: "Web AND mining" ; Search Engine: Google; Term source: Anchor texts that link to the top 10 valid pages.

Terms clustering by using the conventional definition and sIB

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
data	Web	index	art	design
discovery	mining	company	clip	intelligence
knowledge	text	listing	[illegible]	
resource	reference	major		
tech	information	list		
warehouse	retrieval	portal	link	
guide	extraction			
center	institute			
	technology			
	workshop			
	usage			
	analysis			
	proceeding			

Relate to topic "Web mining"
Relate to topic "mining institute"
Relate to topic "workshop"

Note: There are three topical terms including in the same cluster (cluster 2).

Terms clustering by using our proposed definition and sIB

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
mining(9)	mining(13)	Web(6)	mining(8)	Web(12)
Web(6)	Web(12)	mining(5)	Web(6)	mining(3)
data	text	workshop	institute	design
resource	reference	company	technology	intelligence
tech	information	index	art	designer
map	retrieval	usage	clip	
discovery	extraction	analysis	collection	
knowledge	list	proceeding	net	
road	portal	listing	links	
guide		major	link	
warehouse				
center				

Note: Terms relating to "Web mining", "mining institute" and "workshop" are divided into three different clusters.

(2) Experiment and Evaluation in August, 2005

Experiment: Original query: "Web AND mining" ; Search Engine: Google; Term source: 100 extended anchor texts that link to each one of the top 10 valid pages.

Evaluation: Collect the top 30 documents by using the extended query term sequences corresponding to each clusters. Compute intracluster and intercluster similarities by using equations (4)-(7).

Experiment:

Term clustering (Old 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

Term clustering (New 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

Evaluation:

Intracluster and intercluster similarities (Old definition)

Similarity	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Cluster 1	0.647	0.704	2.240	1.557	1.070
Cluster 2	0.990	1.314	1.329	1.162	0.957
Cluster 3	3.420	0.483	0.775	1.045	0.558
Cluster 4	0.538	0.600	0.694	1.679	0.656
Cluster 5	0.334	0.606	1.067	0.759	1.091

Intracluster and intercluster similarities (New definition)

Similarity	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Cluster 1	3.180	0.946	1.668	0.796	0.712
Cluster 2	0.756	1.109	0.786	0.600	0.574
Cluster 3	0.588	0.644	2.781	0.273	0.382
Cluster 4	0.421	1.672	0.527	0.646	0.380
Cluster 5	1.055	1.145	0.871	0.468	2.144

Average intracluster and intercluster similarities

	Intracluster similarity	Intercluster similarity
Old Definition	1.101	1.038
New definition	1.972	0.763
Improvement	79.1% ↑	36.0% ↓

CONCLUSIONS:

(1) We proposed a topic-oriented query expansion model and employed IB in this model.
(2) In order to treat the multiplicity of the query term meanings at the conceptual level. we have proposed a new definition of the term-term similarity matrix.
(3) We have proposed intracluster and intercluster similarity measures to evaluate the relevance between a topic of a cluster and the retrieved documents in the cluster itself as well as the documents in the other clusters.
Experimental results and evaluations showed that the obtained candidate terms using the new definition of the term-term similarity matrix in each cluster are almost of higher topic relevance than the obtained ones using the old definition.

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)

BIBLIOGRAPHY:

[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.

printed by MegaPrint INC. www.postersession.com