SEM12 at the NTCIR-10 INTENT-2 English Subtopic Mining Subtask

Md. Zia Ullah
Dept. of Computer Science and Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-Cho, Toyohashi,
441-8580, Aichi, Japan
arif@kde.cs.tut.ac.jp

Masaki Aono†
Dept. of Computer Science and Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-Cho, Toyohashi,
441-8580, Aichi, Japan
aono@tut.jp

Md. Hanif Seddiqui*
Dept. of Computer Science and Engineering,
University of Chittagong,
Hathazari-4331, Chittagong,
Bangladesh
hanif@cu.ac.bd

ABSTRACT

Users express their information needs in terms of queries in search engines to find some relevant documents on the Internet. However, search queries are usually short, ambiguous and/or underspecified. To understand user's search intent, subtopic mining plays an important role and has attracted attention in the recent years. In this paper, we describe our approach to identifying, and then ranking user's intents for a query (or topic) from query logs, which is an english subtopic mining subtask of the NTCIR-10 Intent-2 task. We extract subtopics that are semantically and lexically related to the topic, and measure their weights based on co-occurrence of a subtopic across search engine query logs, and edit distance between a topic and a subtopic. These weighted subtopic strings are ranked to represent themselves as the candidates of subtopics (or intents). In the experiment section, we show the revised subtopic mining results of our method evaluated by the organizers. The best performance of our system achieves an I-rec@10 (Intent Recall) of 0.3780, a *D-nDCG@10* of 0.4250, and a *D#-nDCG@10* of 0.4014.

Team Name

SEM12

Subtasks

English Subtopic Mining

Keywords

intents, subtopics, query logs, co-occurrence, edit distance

1. INTRODUCTION

Users are used to using search engines to find information on the Web. When an information-need is being formulated in users' mind, queries in the form of a sequence of words will be typed into the search box, ideally, the search engine should respond with a ranked list of snippet results that best meets the needs of users. A query is classified into two types, one is "faceted" and the other is "ambiguous". The search intent of faceted queries is usually clear, so that the search engine can report good quality results. However, information retrieval systems often fail to capture users' search intents exactly if a submitted query is ambiguous. Because an ambiguous query has more than one interpretation and different users have different intents for the same query, which corresponds to different subtopics. For example, for an ambiguous query "tiger", the search engine should be able to collect documents that are highly/marginally relevant to the intents "tiger woods", "tiger airways", "tiger animal (specially royal bengal tiger)", "tiger company products", and "tiger oracle database". Moreover, if the search query log suggests that users are more likely to search for "tiger woods" than for "tiger animal" and others, the search engine may choose to return documents relevant to the former than ones relevant to the laters. As a result, it has been recognized as a crucial part of effective information retrieval to understand users' information needs or intents that underlies the submitted query and diversify the results retrieved for ambiguous query, maximizing the satisfaction of users with different intents.

The INTENT task in NTCIR-10 is dealing with the above problem via two subtasks. The first subtask is how to mine the underlying intents/subtopics, and the second subtask is how to selectively diversify search results. We participated in the former subtask, which is also known as english subtopic mining subtask of the NTCIR-10 Intent-2 Task [8] and propose a method to mine the subtopics of each query issued by users. The remainder of this paper is organized as follows: section 2 describes the systematic review of the related work while NTCIR Subtopic Mining Subtask is defined in section 3. We introduced our approach in section 4. Section 5 includes the overall experiments and the results we obtained. Finally, concluding remarks and some future directions of our work are described in Section 6.

2. RELATED WORK

Queries are usually short, ambiguous and/or underspecified. To perceive the meanings of queries, researchers define taxonomies and classify queries into predefined categories. Song et al. divided queries into 3 categories [9]: ambiguous queries, which have more than one meaning; board queries, which covers a variety of subtopics; and clear queries, which have a specific meaning or narrow topics. At the query level, Broder [2] divided query intent into navigational, informational and transactional types. Nguyen and Kan [5] classified queries into four general facets of ambiguity, authority,

temporal sensitivity and spatial sensitivity. Boldi et al. [1] created query-flow graph with query phrase nodes and used them for query recommendation. Query suggestion or query recommendation is a key technique for generating alternative queries to help users drill down to a subtopic of the original query [10, 4]. Different from query suggestion or query completion, subtopic mining focuses more on the diversity of possible subtopics of the original query rather than inferring relevant queries. Jian Hu [3] integrated the knowledge contained in Wikipedia to predict the possible intents for a given query. A number of intent seed are iteratively propagated through Wikipedia structure with Markov random walk. Filip Radlinkshi [7] proposed an approach for inferring query intents from reformulations and clicks. For an input query, the click and reformulation information are combined to identify a set of possible related queries to construct an undirected graph. An edge is introduced between two queries if they were often clicked for the same documents. Finally, random walk similarity is used to find intent cluster. At the session level, Radlinski and Joachims [6] mined intent from query chains and used it for learning to rank algorithm.

3. NTCIR SUBTOPIC MINING SUBTASK

The NTCIR-10 Intent-2 subtopic mining task is motivated to encourage research on developing and evaluating algorithms to find query intents. The task is defined as follows:

• for a given query, the system should return a ranked list of possible "subtopic strings" that covers as many search intents as possible.

What is a subtopic string? *A subtopic string of a given query is a query that specializes and/or disambiguates the search intent of the original query. If a string returned in response to the query does neither, it is considered incorrect.*

e.g.
original query: "apple" (ambiguous)
subtopic string: "apple iPhone 5"
incorrect: "apple apple" (does not disambiguate; does not specialize)

e.g.
original query: "tutorial on programming" (underspecified)
subtopic string: "tutorial on programming in java"
incorrect: "tutorial programming" (does not specialize)

4. OUR APPROACH

In the subtopic mining subtask, we utilize the query log as the only information resource. Our assumption about search engine query is that *some intents are more likely than others.* This assumption is implemented by using the co-occurrence frequency of subtopics across query logs. Our method consist of two steps:

• Firstly, aggregate the subtopics and find the co-occurrence frequency of the subtopics for each topic from all four search query logs.

• Secondly, rank the subtopics by sorting in descending order based on the co-occurrence frequency of subtopics, and if there is a tie, then apply the Edit Distance between a subtopic and the original topic to identify the most relevant subtopic for finding the relevant query intents.

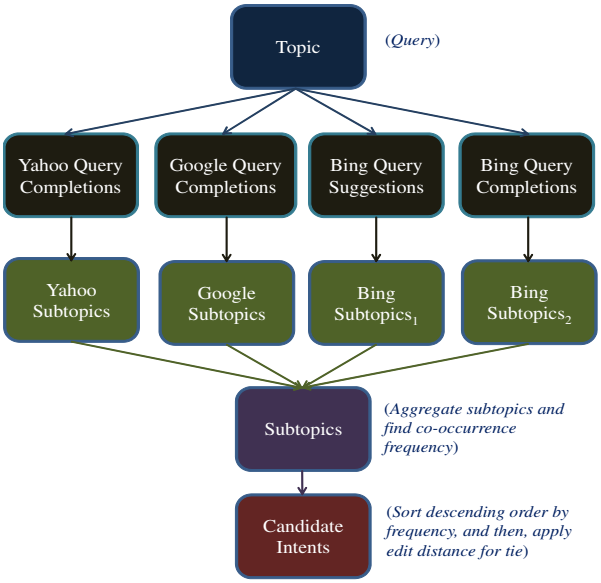

Figure 1: Subtopic mining workflow

Our approach is depicted in fig. 1. It is shown that subtopics with their co-occurrence frequencies are measured from the query logs for each topic. Ranked list of candidate intents are found by sorting the subtopics based on their frequencies, and if there is a tie, then apply Edit Distance between each subtopic string and the original topic string.

The whole procedures of subtopic mining is articulated in algorithm 1. In this algorithm, there are two major steps. In the first step, all the subtopics for a single topic are aggregated from all four search query logs. Then, we finds the co-occurrence frequency of each subtopic by exact string matching. In this step, hashing is used for counting the frequency of subtopics. In the second step, for making the ranked list of subtopics, we sort the subtopics based on their co-occurrence frequency, and if there is tie, then we apply the Edit Distance between a subtopic string and the original topic string.

Given two character strings s_1 and s_2, the Edit Distance between them is the minimum number of edit operations required to transform s1 into s2. The edit operations allowed in our system are:
(1) Insert a character into a string
(2) Delete a character from a string
(3) Replace a character of a string by another character

4.1 Ranking Subtopics

Our system ranks the subtopics of each topic for generating ranked list of query intents. After measuring the co-occurrence frequency of all subtopics for a single topic, we sort the subtopics in descending order of frequency. If the frequency is same for two subtopics string st_1 and st_2, we consider the Edit Distance between each subtopic st and the original topic t. If EditDistance of (st_2, t) is lower than

Algorithm 1: Subtopic Mining($\mathcal{T}$,$\mathcal{QL}$)
A naïve algorithm for aggregating and mining subtopics

Input: Topic ($\mathcal{T}$), Search Engine Query Logs ($\mathcal{QL}$)
Output: Candidate Intents ($\mathcal{CI}$)

```
   /* Aggregate subtopics                                   */
1  Subtopics ←getAllSubtopics(QL,T)
2  CI ← ∅
3  ST ← ∅
   /* Co-occurrence frequency                               */
4  for subtopic st_i ∈ Subtopics do
5      (st_i, freq) ← getScorePair(ST, st_i)
6      if freq is null then
7          putScorePair (ST, (st_i, 1))
8      else
9          putScorePair (ST, (st_i, freq+1))
   /* Rank the subtopics                                    */
10 CI ← Sort ST in descending order using frequency, and
     if there is a tie, then apply edit distance between a
     subtopic st and topic T
11 return CI
```

(st_1, t), we rank the st_2 in top than st_1.

5. EXPERIMENTS

5.1 Runs

We submitted 5 runs for this task.

(1) We aggregated intent candidates from all search engine query logs introduced in 4, and sort the intent candidates in dictionary order.

(2) We aggregated intent candidates from all search engine query logs, and sort the intent candidates based on the co-occurrence frequency.

(3) Similar to (2), but during sorting the intent candidates, if there is a tie, then we apply dictionary ordering between the subtopics.

(4) Similar to (1), but we sort the intent candidates based on Edit Distance between intent candidate and the original query.

(5) Similar to (4), but during sorting the intent candidates, if there is a tie, then we choose the intent candidate with higher co-occurrence frequency than other.

5.2 Experimental Results

In fig. 3, the revised experimental results of our runs for top 10 intent candidates is depicted, and the revised experimental results of all participants is depicted in fig. 2. Our best values of *D#-nDCG@10* is 0.4014 for SEM12-S-E-2A, *D-nDCG@10* is 0.4250 for SEM12-S-E-2A, and *I-rec@10* is 0.3780 for SEM12-S-E-1A. From the obtained results in fig. 3, we can draw the following conclusions for the proposed approach: (1) *D#-nDCG@10*, *D-nDCG@10*, and *I-rec@10* are average compared with the top results. One reason is

run name	I-rec@10	D-nDCG@10	D#-nDCG@10
THUIR-S-E-4A	0.4364	0.5062	0.4713
THUIR-S-E-1A	0.4512	0.4775	0.4644
THUIR-S-E-5A	0.4253	0.4893	0.4573
THUIR-S-E-2A	0.4333	0.4795	0.4564
THCIB-S-E-1A	0.4431	0.4657	0.4544
THUIR-S-E-3A	0.4346	0.4726	0.4536
THCIB-S-E-2A	0.4308	0.4744	0.4526
hultech-S-E-1A	0.3680	0.5368	0.4524
KLE-S-E-4A	0.4457	0.4401	0.4429
THCIB-S-E-3A	0.4248	0.4557	0.4403
THCIB-S-E-4A	0.4100	0.4521	0.4310
THCIB-S-E-5A	0.4144	0.4441	0.4292
hultech-S-E-4A	0.3688	0.4807	0.4248
KLE-S-E-2A	0.4292	0.4159	0.4225
SEM12-S-E-2A	**0.3777**	**0.4250**	**0.4014**
SEM12-S-E-1A	**0.3780**	**0.4233**	**0.4007**
ORG-S-E-4A	0.3815	0.3829	0.3822
ORG-S-E-3A	0.3841	0.3735	0.3788
KLE-S-E-3A	0.3676	0.3661	0.3668
SEM12-S-E-4A	**0.3727**	**0.3471**	**0.3599**
SEM12-S-E-5A	**0.3659**	**0.3445**	**0.3552**
KLE-S-E-1A	0.3529	0.3540	0.3535
SEM12-S-E-3A	**0.3403**	**0.3573**	**0.3488**
ORG-S-E-5A	0.3181	0.3365	0.3273
ORG-S-E-2A	0.3268	0.3231	0.3250
hultech-S-E-3A	0.3045	0.3345	0.3195
ORG-S-E-1A	0.2787	0.3068	0.2927
hultech-S-E-2A	0.2697	0.2986	0.2841
TUTA1-S-E-1A	0.2181	0.2577	0.2379
LIA-S-E-4A	0.2000	0.2753	0.2376
TUTA1-S-E-2A	0.1865	0.2327	0.2096
LIA-S-E-2A	0.0328	0.0474	0.0401
LIA-S-E-1A	0.0291	0.0420	0.0355
LIA-S-E-3A	0.0377	0.0329	0.0353

Figure 2: English Subtopic Mining runs ranked by mean D#-nDCG@10 over 50 topics. Our runs are shown in bold.

Runs	I-rec@10	D-nDCG@10	D#-nDCG@10
SEM12-S-E-1A	**0.3780**	0.4233	0.4007
SEM12-S-E-2A	0.3777	**0.4250**	**0.4014**
SEM12-S-E-3A	0.3403	0.3573	0.3488
SEM12-S-E-4A	0.3727	0.3471	0.3599
SEM12-S-E-5A	0.3659	0.3445	0.3552

Figure 3: Evaluation results of subtopic mining runs. The highest value in each column is shown in bold.

that our proposed approach has less merits to utilize the co-occurrence frequency of subtopics. Another reason is that we rely mainly on query logs. Moreover, other resources might also be integrated.

6. CONCLUSIONS

This paper described an approach to identifying candidate user's intents from search engine query logs. Firstly, we extracted the subtopics which are semantically and lexically related to the original query, and measured their weights based on the co-occurrence frequency of subtopics. Secondly, the subtopics are ranked by sorting in descending order of frequency, and if there is a tie, then apply the Edit Distance between a subtopic string and the original query string to identify the most relevant subtopic for finding the query intents. Our system achieves an average results in the official evaluation, especially on the *D#-nDCG@10* metric. However, we need to improve the overall relevance performance across intents. In the future, we will improve the subtopic ranking algorithm and introduce more features to help ranking subtopics. We will also try to apply clustering algorithms, and use ensemble learning methods to combine together which might be more effective. Some other semantic similarity measures will be tried by organizing these subtopics into hierarchy structure according to their semantic relationships. Future directions also include how to integrate more knowledge resources into the system further, such as wikipedia, and how to extend this work to diversify web search results with taxonomies like Open Directory Project(ODP).

7. ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid(C) 23500119.

8. REFERENCES

[1] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 609–618. ACM, 2008.

[2] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[3] J. Hu, G. Wang, F. Lochovsky, J.-t. Sun, and Z. Chen. Understanding user's query intent with wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 471–480. ACM, 2009.

[4] Q. Mei, D. Zhou, and K. Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 469–478. ACM, 2008.

[5] B. V. Nguyen and M.-Y. Kan. Functional faceted web query analysis. In *WWW2007: 16th International World Wide Web Conference*, 2007.

[6] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248. ACM, 2005.

[7] F. Radlinski, M. Szummer, and N. Craswell. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*, pages 1171–1172. ACM, 2010.

[8] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, R. Song, M. P. kato, and M. Iwata. Overview of the ntcir-10 intent-2 task. *In Proc. of NTCIR*, pages 1–26, 2013.

[9] R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying ambiguous queries in web search. In *Proceedings of the 16th international conference on World Wide Web*, pages 1169–1170. ACM, 2007.

[10] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web*, pages 1039–1040, 2006.