The KLE's Subtopic Mining System for the NTCIR-9 INTENT Task

Se-Jong Kim, Hwidong Na, and Jong-Hyeok Lee
Pohang University of Science and Technology (POSTECH), Korea

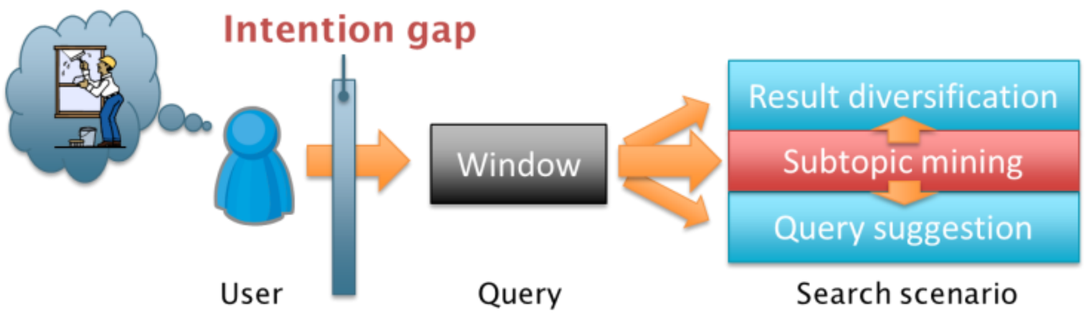
Introduction

Intention gap: This gap is a discrepancy between a user's search intent and a query. Because of the intention gap, to clearly express users' search intents in keywords is not trivial for them. This situation brings about short and vague queries that are ambiguous and broad.

Ambiguous/broad queries: If queries are ambiguous, users may seek for different senses, and if queries are broad, users may get interested in different aspects.

Subtopic mining: Subtopics are senses of a topic (or query) and aspects of each sense. Subtopic mining is that mines users' underlying subtopics of a topic.

Ex. Topic: windows
Sense: Microsoft Windows, house windows
Aspect of Microsoft Windows: Windows7, Windows update



Method

Step 1. Refinement of data: We deleted the fields of date/time and ranks from the original query log (SogouQ), and remained the data lines whose click-sequences are 1 or 2. We also removed unnecessary characters and the white space from the query log.

Step 2. Finding possible subtopics: If the query in the refined query log matched up to all bigrams of the topic and was not same to the topic, we considered the query as the possible subtopic (full-match-subtopic).

If the query matched up to the front/back bigram of the topic, we extracted the text string that appeared before/after this bigram from the query and attached the extracted text string to the front/back of the topic (front/back-match-subtopic).

Step 3. Estimating scores of subtopics: We estimated scores(*Score*) of the subtopics using some equations based on interest and clearness. The interest(*IS*) was measured by the number of distinct session IDs for the subtopic. The clearness(*CS*) was measured by the reciprocal number of distinct clicked URLs that satisfied the click-sequence 1 or 2 for the subtopic.

$$Int(s) = \begin{cases} \frac{\sum_{s' \in OQ_s} \#ID_{s'}}{\#OQ_s}, & If\ s \in S_{front}\ or\ s \in S_{back} \\ \#ID_s, & otherwise \end{cases}$$

$$IS(s) = \frac{Int(s)}{\sum_{p \in S_{all}} Int(p)}$$

$$Amb(s) = \begin{cases} \frac{\sum_{s' \in OQ_s} \#URL_{s'}}{\#OQ_s}, & If\ s \in S_{front}\ or\ s \in S_{back} \\ \#URL_s, & otherwise \end{cases}$$

$$CS(s) = \frac{1/Amb(s)}{\sum_{p \in S_{all}} (1/Amb(p))}$$

$$Score(s) = \lambda IS(s) + (1-\lambda) CS(s)$$

s : a subtopic of the topic
S_{all} : a set of all subtopics of the topic
S_{front} : a set of the front-match-subtopics
S_{back} : a set of the back-match-subtopics
ID_s : a set of session IDs for s
OQ_s : a set of original queries in the refined query log for s

$\#ID_s$: the number of elements in ID_s
$\#OQ_s$: the number of elements in OQ_s
URL_s : a set of clicked URLs for s
$\#URL_s$: the number of elements in URL_s
λ, 1- λ : weights for equations

Result

We mined subtopics for 100 Chinese topics. We used only the given query log data. The λ of KLE-S-C-1 was 0.7, the λ of KLE-S-C-2 was 0.5 and the λ of KLE-S-C-3 was 0.3.

Run	*l*	I-rec@*l*	D-nDCG@*l*	D#-nDCG@*l*
KLE-S-C-1	10	0.3162	0.4466	0.3814
	20	0.4443	0.4326	0.4385
	30	0.4769	0.3709	0.4239
KLE-S-C-2	10	0.3162	0.4464	0.3813
	20	0.4443	0.4329	0.4386
	30	0.4769	0.3712	**0.4241**
KLE-S-C-3	10	0.3185	0.4461	**0.3823**
	20	0.4482	0.4344	**0.4413**
	30	0.4776	0.3677	0.4226

Discussion

The clearness score was more useful than the interest score in the task because KLE-S-C-3 which had a high-weight for the clearness score was the best of our runs for l = 10 and 20. However, the evaluated values were too low. There were few full-match-subtopics for the given topics. Furthermore, we could not mine subtopics for four topics in 100 Chinese topics. Although the number of ideal subtopics in the task was 10,000, we just mined 6,256 subtopics. The reasons of these problems were limited resources and the simple method that generated subtopics and estimated scores of the subtopics.