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

Se-Jong Kim
Pohang University of Science and Technology (POSTECH)
sejong@postech.ac.kr

Hwidong Na
Pohang University of Science and Technology (POSTECH)
leona@postech.ac.kr

Jong-Hyeok Lee
Pohang University of Science and Technology (POSTECH)
jhlee@postech.ac.kr

ABSTRACT

This paper describes our subtopic mining system for the NTCIR-9 INTENT task. We propose a method that mines subtopics for each topic only using the given Chinese query log. Our method finds possible subtopics and estimates scores of them based on interest and clearness. In the Chinese subtopic mining, our best values of D#-nDCG were 0.3823 for l = 10, 0.4413 for l = 20 and 0.4241 for l = 30.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

intent, subtopic, ambiguity, diversity

Team Name

[KLE][POSTECH]

Subtasks/Languages

[Chinese Subtopic Mining]

1. INTRODUCTION

An intention 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. If queries are ambiguous, users may seek for different senses, and if queries are broad, users may get interested in different aspects.

As one of the solutions for these problems, there is subtopic mining that mines users' underlying subtopics of a topic (or query). Subtopics are senses of a topic and aspects of each sense. If a topic is "windows," senses are "Microsoft Windows" and "house windows," and aspects of "Microsoft Windows" are "Windows7" and "Windows update." Subtopic mining is useful to improve the results of various search scenarios like result diversification and query suggestion (Figure 1).

The related approaches of subtopic mining are query expansion, term selection, query suggestion and term disambiguation. In the query expansion approach, Xu, Croft [1], Lam-Adesina and Jones [2] found keywords from top-ranked documents retrieved by the initial query. In the term selection approach, Carpineto et al. [3] selected keywords that maximized the divergence between language model defined by top-ranked documents and that defined by entire documents. In the query suggestion approach, Huang et al. [4] selected query pairs that frequently co-occur in same search sessions, and Beeferman, Berger [5] and Baeza-Yates et al. [6] found similar queries which share a large number of clicked URLs from click-through data. In the term disambiguation approach, Sanderson [7] and Song et al. [8] extracted term lists from Wikipedia disambiguation pages.

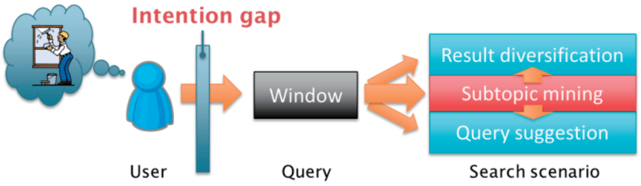

Figure 1. Intention gap and subtopic mining.

This paper describes our subtopic mining system for the NTCIR-9 INTENT task. We propose a method that mines subtopics for each topic from the given Chinese query log (SogouQ). We do not use any document collections, commercial search engines and other resources. The proposed method finds possible subtopics and estimates scores of them based on interest and clearness. A description of the proposed method is given in Section 2. Our results are given in Section 3. Section 4 discusses the results and concludes this paper.

2. METHOD

2.1 Overview

Our method consisted of three steps. The first step was to refine the query log because there were unnecessary fields and characters. The second step was to find possible subtopics which would appear fully or partially in the refined query log. The final step was to estimate scores of subtopics using some assumptions for the interest and clearness of them.

2.2 Refinement of Data

The query log data SogouQ includes session IDs, queries, clicked URLs, click-sequences for the clicked URLs within the session, date/time and ranks in the result lists. Our method does not need the data for date/time and ranks. If the value of the click-sequence is not 1 or 2, this data line is not useful in the proposed method. Therefore, we deleted the fields of date/time and ranks from the original query log, and remained the data lines whose click-sequences are 1 or 2. We also removed unnecessary characters like '(,' '),' '{,' '},' '[,' '],' '_,' '+' and the white space from the query log.

2.3 Finding Possible Subtopics

We found possible subtopics for each topic. First of all, we divided the topic into bigrams. 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. We called the type of this subtopic the full-match-subtopic. If the query matched up to the front bigram of the topic, we extracted the text string that appeared before this front bigram from the query and attached the extracted text string to the front of the topic. We called the type of this generated subtopic the front-match-subtopic. If the query matched up to the back bigram of the topic, we extracted the text string that appeared after this back bigram from the query and attached the extracted text string to the back of the topic. We called the type of this generated subtopic the back-match-subtopic (Figure 2).

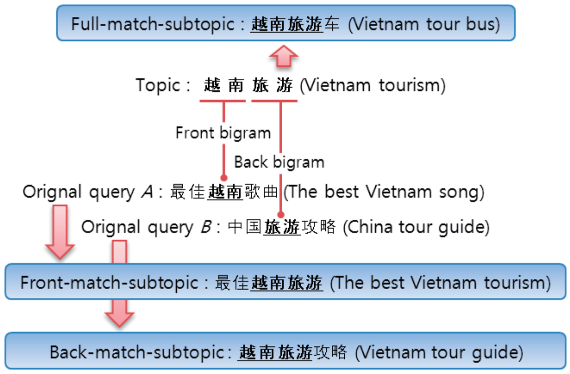
Full-match-subtopic : 越南旅游车 (Vietnam tour bus)
Topic : 越 南 旅 游 (Vietnam tourism)
Front bigram
Back bigram
Orignal query A : 最佳越南歌曲 (The best Vietnam song)
Orignal query B : 中国旅游攻略 (China tour guide)
Front-match-subtopic : 最佳越南旅游 (The best Vietnam tourism)
Back-match-subtopic : 越南旅游攻略 (Vietnam tour guide)

Figure 2. Examples of possible subtopics.

2.4 Estimating Scores of Subtopics

We estimated scores of the found subtopics using some equations based on interest and clearness. There were two assumptions. The first assumption was that a score of a subtopic increased by the interest. The interest was measured by the frequency of the subtopic that had a different session ID. The second assumption was that a score of a subtopic increased by the clearness which was measured by distinct clicked URLs that satisfied the click-sequence 1 or 2.

From the first assumption, we defined the interest factor equation (*Int*) and the interest score equation (*IS*) as:

$$Int(s) = \begin{cases} \frac{\sum_{s' \in OQ_s} \#ID_{s'}}{\#OQ_s}, & If\ s \in S_{front}\ or\ s \in S_{back} \\ \#ID_s & ,\ otherwise \end{cases}, \quad (1)$$

$$IS(s) = \frac{Int(s)}{\sum_{p \in S_{all}} Int(p)}, \quad (2)$$

where *s* is a subtopic of the given topic; S_{all} is a set of all subtopics of the given topic; S_{front} is a set of the front-match-subtopics; S_{back} is a set of the back-match-subtopics; ID_s is a set of session IDs for the subtopic *s*; OQ_s is a set of original queries in the refined query log for the subtopic *s*; $\#ID_s$ is the number of elements in ID_s and $\#OQ_s$ is the number of elements in OQ_s. In (1) and (2), we got the interest factors for each type of subtopics and normalized the scores. If the value of *IS* for some subtopic is smaller than 2, we deleted the subtopic from the results.

From the second assumption, we defined the ambiguity factor equation (*Amb*) and the clearness score equation (*CS*) as:

$$Amb(s) = \begin{cases} \frac{\sum_{s' \in OQ_s} \#URL_{s'}}{\#OQ_s}, & If\ s \in S_{front}\ or\ s \in S_{back} \\ \#URL_s & ,\ otherwise \end{cases}, \quad (3)$$

$$CS(s) = \frac{1/Amb(s)}{\sum_{p \in S_{all}} (1/Amb(p))}, \quad (4)$$

where URL_s is a set of clicked URLs that satisfied the click-sequence 1 or 2 for the subtopic *s* and $\#URL_s$ is the number of elements in URL_s. We got the ambiguity factors for each type of subtopics from (3). In (4), we considered the reciprocal number of the ambiguity factor as the clearness factor and normalized the scores.

To combine the interest score and the clearness score, we defined the subtopic score equation (*Score*) as:

$$Score(s) = \lambda IS(s) + (1 - \lambda) CS(s), \quad (5)$$

where λ and 1- λ are weights for equations.

3. RESULT

3.1 Overview

We mined subtopics for 100 Chinese topics. We used only the given query log data SogouQ that was one-month queries and click-through data in 2008. Our run names were "KLE-S-C-1," "KLE-S-C-2" and "KLE-S-C-3." We got 6,256 subtopics for 96 topics and estimated scores of these subtopics through changing weights for the proposed equations (Figure 3). 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.

0005;0;红酒面膜 (red wine mask);1;0.0722393652;Run1

0005;0;红酒知识 (red wine knowledge); 2;0.0241642789;Run1

0005;0;法国红酒 (French red wine);3;0.0221489971;Run1

0005;0;红酒面膜屈臣氏 (red wine mask Watson); 4;0.0181470869;Run1

0005;0;自制红酒面膜 (homemade red wine mask); 5;0.0121776504;Run1

Figure 3. The sample of KLE-S-C-1 for 红酒 (red wine).

3.2 Chinese Subtopic Mining Results

To evaluate the result, we used I-rec which measured diversity, D-nDCG which measured overall relevance across intents, and D#-nDCG which was a simple average of I-rec and D-nDCG [9, 10]. The D#-nDCG@10 values of KLE-S-C-1, KLE-S-C-2 and KLE-S-C-3 were 0.3814, 0.3813 and 0.3823 respectively (Table 1). The D#-nDCG@20 values were 0.4385, 0.4386 and 0.4413 (Table 2). The D#-nDCG@30 values were 0.4239, 0.4241 and 0.4226 (Table 3). Our best values of D#-nDCG@10, D#-

nDCG@20 and D#-nDCG@30 were 0.3823 of KLE-S-C-3, 0.4413 of KLE-S-C-3 and 0.4241 of KLE-S-C-2 respectively [10].

Table 1. The mean intent recall, D-nDCG and D#-nDCG values for *l* = 10 for three runs

Run	I-rec@10	D-nDCG@10	D#-nDCG@10
KLE-S-C-1	0.3162	0.4466	0.3814
KLE-S-C-2	0.3162	0.4464	0.3813
KLE-S-C-3	0.3185	0.4461	0.3823

Table 2. The mean intent recall, D-nDCG and D#-nDCG values for *l* = 20 for three runs

Run	I-rec@20	D-nDCG@20	D#-nDCG@20
KLE-S-C-1	0.4443	0.4326	0.4385
KLE-S-C-2	0.4443	0.4329	0.4386
KLE-S-C-3	0.4482	0.4344	0.4413

Table 3. The mean intent recall, D-nDCG and D#-nDCG values for *l* = 30 for three runs

Run	I-rec@30	D-nDCG@30	D#-nDCG@30
KLE-S-C-1	0.4769	0.3709	0.4239
KLE-S-C-2	0.4769	0.3712	0.4241
KLE-S-C-3	0.4776	0.3677	0.4226

4. DISCUSSION AND CONCLUSION

This paper proposed a method that mined subtopics from the query log using the interest score and the clearness score. In the Chinese subtopic mining, our best values of D#-nDCG were 0.3823 for *l* = 10, 0.4413 for *l* = 20 and 0.4241 for *l* = 30. 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 any subtopics for 减肥粥 (porridge diet), 十二生肖来历 (Zodiac origin), 植树节的来历 (Arbor Day origin) and 张含韵 (Zhang Han Yun) 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. Therefore, to overcome the data sparseness problem and the simple algorithm, we will use various resources or other query logs that were collected long-period, and have a full study about subtopic mining.

5. ACKNOWLEDGMENTS

This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST No. 2011-0017960), and in part by the BK 21 Project in 2011.

6. REFERENCES

[1] J. Xu and W. B. Croft. Query Expansion using Local and Global Document Analysis, In *Proceedings of ACM SIGIR 1996*, pages 4-11, 1996.

[2] A. M. Lam-Adesina and G. J. F. Jones. Applying Summarization Techniques for Term Selection in Relevance Feedback, In *Proceedings of ACM SIGIR 2001*, pages 1-9, 2001.

[3] C. Carpineto, R. De Mori, G. Romano and B. Bigi. An Information-Theoretic Approach to Automatic Query Expansion, *ACM Transactions on Information Systems*, Vol. 19, No. 1, pages 1-27, 2001.

[4] C. K. Huang, L. F. Chien and Y. J. Oyang. Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs, *Journal of the American Society for Information Science and Technology*, Vol. 54, Issue 7, pages 638-649, 2003.

[5] D. Beeferman and A. Berger. Agglomerative Clustering of a Search Engine Query Log, In *Proceedings of ACM SIGKDD* 2000, pages 407-416, 2000.

[6] R. Baeza-Yates, C. Hurtado and M. Mendoza. Query Recommendation Using Query Logs in Search Engines, *EDBT 2004*, LNCS 3268, pages 588-596, 2004.

[7] M. Sanderson. Ambiguous Queries: Test Collections Need More Sense, In *Proceedings of ACM SIGIR 2008*, pages 499-506, 2008.

[8] R. Song, D. Qi, H. Liu, T, Sakai, J. Y. Nie, H. W. Hon and Y. Yu. Constructing a Test Collection with Multi-Intent Queries, *EVIA 2010*, pages 51-59, 2010.

[9] T. Sakai and R. Song. Evaluating Diversified Search Results Using Per-Intent Graded Relevance, In *Proceedings of ACM SIGIR 2011*, pages 1043-1052, 2011.

[10] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang and N. Orii. Overview of the NTCIR-9 INTENT Task, 2011.