The KLE's Subtopic Mining System for the NTCIR-10 INTENT-2 Task

Se-Jong Kim
Pohang University of Science and Technology (POSTECH)
sejong@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-10 INTENT-2 task. We propose a method that mines subtopics using simple patterns and the hierarchical structure of candidate strings based on the clusters of relevant documents using the provided web documents and official query suggestions. We extracted various candidate strings using simple patterns based on new query-types and POS tags. We constructed the hierarchical structure of the candidate strings according to the proposed process, and ranked them as subtopics using the document coverage and frequency information for each group of the candidate strings in the area satisfying the diversity requirement of the hierarchical structure.

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

General Terms
Experimentation

Keywords
intent, subtopic, diversity, pattern, hierarchical structure

Team Name
[KLE][POSTECH]

Subtasks
[English Subtopic Mining][Japanese Subtopic Mining]

1. INTRODUCTION
Many web queries are short and unclear. Some users do not choose appropriate words for a web search, and others omit specific terms needed to clarify search intents, because it is not easy for users to express their search intents explicitly through keywords. This intention gap between users' search intents and queries results in queries which are ambiguous and broad. For ambiguous queries, users may get results quite different from their intents; for broad queries, results may not be as specific as users expect.

As a solution for these problems, subtopic mining is proposed, which can find the possible subtopics (subtopic strings) for a given query and return a ranked list of them in terms of the relevance to the query, popularity and diversity of subtopics (Figure 1). According to the NTCIR-9 subtopic mining task [1], a subtopic of a given query is a query that specifies and disambiguates the search intent of the original query. For example, if a query is "chocolate," its specific hyponyms "valentine chocolate" and "white chocolate" can be subtopics. Subtopic mining can be used to improve the results of various search scenarios, such as query suggestion and result diversification.

Figure 1. A flow from a user's intent to subtopic mining.

The NTCIR-9 subtopic mining task motivated various methods for the Chinese and Japanese languages. To achieve high-level performance, some methods [2-5] used suggested queries from major web search engines (Baidu, Bing, Google, and Yahoo), and some others [6-8] used top-ranked documents obtained from search engines. In addition to the resources provided, query logs, web documents, and online encyclopedias were used [1, 4, 5, 7], and anchor texts and URLs were utilized [1, 6].

This paper describes our subtopic mining system for the NTCIR-10 INTENT-2 task. We propose a method that mines subtopics using simple patterns and the hierarchical structure of candidate strings based on the clusters of relevant documents. We use only the provided web documents and official query suggestions for the English and Japanese languages. To reduce data sparseness, we extract fully or partially original strings as candidate strings from the web documents using the simple patterns. To automatically decide the scope of subtopics, we first select more relevant subtopics (disambiguated and initially specified search intents) using the hierarchical structure of candidate strings, and exclude the subtopics that point to non-existent information using the document coverage.

A description of the proposed method is given in Section 2. In Section 3, our results are presented, and in the final section, we give the discussion and conclusion.

2. METHOD

2.1 Overview
Our method consisted of three steps. The first step was to extract candidate strings. We made new query-types from the original query, and found various candidate strings using simple patterns based on each query-type and POS tags. The second step was to construct the hierarchical structure of candidate strings. We proposed a three-depth hierarchical structure of candidate strings

to select more relevant subtopics with the high popularity and improve the diversity of subtopics. The third step was to rank candidate strings as subtopics. We ranked candidate strings using the document coverage and frequency information for each group of the candidate strings in the area satisfying the diversity requirement of the hierarchical structure.

2.2 Extracting Candidate Strings

We can specify words using other words that co-occurred with the target words. Because a subtopic is a specified string (words) that reflects a user's search intent when inputting the query, we can also find candidate strings as subtopics using several words that co-occurred with the query in the given documents. From this characteristic of subtopics, we created simple patterns to extract candidate strings:

• *Pattern 1*: (adjective)?(noun)+(non-noun)*(query)(non-noun)*(adjective)?(noun)+

• *Pattern 2*: (query)(non-noun)*(adjective)?(noun)+

• *Pattern 3*: (adjective)?(noun)+(non-noun)*(query)

where the ? operator indicates there is zero or one preceding element; the + operator indicates there are one or more preceding elements; and the * operator indicates there are zero or more preceding elements. Because *Patterns 1-3* covered original phrases with the whole query and modifiers in the documents, we could find relevant and understandable candidate strings by applying these patterns sequentially from the top 1,000 relevant documents retrieved by BM25 model in English [9], and the non-diversified baseline Document Ranking run in Japanese [10]. In particular, we generated several web documents in which the *i*-th item of each official query suggestion appeared $11 - i$ times, and extracted candidate strings using the patterns from these documents.

However, if a query consisted of more than two words, we could not thoroughly extract various candidate strings using only *Patterns 1-3* from the documents, because the number of candidate strings that fully matched the query decreased. Therefore, we made two new query-types from the original query to extract various candidate strings. One was q_{left} and the other was q_{right}. For phrases which were the remaining words after consecutively removing the right words of the query, we retrieved the top 200 relevant documents for each, and compared these documents with the top 200 relevant documents for the original query. If the relevant documents for one of the phrases covered more than 100 documents in the relevant documents for the original query, we considered this phrase as an alternative query. Among alternative queries that covered the most documents, we selected the shortest alternative query as new query q_{left}. If none of the phrases satisfied this condition, we selected the longest phrase as q_{left}. For phrases which were the remaining words after consecutively removing the left words of the query, we selected q_{right} by applying the process mentioned above. Using q_{left} and q_{right}, we created new simple patterns:

• *Pattern 4*: (adjective)?(noun)+(non-noun)*($\underline{q_{left}}$)(word)*($\underline{q_{right}}$)(non-noun)*(adjective)?(noun)+

• *Pattern 5*: ($\underline{q_{left}}$)(word)*($\underline{q_{right}}$)(non-noun)*(adjective)?(noun)+

• *Pattern 6*: (adjective)?(noun)+(non-noun)*($\underline{q_{left}}$)(word)*($\underline{q_{right}}$)

• *Pattern 7*: ($\underline{q_{right}}$)(non-noun)*(adjective)?(noun)+

• *Pattern 8*: (adjective)?(noun)+(non-noun)*($\underline{q_{left}}$)

We found various candidate strings using these new patterns, and replaced the parts of candidate strings corresponding to the underlined patterns with the original query. Even if these candidate strings that partially matched the query were not very relevant to the query, we could reduce data sparseness.

Lastly, to avoid redundancy of candidate strings, we merged them based on the meaningful keyword sets. We regarded the lemmas of noun phrases at the start or end of each candidate string as the important words that decided the meaning of the candidate string. We named the set of these words the meaningful keyword set for each candidate string. If the meaningful keyword sets of candidate strings contained identical words, these candidate strings were similar and we merged them. At this time, we selected the shortest fully or partially original string for each meaningful keyword set as the only candidate string.

2.3 Constructing Hierarchical Structure

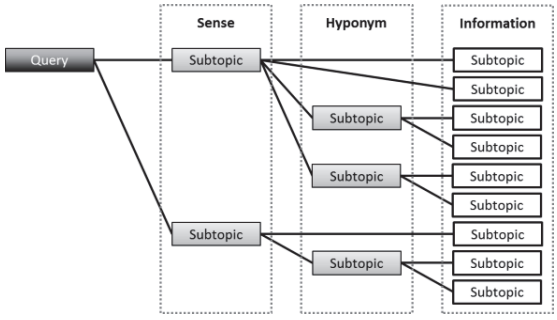

Figure 2. The ideal hierarchical structure of subtopics.

Subtopics of a given query have their hierarchical structure (Figure 2). If a query is ambiguous, more than two senses can be subtopics of the query and these subtopics are semantic children of the query in the hierarchical structure. Hyponyms of senses and titles for related information of senses can be subtopics of the query, and these subtopics are children of senses. Titles for related information of hyponyms can be also subtopics of the query, and these subtopics are children of hyponyms. For these reasons, subtopics of the query can be repeatedly increased.

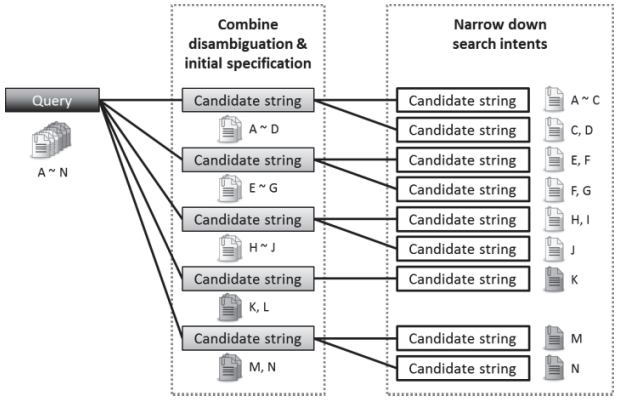

Figure 3. The hierarchical structure of candidate strings and documents extracting them.

To decide the appropriate mining scope of subtopics, we constructed the three-depth hierarchical structure of extracted candidate strings. We proposed the simplified hierarchical

structure (Figure 3) that placed candidate strings corresponding to more relevant subtopics, which are disambiguated and initially specified search intents such as senses and hyponyms, in the same level of the hierarchical structure, because we had to consider various semantic relation extraction methods based on semantic resources to construct the ideal hierarchical structure. The root was the given query, children of the root were candidate strings as more relevant subtopics, and leaves were candidate strings with a search intent similar to that of each parent.

As one child of the root, we selected the candidate string *st* satisfying the best popularity and document representativeness first, using the restricted corpus term frequency over inverse document frequency (*CTFIDF*):

$$CTFIDF(st) = freq(st, R_{query}) \cdot log \frac{|R_{query}|}{|D(st, R_{query})|}, \quad (1)$$

where R_{query} is the set of the top 1,000 relevant documents for the query; $freq(st, R_{query})$ is the frequency of *st* in R_{query}; and $D(st, R_{query})$ is the set of documents extracting *st* in R_{query}. To keep a balance between the diversity and the popularity, the first step, in which we selected the candidate string with the maximum value of (1) first, was important, because (1) could directly measure the general popularity and document representativeness of candidate strings using basic frequency information. To find other children of the root satisfying high diversity, we selected candidate strings using the cluster entropy (*CE*) [8]:

$$CE(st, P) = -\sum_{st' \in ST, st' \neq st} \frac{|D(st,P) \cap D(st',P)|}{|D(st,P)|} \cdot log \frac{|D(st,P) \cap D(st',P)|}{|D(st,P)|}, \quad (2)$$

where *P* is the set of the top 200 relevant documents for the parent query of *st*, or the set of documents extracting the parent of *st*; *ST* is the set of unselected candidate strings that appear in at least two documents in *P*; and $D(st, P)$ is the set (cluster) of documents extracting *st* in *P*. For various clusters of documents extracting candidate strings, we selected continuously optimal document clusters satisfying the diversity using (2), and returned their candidate strings. If the values of (2) for candidate strings were the same, we selected the candidate string with the largest value of (1). If the accumulated set of documents of the previously selected clusters included all documents of the newly selected cluster, we skipped the candidate string of this cluster and stopped the process for children of the root. For leaves in the hierarchical structure, we ensured the diversity of candidate strings by repeating the above process using the selected candidate strings instead of the root.

2.4 Ranking Subtopics

To rank candidate strings as subtopics in the hierarchical structure, we defined the document coverage (*DC*):

$$DC(st) = \sum_{doc \in (HR_{st} \cap HR_{query})} DocScore(doc), \quad (3)$$

where HR_{st} is the set of the top 200 relevant documents for *st*; HR_{query} is the set of the top 200 relevant documents for the query; and $DocScore(doc)$ is the ranking score of document *doc* for the query. (3) measured the popularity of candidate strings by calculating how many highly relevant documents of each candidate string covered the documents that reflected search intents with high popularity for the query. If the value of (3) for a candidate string was 0, the candidate string would be excluded.

However, (3) could have given a biased result because this measure considered only limited documents for the query and candidate strings. Therefore, to compensate for this limitation, we applied (1) considering the top 1,000 relevant documents to the final score. The score (*Score*) obtained by combining (1) and (3) was:

$$Score(st) = \frac{DC(st)}{average\ of\ DCs} + \frac{CTFIDF(st)}{average\ of\ CTFIDFs}. \quad (4)$$

We returned the ranked list of candidate strings using (4) for each group of the candidate strings in the area satisfying the diversity requirement of the hierarchical structure (Figure 4). We ranked more relevant candidate strings in the area satisfying the diversity requirement for the query first, using (4). We applied (4) sequentially from the area for the candidate string with the highest value of (4) among the more relevant candidate strings to the area for the candidate string with the lowest value of (4).

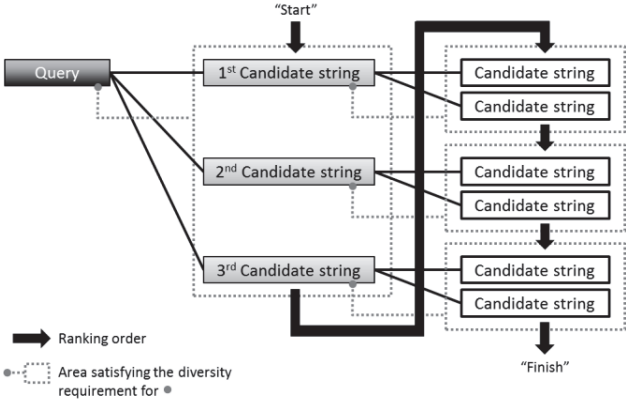

Figure 4. The ranking order and the area satisfying the diversity requirement in the hierarchical structure.

3. RESULT

3.1 Overview

We mined subtopics for 50 English queries (topics) of TREC 2012 and 100 Japanese queries of the NTCIR-10 subtopic mining task. We used only the English web document collection TREC Category B, the Japanese web document collection ClueWeb09-JA, and the official query suggestions. To perform word segmentation and identify noun phrases, we used the English Stanford POS tagger[1] and the Japanese MeCab POS tagger[2]. Our run names were "KLE-S-E(English)/J(Japanese)-1A/B," "KLE-S-E/J-2A/B," "KLE-S-E/J-3A/B," and "KLE-S-E/J-4A/B." KLE-S-E/J-1A/B used only the provided web documents and applied (3). KLE-S-E/J-2A/B also used only the provided web documents and applied (4). KLE-S-E/J-3A/B used the provided web documents and official query suggestions, and applied (3). KLE-S-E/J-4A/B used the provided web documents and official query suggestions, and applied (4). To evaluate the results, 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 [10, 11]. The number of top ranked subtopics we evaluated was $l = 10$.

[1] http://nlp.stanford.edu/software/tagger.shtml

[2] http://mecab.sourceforge.net

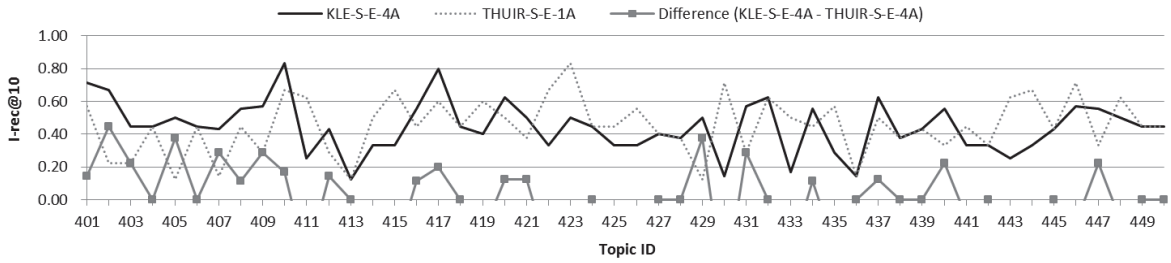

Figure 5. A comparison of I-rec@10 of KLE-S-E-4A and THUIR-S-E-1A for each topic ID.

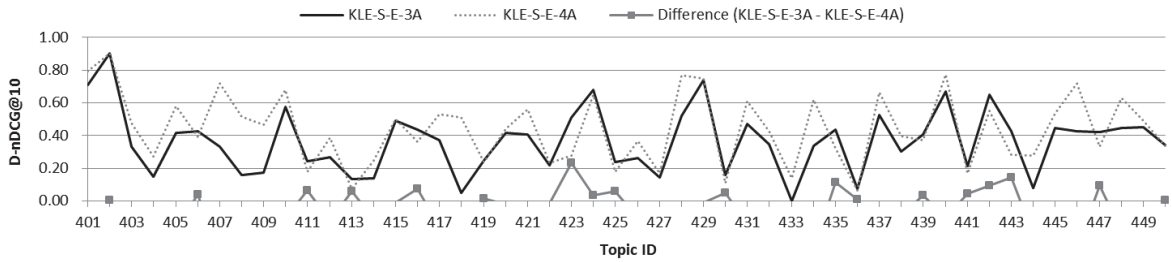

Figure 6. A comparison of D-nDCG@10 of KLE-S-E-3A and KLE-S-E-4A for each topic ID.

3.2 Revised English Subtopic Mining Results

In English, our mean D#-nDCG@10 values of KLE-S-E-1A, KLE-S-E-2A, KLE-S-E-3A, and KLE-S-E-4A were 0.3535, 0.4225, 0.3668, and 0.4429, respectively (Table 1). Our best value of mean D#-nDCG@10 was 0.4429 of KLE-S-E-4A [10]. However, the differences among our four runs in I-rec, D-nDCG, and D#-nDCG were not statistically significant, and KLE-S-E-4A also was statistically indistinguishable from THUIR-S-E-4A (top run in terms of D#-nDCG) in terms of I-rec, D-nDCG, and D#-nDCG (two-sided randomized Tukey's HSD at $\alpha = 0.05$) [10].

Table 1. Runs sorted by mean D#-nDCG for *l* = 10 in English (*revised*).

Run	Mean I-rec@10	Mean D-nDCG@10	Mean D#-nDCG@10
KLE-S-E-4A	0.4457	0.4401	0.4429
KLE-S-E-2A	0.4292	0.4159	0.4225
KLE-S-E-3A	0.3676	0.3661	0.3668
KLE-S-E-1A	0.3529	0.3540	0.3535

3.3 Revised Japanese Subtopic Mining Results

In Japanese, our mean D#-nDCG@10 values of KLE-S-J-1B, KLE-S-J-2B, KLE-S-J-3B, and KLE-S-J-4B were 0.2632, 0.1851, 0.2628, and 0.1917, respectively (Table 2). Our best value of mean D#-nDCG@10 was 0.2632 of KLE-S-J-1B [10]. KLE-S-J-1B was statistically distinguishable from KLE-S-J-2B and KLE-S-J-4B in terms of D-nDCG. However, the differences in the other parts of our four runs were not statistically significant, and KLE-S-J-1B was statistically indistinguishable from ORG-S-J-3A (top run in terms of D#-nDCG) in terms of I-rec, D-nDCG, and D#-nDCG [10].

Table 2. Runs sorted by mean D#-nDCG for *l* = 10 in Japanese (*revised*).

Run	Mean I-rec@10	Mean D-nDCG@10	Mean D#-nDCG@10
KLE-S-J-1B	0.2607	0.2656	0.2632
KLE-S-J-3B	0.2529	0.2726	0.2628
KLE-S-J-4B	0.2146	0.1687	0.1917
KLE-S-J-2B	0.2034	0.1667	0.1851

4. DISCUSSION AND CONCLUSION

This paper proposed a method that mines subtopics using simple patterns and the hierarchical structure of candidate strings using the provided resources for English and Japanese. In the English subtopic mining, our best values of mean I-rec, D-nDCG, and D#-nDCG were 0.4457, 0.4401, and 0.4429 for l = 10, respectively. In particular, the mean I-rec value placed second in the task. Furthermore, in I-rec@10 of our best run (KLE-S-E-4A), 34 queries showed a better or the same performance as the top run (THUIR-S-E-1A) in terms of I-rec (Figure 5). If intents of a query were relatively few, our method could extract various subtopics but there were several non-relevant subtopics; if intents of a query were many, our method could find various and relevant subtopics using candidate strings which partially matched the query and their hierarchical structure (Table 3). For each group of runs applying same ranking method, the run that used the official query suggestions together was better than the other because the query suggestions could give good candidate strings reflecting various search intents by query logs. Meanwhile, the runs which used only *DC* for ranking showed the lower results than others. If search results for candidate strings were similar to that for the original query, the values of *DC* increased. This occurred because useless terms, such as stop words, appeared frequently as meaningful keywords in candidate strings. Checking D-nDCG, however, we found that our method could partially ascertain queries at a higher performance level than others that used *CTFIDF* together (Figure

6). For the queries that yielded a high performance by using *DC* only, the analysis showed that gaps between subtopics' values of *DC* for the queries were similar to those for other queries, while the gaps between values of *CTFIDF* were relatively large. This means that only some of the subtopics for the queries had high document representativeness; in other words, each of the many documents for the queries included more than two subtopics with different search intents. In short, we can say that *DC* has an advantage when a certain document has multi-intent, while *CTFIDF* is favorable when a document holds a one intent condition. Therefore, we expect to obtain a high subtopic mining performance by giving appropriate weights to *DC* and *CTFIDF* based on a judgment as to whether related documents have multi-intent by using *CTFIDF*.

Table 3. Examples of subtopics mined by KLE-S-E-4A.

Query (topic ID)	|Intent|	Rank. subtopic (intent number or non-relevant '-')
angular cheilitis (402)	9	1. angular cheilitis treatment (3) 2. angular cheilitis cure (3) 3. angular cheilitis causes (2) 4. angular cheilitis candidiasis (5) 5. angular cheilitis crustosa factitia 0 images (8) 6. angular cheilitis granulomatosa (5) 7. angular cheilitis Wikipedia (8) 8. angular cheilitis solution (3) 9. definition of angular cheilitis (4) 10. angular cheilitis symptoms (1)
grilling (410)	6	1. barbecues grilling (2) 2. grilling recipes (1) 3. grilling tips (3) 4. grilling flank steak (6) 5. grilling tools (-) 6. grilling chicken (-) 7. grilling corn (-) 8. grilling salmon (5) 9. grilling shrimp (5) 10. grilling plank (-)
barbados (417)	5	1. barbados tourism (2) 2. barbados news (-) 3. barbados map (5) 4. barbados blog (-) 5. barbados hotels (1) 6. barbados weather (-) 7. entry to barbados (-) 8. barbados underground (-) 9. barbados tours (2) 10. barbados real estate (3)
brooks brothers clearance (440)	9	1. garth brooks brothers clearance (8) 2. brooks brothers clearance center (3) 3. brooks brothers clearance sale (6) 4. brooks brothers clearance items (2) 5. brooks brothers clearance outlet (4) 6. brooks brothers clearance event (-) 7. brooks brothers clearance store (4) 8. brooks brothers clearance fee (-) 9. brooks brothers clearance ties (2) 10. brooks brothers clearance shoes (2)

In the Japanese subtopic mining, our best values of mean I-rec, D-nDCG, and D#-nDCG were 0.2607, 0.2726, and 0.2632 for l = 10, respectively. These evaluated values were too low because we could not extract appropriate candidate strings as subtopics in the first step. There were unexpected low relevance and non-understandable candidate strings such as "金魚のこと (thing of goldfish)" and "のに麻雀 (even though mah-jong)." Therefore, to obtain more refined candidate strings, we will modify our method by considering segmentation and useless term detection. As for future work, we will combine our method with approaches based on open resources, and research evaluation issues in detail.

5. ACKNOWLEDGMENTS

This work was supported by the Korea Ministry of Knowledge Economy (MKE) under Grant No.10041807, in part by the National Korea Science and Engineering Foundation (KOSEF) (NRF-2010-0012662).

6. REFERENCES

[1] 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, Proceedings of NTCIR-9 Workshop Meeting, pages 82-105, 2011.

[2] R. L. T. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification, Proceedings of the 19th International Conference on World Wide Web, pages 881-890, 2010.

[3] R. L. T. Santos, C. Macdonald, and I. Ounis. University of glasgow at the ntcir-9 intent task, Proceedings of NTCIR-9 Workshop Meeting, pages 111-115, 2011.

[4] Y. Xue, F. Chen, T. Zhu, C. Wang, Z. Li, Y. Liu, M. Zhang, Y. Jin, and S. Ma. Thuir at ntcir-9 intent task, Proceedings of NTCIR-9 Workshop Meeting, pages 123-128, 2011.

[5] S. Zhang, K. Lu, and B. Wang. Ictir subtopic mining system at ntcir-9 intent task, Proceedings of NTCIR-9 Workshop Meeting, pages 106-110, 2011.

[6] J. Han, Q. Wang, N. Orii, Z. Dou, T. Sakai, and R. Song. Microsoft research asia at the ntcir-9 intent task, Proceedings of NTCIR-9 Workshop Meeting,pages 116-122, 2011.

[7] X. Jiang, X. Han, and L. Sun. Iscas at subtopic mining task in ntcir9, Proceedings of NTCIR-9 Workshop Meeting, pages 168-171, 2011.

[8] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma, and J. Ma. Learning to cluster web search results, Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 210-217, 2004.

[9] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends in Information Retrieval, vol. 3 (4), pages 333-389, 2009.

[10] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the ntcir-10 intent-2 task, Proceedings of NTCIR-10 Workshop Meeting, 2013.

[11] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance, Proceedings of the 34th ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1043-1052, 2011.