Microsoft Research Asia at the NTCIR-10 Intent Task

Kosetsu Tsukuda
Kyoto University
tsukuda@dl.kuis.kyoto-u.ac.jp

Zhicheng Dou
Microsoft Research Asia
zhichdou@microsoft.com

Tetsuya Sakai
Microsoft Research Asia
tesakai@microsoft.com

ABSTRACT

Microsoft Research Asia participated in the Subtopic Mining subtask and Document Ranking subtask of the NTCIR-10 INTENT Task. In the Subtopic Mining subtask, we mine subtopics from query suggestions, clickthrough data and top results of the queries, and rank them based on their importance for the given query. In the Document Ranking subtask, we diversify top search results by estimating the intent types of the mined subtopics and combining multiple search engine results. Experimental results show that our best Japanese subtopic mining run is ranked No. 2 of all 14 runs in terms of D♯-nDCG@10. All of our Japanese document ranking runs outperform the baseline ranking without diversification.

Team Name

MSINT

Subtasks

Subtopic Mining (Japanese) and Document Ranking (Japanese)

Keywords

Query Intent, Subtopics, Diversity

1. INTRODUCTION

The MSINT team participated in the Subtopic Mining subtask and Document Ranking subtask of the NTCIR-10 INTENT Task [16]. The goal of the Subtopic Mining subtask is to return a ranked list of subtopic strings for a given query. In the Subtopic Mining subtask, we extract subtopics from query suggestions, clickthrough data and top results of the queries. Then we estimate the importance of each subtopic based on the hit count or weighted search result overlap.

The goal of the Document Ranking subtask is to diversify search results based on the mined subtopics. In the Document Ranking subtask, we propose a diversification framework that builds on the one proposed by Dou *et al.* [7]. Our framework reflects the *intent types* (informational vs. navigational [4]). If it is possible to distinguish between informational and navigational intents, search engines can aim to return one best URL for each navigational intent, while allocating more space to the informational intents within the SERP [14, 15]. For example, consider a navigational intent "I want to go to the Red Cliff movie website": the user probably wants one particular URL for this intent, so the search engine probably should try to allocate more space to the other more informational intents, for which more relevant documents basically means more informativeness. In addition to the intent types, the original search results for a query also play an important role for search result diversification. We combine multiple search engine results to obtain more relevant Web pages for a query.

Figure 1 shows the overview of our proposed system with some examples. Here, the query is "red cliff", and our first step is to obtain subtopics such as "red cliff review" and "red cliff homepage" from query suggestions, query logs and search results. For simplicity, the figure shows only up to four subtopics obtained from each resource. Our second step clusters all of the obtained subtopics. In this example, we have three clusters and each cluster represents an intent. Then our third step estimates the importance of each subtopic. In cluster I_1, for example, the degrees of importance of "red cliff review" and "red cliff critique" are 0.62 and 0.47, respectively. In this case, we take "red cliff review" as a representative subtopic of I_1 and the degree of importance of I_1 is 0.62. In the fourth step, we estimate intent types of each representative subtopic: we classify subtopics "red cliff review" and "red cliff movie" to informational and "red cliff HP" to navigational. Our final step generates a diversified search result. In this step, we optionally combine multiplel search engine results of the query "red cliff." In this example, documents which are relevant to the most important intent I_1 are allocated more space and ranked higher than those which are relevant to I_2 and I_3. Only one document which is relevant to I_2 is included in the result because I_2 is classified to navigational.

2. SUBTOPIC MINING

2.1 Subtopic Mining Resources

Our subtopic mining component mines subtopics of a given query from three different resources, as described below.

2.1.1 Query Suggestions

Query suggestions obtained from WSEs are an easy and effective choice for obtaining subtopics. Santos, Macdonald and Ounis [18] evaluated the effectiveness of two kinds of query suggestions for the purpose of search result diversification: "suggested queries," (a.k.a. query autocompletions) which are presented in a drop-down list within the search box as the user enters a query, and "related queries," which are shown alongside a ranked list of URLs within a SERP. As their results suggest that suggested queries are more effective for search result diversification, we also decided to use suggested queries rather than related queries.

2.1.2 Clickthrough Data

Another popular resource for obtaining subtopics is clickthrough data. In our method, we first obtained data that consists of approximately 14.8 million Japanese queries from Bing over a one month period (April 2012). Then, for each original query q, we used the

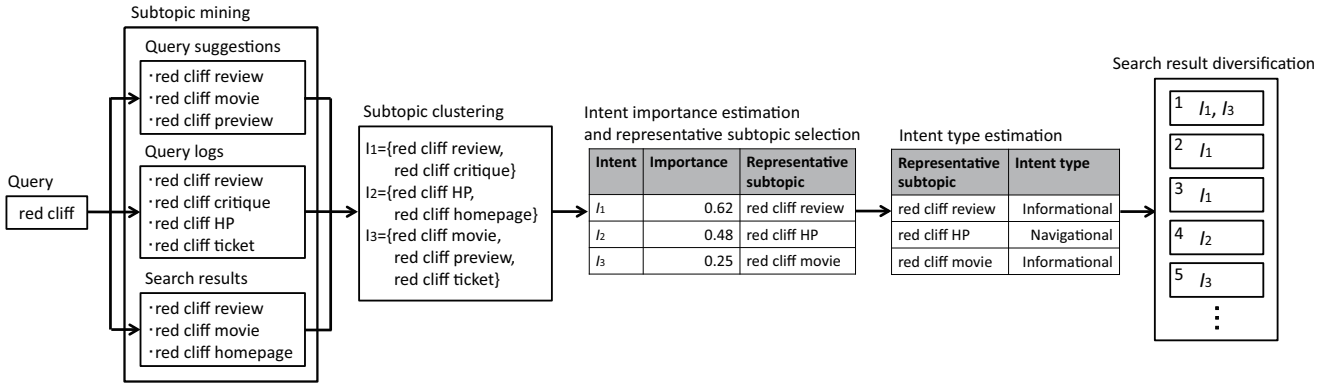

Figure 1: Overview of our method.

following simple filters for obtaining candidate subtopics: extract all queries that (1) were issued by at least five unique users; and (2) are of the form "q plus an additional keyword." The first condition is designed to avoid subtopics that are too obscure; the second condition was devised based on the observation that most of the subtopics submitted by the NTCIR-9 INTENT Japanese Subtopic Mining participants conformed to this style [1]. Thus, for example, given a query "red cliff," we might obtain "red cliff review" or "red cliff dvd" from the clickthrough data, but not "review of red cliff."

2.1.3 Search Result Clusters

While either query suggestions or clickthrough data may work for simple phrase queries, these resources may not help when the original query is more complex. For example, if the original query is a natural language query, WSEs may fail to return query suggestions, and even clickthrough data may not contain natural language queries that completely subsume the original query. We therefore follow Zeng *et al.* [20] and use search result clusters for mining subtopic candidates. In their method, top N search results for the original query are grouped into K clusters based on key phrases (n-grams) extracted from snippets, where the following features are used for clustering: (1) phrase frequency and inverted document frequency, (2) intra-cluster similarity, (3) cluster entropy; and (4) phrase independence. But as our preliminary experiment suggested that the fourth feature is harmful, we dropped this feature. As for the parameters, we used $N = 200$ and $K = 10$, following Zeng *et al.* [20].

The above method obtains words such as "reviews" and "dvd": we thus add the original query to the mined words to form subtopics such as "red cliff reviews." Also, the above method requires a search engine for obtaining a ranked list of URLs with snippets for a given query. For this purpose, we used Microsoft's internal web search platform WebStudio [2] which was also used by Han *et al.* [9] in their participation at the NTCIR-9 INTENT task. The index web corpus is the Japanese portion of Clueweb09. Unless otherwise noted, this is the search platform we use for creating document rankings throughout this paper.

2.2 Subtopic Clustering

Having obtained candidate subtopics for a given query, the next step is to cluster subtopics in order to identify the *intents*.

As Dou *et al.* [7] reported that combining subtopics from multiple sources is useful for discovering user intents, we first pool all subtopics extracted from query suggestions, clickthrough data and search result clusters. Recall that not all of our subtopics are head queries: thus click-based clustering methods [2, 6, 10] would not work for this purpose. Instead, we use a simple clustering approach based on search result contents.

First, we extract all terms from the titles and snippets in the top l web pages returned for each subtopic, using Bing API [3]. Then, we create a feature vector for each subtopic, where each element represents the tf-idf value for an extracted term. Here, "tf" is the total frequency of the term within the top l result (titles and snippets only) for the subtopic; "df" is the number of subtopics whose search results contain the term. By assuming that subtopics that share the same intent have similar search results, we can apply a clustering algorithm to the subtopics represented as vectors.

We apply the well-known Ward's method [19] for clustering subtopics. The distance between clusters C_1 and C_2, denoted by $D(C_1, C_2)$, is calculated as:

$$D(C_1, C_2) = E(C_1 \cup C2) - E(C_1) - E(C_2), \quad (1)$$

$$E(C_i) = \sum_{\boldsymbol{x} \in C_i} (d(\boldsymbol{x}, \boldsymbol{c}_i))^2, \quad (2)$$

where $\boldsymbol{c}_i$ is the centroid of C_i, given by $\sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}/|C_i|$, and $d(\boldsymbol{x}, \boldsymbol{c}_i)$ is the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{c}_i$. In our case, each element $\boldsymbol{x} \in C_i$ corresponds to a subtopic.

As Ward's method is a hierarchical agglomerative clustering (HAC) method, it requires a stopping condition for obtaining a set of clusters. Obviously, if we set a loose threshold, subtopics that represent different intents may end up in the same cluster. Conversely, if we set a tight threshold, subtopics that represent the same intent may end up in different clusters. We set a threshold based on the average distance between every pair of subtopics:

$$d_{avg}(q) = \frac{2}{n(n-1)} \sum_{\boldsymbol{x_i}, \boldsymbol{x_j} \in C} d(\boldsymbol{x_i}, \boldsymbol{x_j}), \quad (3)$$

where C is the set of all subtopics obtained through subtopic mining (Sections 2.1.1-2.1.3) and $n = |C|$. We stop clustering the subtopics when the minimum distance between two clusters is less

[1] In the ongoing NTCIR-10 INTENT-2 task, participants are explicitly encouraged to submit subtopics of this form. See `http://research.microsoft.com/en-us/people/tesakai/intent2.aspx`.

[2] WebStudio platform: `http://research.microsoft.com/en-us/projects/webstudio/`.

[3] `http://msdn.microsoft.com/en-us/library/dd251056.aspx`

than $d_{avg}(q) * h$. The value of h is set to 0.3 based on our preliminary experiment.

2.3 Intent Importance Estimation

Having obtained clusters of subtopics, we first estimate the importance of each subtopic. Then, the most important subtopic from each cluster is taken as a *representative subtopic*, which we regard as a representation of a particular intent. By intent importance ranking, we mean ranking these representative subtopics. Only the representative subtopics are used for diversifying the search result.

Below, we describe two methods that we experimented with for estimating the importance of each subtopic: *hit count* and *weighted search result overlap*.

2.3.1 Hit Count

Our first method for estimating the importance of a subtopic is based on the intuition that, if a subtopic is frequently mentioned in web pages, then that subtopic should be important. Recall that our subtopics are always superstrings of the original query: thus, a page that contains a subtopic string also contains the original query. Let $Hit(c_i)$ denote the web hit count of a subtopic c_i, obtained by Bing API. This method calculates the importance of a subtopic c_i given a query q as:

$$P_{hit}(c_i|q) = \frac{Hit(c_i)}{\sum_{c_j \in C} Hit(c_j)}. \quad (4)$$

2.3.2 Weighted Search Result Overlap

Our second method is based on the overlap between a SERP for the original query and a SERP for each subtopic. The basic assumption behind it is that, if a subtopic is important for a given query, then the two SERPs have many URLs in common. We also assume that the overlap between the sets of URLs near top ranks is more important than that between those at low ranks. Let $D_k(q)$ and $D_k(c_i)$ denote the set of top k retrieved URLs for q and c_i, respectively. This method calculates the importance of a subtopic c_i given a query q as:

$$P_{wov}(c_i|q) = \sum_{d \in D_k(c_i) \cap D_k(q)} \frac{1}{rank(q,d)} \quad (5)$$

where $rank(q,d)$ is the rank of the document d in the ranked list for q.

2.4 Submitted Runs

We submit the following five runs for the Japanese Subtopic Mining subtask:

- MSINT-S-J-1B: Use query suggestions, clickthrough data and top results of the queries. Rank subtopics based on the weighted overlap of search results.
- MSINT-S-J-2B: Use query suggestions, clickthrough data and top results of the queries. Rank subtopics based on hit count.
- MSINT-S-J-3A: Use query suggestions, clickthrough data. Rank subtopics based on the weighted overlap of search results.
- MSINT-S-J-4A: Use query suggestions. Rank subtopics based on the weighted overlap of search results.
- MSINT-S-J-5B: Use top results of the queries. Rank subtopics based on the weighted overlap of search results.

3. INTENT TYPE ESTIMATION

Since Broder [4] proposed his taxnomy of search intents (navigational, informational and transactional), some researchers have addressed the problem of classifying queries into intent types, especially for the first two intent types [8, 11, 12]. In contrast to their faithful interpretation of "navigational" ("*The immediate intent is to reach a particular site*" [4]), we adopt a broader interpretation for the purpose of search result diversification, following Sakai and Song [17]. To be more specific, in addition to *homepage finding* intents, we also consider *single answer finding* intents as navigational. For example, if the user submits a query "president obama full name," probably exactly one good web page that answers this question suffices for this intent, and any additional web pages that contain the same information would be redundant. Thus, even though this is not a homepage finding intent, the single answer finding intent is similar to it in that it also requires exactly one web page. From the viewpoint of optimizing the SERP, these two types of intents can both be regarded as navigational and be treated differently from informational intents, as we generally want as many relevant (but nonredundant) documents as possible for the latter.

We use Support Vector Machine (SVM) developed by Boser, Guyon and Vapnik [3] with RBF (Radial Basis Function) kernel to classify representative subtopics into navigational and informational intent types. Effective classification features used in previous studies include click entropy [8], anchor-link distribution [12] and query term distribution [11], but these are not suitable for our purpose for the following two reasons. First, as not all of the representative subtopics are head queries, statistics such as click entropy are not so reliable. Second, while these methods may be suitable for separating homepage finding intents from informational intents, they are probably not for separating single answer finding intents from informational intents. For example, while many users click the same URLs for homepage finding intents, different users may click different URLs to find the answer to the aforementioned question: "president obama full name," just like with informational intents.

In order to solve the above two problems, we propose two categories of features for SVM below: click features and character type features. The latter category of features was designed for Japanese queries and is language-dependent; the other parts of the proposed framework are basically language-independent.

3.1 Click Features

Our first category of features for intent type estimation is based on clickthrough data. Recall that not all of our subtopics are head queries, and that therefore looking for occurrences of the subtopics in the clickthrough data would not work. Instead, we assume that the rightmost term (or *tail term*) of a query is often useful for estimating query intent types. For example, suppose that the user wants to read reviews of the movie Red Cliff: we assume that the user is likely to enter "red cliff review" rather than "review red cliff." Here, the tail term "review" suggests that the intent is informational: the user wants many relevant documents. Similarly, if the user wants to visit the Red Cliff official homepage, we assume that the user will enter "red cliff homepage": again, the tail term suggests that the intent is navigational. (Note that these examples are provided for the English speaking reader: the actual queries and subtopics we currently handle are in Japanese.) Note that while the occurrrences of "red cliff review" may not be frequent in the clickthrough data, those of "review" probably are. Thus we try to avoid the sparsity problem.

More specifically, given a subtopic c, we first extract its tail term t. (If c consists of one term, then t is equal to c.) Then, we ex-

tract all queries that contain t as a tail term from the clickthrough data. As each record in our clickthrough data contain a user id, a query, a clicked URL and its position, we can compute the following feature for t: (1) Average number of clicked pages per query per user; (2) Average number of unique clicked URLs per query; (3) Average rank of the first clicked web page for each query for each user; (4) Average rank of the last clicked web page for each query for each user; and (5) Average rank of any clicked web pages for each query for each user. The first feature represents how many pages are clicked after a user issues a query; if this is small, the query whose tail term is t may be navigational. The second feature approximates the number of relevant URLs for a query containing t; this should be small at least for homepage finding intents, if not for single answer finding intents. The other three features are to do with clicked ranks: for example, we can hypothesize that many homepage finding intents are easy to satisfy, as search engines often manage to return the home pages near the top ranks. In addition to these five features for t, we also compute the corresponding statistics for the most frequent query that has t as its tail term. Hence we use ten click features in total.

3.2 Character Type Features

Our second category of features for intent type estimation is designed specifically for Japanese, and is based on character types. Unlike English, Chinese and many other languages, the Japanese language uses three distinct character types that are outside the ascii codes: kanji, katakana and hiragana. Kanji, also known as Chinese characters, is an ideogram; Katakana and hiragana are phonograms. Just like our click features, we examine the tail term of a given subtopic as described below.

We observed that when the intent is informational, the tail term tends to be made up from a single character set, e.g. "*joho* (an all-kanji word meaning "information")" and "*osusume* (an all-hiragana word meaning "recommendation")." On the other hand, when the intent is navigational, the tail term tends to be more specific, e.g. "*shin-ruru-kaisetsu* (a kanji-katakana-combined word meaning "explanation of a new rule")." Moreover, we observed that the similar tendency is also seen about a query: e.g. the subtopics of a query "*ketsuekigata* (an all-kanji word meaning "blood type")" often belong to an informational intent, whereas those of a query "*aipoddo-no-tsukaikata* (a kanji-katakana-hiragana-combined word meaning "how to use iPod")" often belong to a navigational intent.

In light of this observation, we count how many times the character types change in the tail term and the original query, and use them as features. For example, the score of *joho* is zero because it is written in kanji only; that of *shin-ruru-kaisetsu* is two because *shin* and *kaisetsu* are written in kanji, whereas *ruru* is written in katakana. Orii, Song and Sakai [13] also used the combination of character types as features for a Japanese question classification task and found it effective.

4. DOCUMENT RANKING

As we mentioned earlier, our proposed diversification framework builds on the one proposed by Dou *et al.* [7], which has been shown to outperform IA-Select [1] and Maximal Marginal Relevance [5]. The framework was also used at the NTCIR-9 INTENT Japanese Document Ranking subtask, where it outperformed other participating teams. In this section, we first describe the method for combining multiple search engine results. We then describe the algorithm by Dou *et al.*, and propose a modification below.

4.1 Combination of Multiple Search Engine Results

Both our framework and Dou's one use the nondiversified ranked list of a query. If many irrelevant documents are included in the list, they reduce the accuracy of diversification. To solve this problem, we hypothesize that if a document is ranked higher in multiple search engine results for a query, the document is surely relevant to the query. We obtain top m documents for a query from WebStudio, Yahoo [4] and Bing, and calculate the socre of each document. In order to obtain Yahoo search results, we use the Yahoo! Web Search API[5]. The score of a document d is given by:

$$s(d) = \frac{1}{r_{webstudio}(d)} + \frac{1}{r_{yahoo}(d)} + \frac{1}{r_{bing}(d)}, \quad (6)$$

where $r_{webstudio}(d)$, $r_{yahoo}(d)$ and $r_{bing}(d)$ denote the rank of d retrieved by WebStudio, Yahoo and Bing, respectively. When d is not included in the top m documents of WebStudio, for example, we set $\frac{1}{r_{webstudio}(d)}$ to 0.

After calculating the scores of all documents, we rank them in descending order of the scores and obtain a ranked list for the query.

4.2 Dou et al.

Let C denote the set of representative subtopics obtained as described in Section 2.3 and let c be a member of C. Using WebStudio (See Section 2.1.3), we first generate a nondiversified ranked list for the original query q and for each representative subtopic c: following Dou et al. [7], we obtain 1,000 URLs for q and 10 URLs for each c. When we use multiple search engine results, we combine them based on the method as described in Section 4.1 instead of using only WebStudio. Let $rank(q, d)$ denote the rank of document d in the nondiversified ranked list of q. According to Dou *et al.*, the relevance score of document d with respect to the original query q is given by:

$$rel(q, d) = \frac{1}{\sqrt{rank(q, d)}}. \quad (7)$$

Similarly, $rel(c, d)$, the relevance score of d with respect to a representative subtopic c is also computed.

Let R be the pool of candidate documents retrieved by the original query q and its subtopics, and let S_n denote the top n documents selected so far. Dou *et al.* [7] employs a greedy algorithm which iteratively selects documents and generates a diversified ranking list. The n-th document is given by:

$$d_{n+1} = \arg\max_{d \in R \setminus S_n} [\rho \cdot rel(q, d) + (1 - \rho) \cdot \Phi(d, S_n, C)], \quad (8)$$

where ρ is the parameter that controls the tradeoff between relevance and diversity; $\Phi(d, S_n, C)$ represents a topic richness score of d given the set S_n of documents already selected:

$$\Phi(d, S_n, C) = \sum_{c \in C} w_c \cdot \phi(c, S_n) \cdot rel(c, d), \quad (9)$$

where w_c is the importance of subtopic c. In this paper, w_c is calculated by one of the three methods described in Section 2.3. $\phi(c, S_n)$ is the discounted importance of subtopic c given S_n, given by:

$$\phi(c, S_n) = \begin{cases} 1 & \text{if } n = 0; \\ \prod_{d_s \in S_n} [1 - rel(c, d_s)] & \text{otherwise.} \end{cases} \quad (10)$$

More details of this framework can be found in Dou *et al.* [7].

[4] http://www.yahoo.com/

[5] http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html

Table 2: Japanese Document Ranking runs ranked by mean D♯-nDCG@10. The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10	DIN-nDCG@10	P+Q
MSINT-D-J-4B	**0.8160**	**0.5029**	**0.6595**	**0.3458**	**0.3666**
MSINT-D-J-3B	0.7809	0.4457	0.6133	0.3182	0.3373
MSINT-D-J-5B	0.7809	0.4397	0.6103	0.3124	0.3282
MSINT-D-J-1B	0.7789	0.4388	0.6089	0.3099	0.3248
MSINT-D-J-2B	0.7655	0.4505	0.6080	0.3159	0.3271
baseline	0.7428	0.4136	0.5782	0.2820	0.3160

Table 1: Japanese Subtopic Mining runs ranked by mean D♯-nDCG@10. The highest value in each column is shown in bold.

run name	I-rec@10	D-nDCG@10	D♯-nDCG@10
MSINT-S-J-4A	**0.2988**	**0.3085**	**0.3036**
MSINT-S-J-1B	0.2969	0.3058	0.3013
MSINT-S-J-3A	0.2746	0.2980	0.2863
MSINT-S-J-2B	0.2659	0.2494	0.2576
MSINT-S-J-5B	0.2370	0.2341	0.2356

4.3 Proposed Framework

As the algorithm by Dou *et al.* does not consider intent types, we modify it in order to make it intent type-aware. We propose two modified methods, but first describe their common features.

In our intent type-aware models, the relevance score with respect to c is given by:

$$rel(c,d) = p_{inf}(c) \cdot rel_{inf}(c,d) + p_{nav}(c) \cdot rel_{nav}(c,d) \quad (11)$$

where $p_{inf}(c)$ ($p_{nav}(c)$) is the probability that c is informational (navigational), as estimated by our SVM-based intent type estimation component. The key here is that the relevance score with respect to c is defined separately depending on intent types. In particular, we define the relevance score for the case where c is navigational as

$$rel_{nav}(c,d) = \begin{cases} 1 & \text{if } rank(c,d) = 1; \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

to reflect the fact that we want exactly one relevant document for such an intent. Whereas, $rel_{inf}(c,d)$, the corresponding score for the informational case, is calculated by equation (7).

4.4 Submitted Runs

We submit the following five runs for the Japanese Document Ranking subtask:

- MSINT-D-J-1B: Use Dou's search result diversification model, considering intent type.
- MSINT-D-J-2B: Use Dou's search result diversification model, considering intent type. Combine search results of WebStudio, Yahoo and Bing.
- MSINT-D-J-3B: Use Dou's search result diversification model, considering intent type. Not diversify search result when topic has a navigational intent.
- MSINT-D-J-4B: Use Dou's search result diversification model. Combine search results of WebStudio, Yahoo and Bing.
- MSINT-D-J-5B: Use Dou's search result diversification model.

5. EXPERIMENTAL RESULTS

5.1 Japanese Subtopic Mining Runs

Table 1 shows the evaluation results of our Subtopic Mining runs. According to Sakai *et al.* [16], our proposed methods are statistically indistinguishable from one another in all metrics, but MSINT-S-J-4A attained best performance throughout all evaluation metrics. This result shows that, as for the data sources of subtopics, using only query suggestions has a better effectiveness than combining some resources such as clickthrough data or top results of the queries. Comparing MSINT-S-J-1B and MSINT-S-J-2B, we see that the weighted overlap of search results can estimate the subtopic importance more precisely than the hit count.

Figure 2 shows the per-topic D♯-nDCG performances for Japanese Subtopic Mining subtask. MSINT-S-J-4A is sperior to the other four methods in some topics such as 0304, 0333, 0334, 0343 and 0364. All queries of the topics consist of only one keyword. For example, Topic 0333 is "moth," Topic 0343 is "sat" and Topic 0364 is "titanic." In such a case, query suggestions provide appropriate and a sufficient number of subtopics, and MSINT-S-J-4A, which uses only query suggestions as a resource, outperforms other methods. On the other hand, when the query is complex, for example "maps of U.S.A." of Topic 0319 and "recipes of oyster" of Topic 0400, MSINT-S-J-5, which uses search result clusters, is superior to MSINT-S-J-4A. These results suggest that if we can use different resources according to the complexity of queries, we can achieve a better result.

5.2 Japanese Document Ranking Runs

Table 2 shows the evaluation results of our Document Ranking runs. In Table 2, "baseline" means the original ranking without diversification. We observe that all of our submitted runs improved upon the baseline. Especially, MSINT-D-J-4B significantly outperforms the baseline.

MSINT-D-J-4B significantly outperforms MSINT-D-J-5B in terms of D♯-nDCG, which shows that the combination of multiple search engines was successful. MSINT-D-J-4B also significantly outperforms MSINT-D-J-2B in terms of D♯-nDCG, which means that our intent type-aware diversification was not successful. This is probably due to the limited accuracy of our intent type classification. There is room for further investigation here. Finally, while the difference between MSINT-D-J-3B and MSINT-D-J-5B is statistically insignificant, selective diversification does have a positive effect, albeit very small. Further investigation is required here too.

Figure 3 shows the per-topic D♯-nDCG performances for Japanese Document Ranking subtask. MSINT-D-J-4B is sperior to the other four methods in some topics such as 0305, 0313, 0324, 0341 and 0381. They are all simple queries, for example "kung fu" of Topic 0305, "volvo" of Topic 0313 and "yahoo" of Topic 0324. It appears that search results are greatly different from one search engine to another for simple queries. Documents retrieved by many search

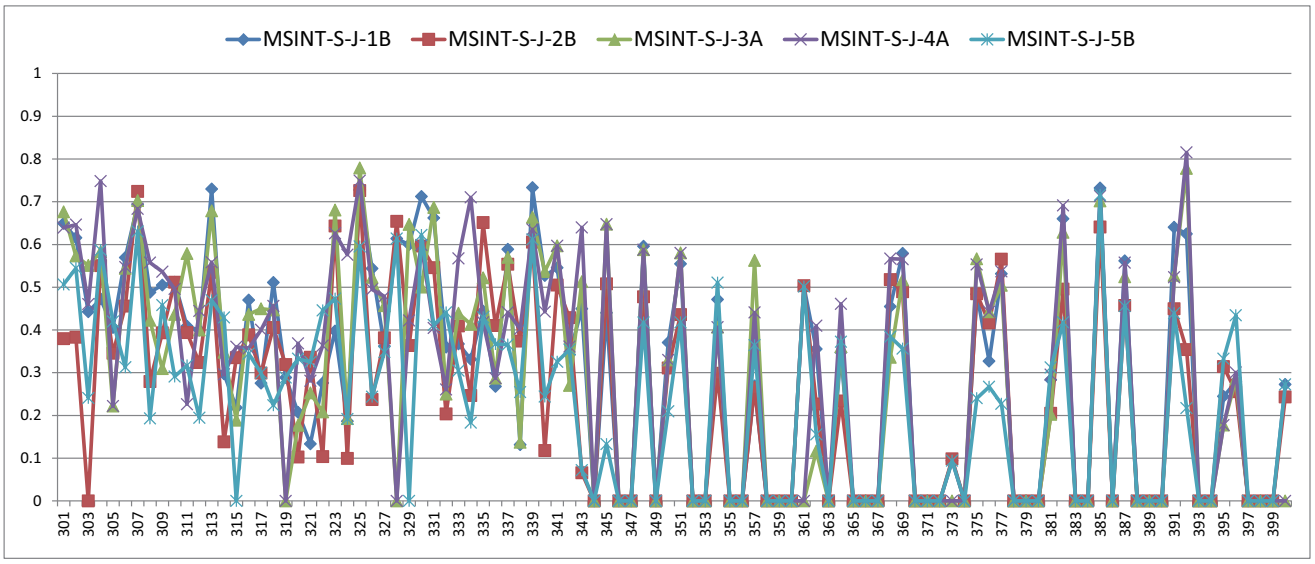

Figure 2: Per-topic D♯-nDCG performances for Japanese Subtopic Mining subtask.

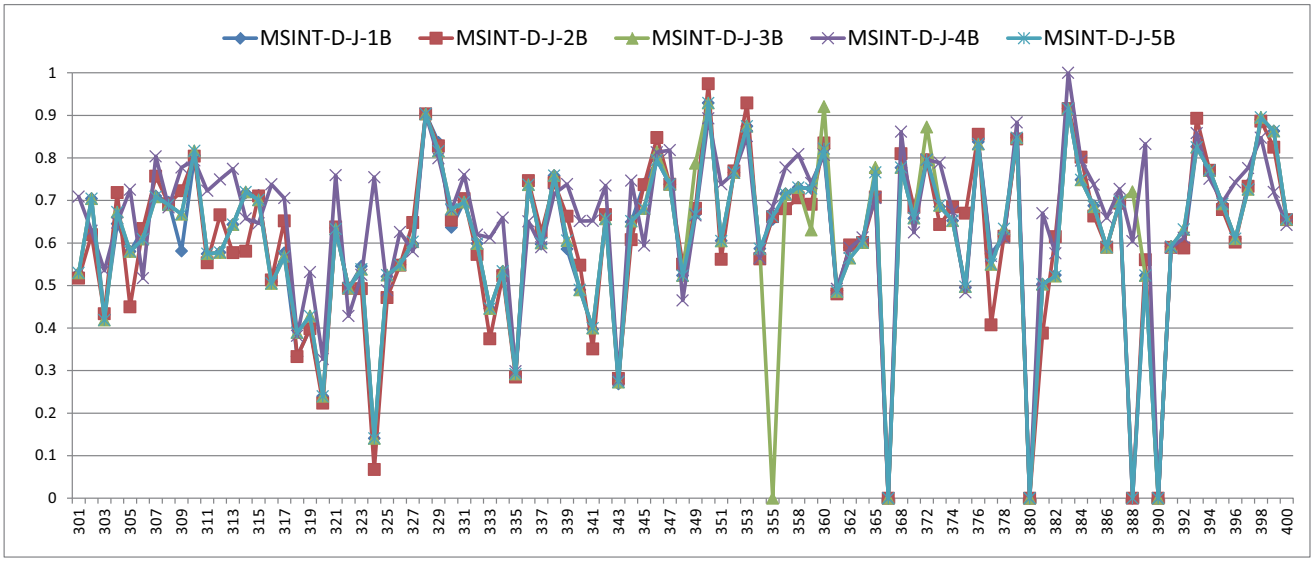

Figure 3: Per-topic D♯-nDCG performances for Japanese Document Ranking subtask.

engines are highly relevant to the query and as a result, MSINT-D-J-4B achieves high D♯-nDCG. As we intended, MSINT-D-J-3B is sperior in some navigational queries such as "official blog of Aya Nakashima" of Topic 360, "homepage of Tokyo Marine & Nichido Systems" of Topic 0372 and "homepage of Hokuetsu Bank" of Topic 0388. In Document Ranking subtask, too, it is important to use different methods according to query types such as query complexity and intent types in order to achieve a better result.

6. CONCLUSIONS

In this paper, we proposed methods for Subtopic Mining subtask and Document Ranking subtask in the NTCIR10 Intent Task. In Subtopic Mining subtask, we mine subtopics from query suggestions, clickthrough data and top results of the queries. In Document Ranking subtask, we diversify top search results by estimating the intent types of the mined subtopics and combining multiple search engine results. Results showed the effectiveness of using query suggestions for the subtopic mining and combining multiple search engine results for the search results diversification.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proc. of ACM WSDM 2009*, pages 5–14, 2009.

[2] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proc. of ACM SIGKDD 2000*, pages 407–416, 2000.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. of COLT 1992*, pages 144–152, 1992.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proc. of ACM SIGIR 1998*, pages 335–336, 1998.

[6] W. S. Chan, W. T. Leung, and D. L. Lee. Clustering search engine query log containing noisy clickthroughs. *Proc. of SAINT 2004*, pages 305–308, 2004.

[7] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proc. of ACM WSDM 2011*, pages 475–484, 2011.

[8] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proc. of WWW 2007*, pages 581–590, 2007.

[9] J. Han, Q. Wang, N. Orii, Z. Dou, S. T., and R. Song. Microsoft research asia at the ntcir-9 intent task. In *Proc. of NTCIR-9*, pages 116–122, 2011.

[10] M. Hosseini, H. Abolhassani, and M. S. Harikandeh. Content free clustering for search engine query log. In *Proc. of SMO 2007*, pages 201–206, 2007.

[11] I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proc. of ACM SIGIR 2003*, pages 64–71, 2003.

[12] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. of WWW 2005*, pages 391–400, 2005.

[13] N. Orii, Y.-I. Song, and T. Sakai. Microsoft research asia at the ntcir-9 1click task. In *Proc. of NTCIR-9*, pages 216–222, 2011.

[14] T. Sakai. Evaluation with informational and navigational intents. In *Proc. of WWW 2012*, pages 499–508, 2012.

[15] T. Sakai. Web search evaluation with informational and navigational intents. In *to appear in Journal of Information Processing, Vol.21, No.1*, 2013.

[16] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the ntcir-10 intent-2 task. In *Proc. of NTCIR-10*, 2013.

[17] T. Sakai and Y.-I. Song. Designing diversity evaluation environments with navigational intents. In *submitted to ACM TOIS*.

[18] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proc. of WWW 2010*, pages 881–890, 2010.

[19] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[20] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster web search results. In *Proc. of ACM SIGIR 2004*, pages 210–217, 2004.