Estimating Intent Types for Search Result Diversification

Kosetsu Tsukuda[1,⋆], Tetsuya Sakai[2], Zhicheng Dou[3], and Katsumi Tanaka[1]

[1] Kyoto University, Japan
{tsukuda,tanaka}@dl.kuis.kyoto-u.ac.jp
[2] Waseda University, Japan
tetsuyasakai@acm.org
[3] Microsoft Research Asia, China
zhichdou@microsoft.com

Abstract. Given an ambiguous or underspecified query, search result diversification aims at accommodating different user intents within a single Search Engine Result Page (SERP). While automatic identification of different intents for a given query is a crucial step for result diversification, also important is the estimation of intent types (informational vs. navigational). 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. In light of the observations, we propose a new framework for search result diversification that is intent importance-aware and type-aware. Our experiments using the NTCIR-9 INTENT Japanese Subtopic Mining and Document Ranking test collections show that: (a) our intent type estimation method for Japanese achieves 64.4% accuracy; and (b) our proposed diversification method achieves 0.6373 in D♯-nDCG and 0.5898 in DIN♯-nDCG over 56 topics, which are statistically significant gains over the top performers of the NTCIR-9 INTENT Japanese Document Ranking runs. Moreover, our relevance oriented model significantly outperforms our diversity oriented model and the original model by Dou *et al.*.

Keywords: Search Result Diversity, Subtopic, Intent Type.

1 Introduction

Given an ambiguous or underspecified query, search result diversification aims at accommodating different user intents within a single Search Engine Result Page (SERP). For example, a query "red cliff" may represent several different search intents, such as "I want to go to the Red Cliff movie website" and "I want to read various reviews of the movie Red Cliff." Given a query, typical diversification algorithms first try to identify these different intents, and then rank documents so that "novel" documents (i.e. those that are dissimilar to the ones ranked above them) are included in the SERP [1,4,5,16,17].

⋆ This research was conducted while the first author was an intern at Microsoft Research Asia.

R.E. Banchs et al. (Eds.): AIRS 2013, LNCS 8281, pp. 25–37, 2013.

While automatic identification of different intents for a given query is a crucial step for result diversification, we argue that also important is the estimation of *intent types* (informational vs. navigational [3]). 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 [13]. For example, consider the aforementioned 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 light of the above observations, we propose a new framework for search result diversification that is *intent type-aware*. The framework comprises the following steps:

Subtopic mining and clustering. We first obtain *subtopics* from query suggestions, query logs and search results. Here, a subtopic is an instance of a representation of a particular search intent given a query, which either disambiguates or specifies the original query[1]. As a single intent may be represented by several different subtopic strings, we automatically cluster the mined subtopics to identify intents. For example, subtopics "red cliff review" and "red cliff critique" may form the "red cliff review" cluster.

Intent importance estimation. Next, we estimate the importance of each intent by utilizing search engine results for the original query as well as those for the subtopics.

Intent type estimation. We also classify each intent to either navigational or informational using Support Vector Machine (SVM), so that we can allocate more space to the informational intents compared to the navigational intents in the SERP. Here, our interpretation of "navigational" is slightly broader than the original definition by Broder [3], as we shall discuss in Section 5.

Document reranking. Finally, we generate a diversified search result by leveraging the intents, estimated intent probabilities and types.

Except for the character type feature used in intent type estimation, our framework is basically language-independent.

Our experiments using the NTCIR-9 INTENT Japanese Subtopic Mining and Document Ranking test collections [18] show that: (a) our intent type estimation method for Japanese achieves 64.4% accuracy; and (b) our proposed diversification method achieves 0.6373 in D♯-nDCG [14] and 0.5898 in DIN♯-nDCG [12] over 56 topics, which are statistically significant gains over the top performers of the NTCIR-9 INTENT Japanese Document Ranking runs[2]. Moreover, our relevance oriented model significantly outperforms our diversity oriented model and the original model by Dou *et al.* [5].

[1] `http://research.microsoft.com/INTENT/`

[2] It should be noted, however, that the official top performers at NTCIR-9 worked under time pressure and that a postmortem comparison of this kind is only indicative.

2 Related Work

2.1 Intent Type Estimation

Lee, Liu and Cho [9] proposed a method for identifying the user goals (informational or navigational) based on user-click behavior and anchor-link distribution. Dou, Song and Wen [6] utilized the click entropy to estimate intent types of queries. These studies concern only head queries, for which reliable statistics can be obtained from clickthrough data. In contrast, we aim to estimate the intent type of *any* given subtopic, and therefore their methods are not directly applicable. Li, Wang and Acero [10] constructed click graphs based on clickthrough data and developed query intent classifiers. In order to compensate for the sparsity of a click graph, they also used the contents of documents. Our approach also utilises both clickthrough data and search engine results, as we shall describe in Section 5.

2.2 Search Result Diversification

Several search result diversification algorithms have been proposed in the literature [1,4,5,16,17]. The common approach is to first identify multiple possible subtopics (or intents) for the given query, and to try to cover as many subtopics as possible with the SERP, by minimizing retrieved redundant documents for each subtopic. State-of-the-art diversification algorithms include *IA-select* by Agrawal *et al.* [1], *xQuAD* by Santos, Macdonald and Ounis [16] and the algorithm by Dou *et al.* [5]. Santos, Macdonald and Ounis [17] also proposed a diversification approach which takes intent types (navigational and informational) into account. However, their approach does not aim to return one best URL for a navigational intent.

Our proposed algorithm uses the algorithm by Dou *et al.* as the starting point.

3 Subtopic Mining and Clustering

3.1 Subtopic Mining Resources

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

Query Suggestions. Query suggestions, which are "suggested queries" (a.k.a. query autocompletions) and "related queries," obtained from WSEs are an easy and effective choice for obtaining subtopics. As Santos, Macdonald and Ounis [16] suggest that suggested queries are more effective for search result diversification, we also decided to use suggested queries rather than related queries. In our experiments, we use the "official" Japanese suggested queries as we shall describe in Section 7.1.

Clickthrough Data. Another popular resource for obtaining subtopics is clickthrough data. In our experiments, 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 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[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. 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. 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[4]. Unless otherwise noted, this is the search platform we use for creating document rankings throughout this paper.

3.2 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.* [5] 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 [2,7] 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[5]. 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. As Ward's method is a hierarchical agglomerative clustering (HAC) method, we stop clustering the subtopics when the minimum distance between two clusters is less than $d_{avg}(q) * h$, where $d_{avg}(q)$ represents the average distance between every pair of subtopics.

In this paper, we empirically set l and h to 200 and 0.3, respectively.

4 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

[3] In the NTCIR-10 INTENT-2 task, participants were explicitly encouraged to submit subtopics of this form. See `http://research.microsoft.com/INTENT/`

[4] `http://research.microsoft.com/en-us/projects/webstudio/`

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

representative subtopic, which we regard as a representation of a particular intent. Only the representative subtopics are used for diversifying the search result.

Our method for intent importance estimation is based on the overlap between a SERP for the original query and a SERP for each subtopic, and the rank information for each subtopic. The assumption is 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 a query q and a subtopic c_i, respectively. This method calculates the importance of c_i given q as:

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

where $rank(q,d)$ is the rank of the document d in the ranked list for q. In this paper, we empirically set k to 200.

5 Intent Type Estimation

Since Broder [3] proposed his taxnomy of search intents (informational, navigational and transactional), some researchers have addressed the problem of classifying queries into intent types, especially for the first two intent types [6,8,9]. In contrast to their faithful interpretation of "navigational" ("*The immediate intent is to reach a particular site*" [3]), we adopt a broader interpretation for the purpose of search result diversification, following Sakai and Song [15]. 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. From the viewpoint of optimizing the SERP, these two types of intents can both be regarded as navigational.

We use SVM with RBF (Radial Basis Function) kernel to classify representative subtopics into navigational and informational intent types. Effective classification features were used in previous studies [6,8,9], 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, 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. Only the latter category of features was designed for Japanese queries and is language-dependent.

5.1 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 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 extract 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 features 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.

5.2 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. "*shinruru-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.

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. Orii, Song and Sakai [11] also used these features for a Japanese question classification task and found it effective.

6 Search Result Diversification

As we mentioned earlier, our proposed diversification framework builds on the one proposed by Dou *et al.* [5], which has been shown to outperform IA-Select [1] and MMR [4]. The framework was also used at the NTCIR-9 INTENT Japanese Document Ranking subtask, where it outperformed other participating teams. We first describe the algorithm by Dou *et al.*, and then propose a few modifications below.

6.1 Dou et al.

Let C denote the set of representative subtopics obtained as described in Section 4 and let c be a member of C. We first generate a nondiversified ranked list for the original query q and for each representative subtopic c: following Dou et al. [5], we obtain 1,000 URLs for q and 10 URLs for each c. 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 $rel(q, d) = 1/\sqrt{rank(q, d)}$. 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.* [5] employs a greedy algorithm which iteratively selects documents and generates a diversified ranking list. The $n + 1$-th document is given by:

$$d_{n+1} = \arg\max_{d \in R \setminus S_n} [\rho \cdot rel(q, d) + (1 - \rho) \cdot \Phi(d, S_n, C)], \tag{2}$$

where ρ is the parameter that controls the tradeoff between relevance and diversity and we use $\rho = 0.3$, following Dou *et al.* [5]; $\Phi(d, S_n, C)$ represents a topic richness score of d given the set S_n:

$$\Phi(d, S_n, C) = \sum_{c \in C} w_c \cdot \phi(c, S_n) \cdot rel(c, d), \tag{3}$$

where w_c is the importance of subtopic c. In this paper, w_c is calculated by the method described in Section 4. $\phi(c, S_n)$ is the discounted importance of subtopic c given S_n:

$$\phi(c, S_n) = \begin{cases} 1 & \text{if } n = 0; \\ \prod_{d_s \in S_n} [1 - rel(c, d_s)] & \text{otherwise.} \end{cases} \tag{4}$$

More details of this framework can be found in Dou *et al.* [5].

6.2 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:

$$rel(c, d) = p_{inf}(c) \cdot rel_{inf}(c, d) + p_{nav}(c) \cdot rel_{nav}(c, d), \tag{5}$$

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

$$rel_{nav}(c,d) = \begin{cases} 1 \text{ if } rank(c,d) = 1; \\ 0 \text{ otherwise,} \end{cases} \quad (6)$$

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, differs according to our two models.

Relevance Oriented Model. In our first model, we let $rel_{inf}(c,d) = 1/\sqrt{rank(c,d)}$ just as in the original model. However, we modify $\phi(c,S_n)$: we still use Equation 4 if c is navigational, but let $\phi(c,S_n) = 1$ regardless of n if c is informational. This is because Equation 4 penalizes "redundant" documents for each c regardless of the intent type. In intent type-aware diversification, multiple relevant documents for an informational intent are not necessarily "redundant."

Diversity Oriented Model. In our second model, we first rerank each ranked list for each informational intent c to obtain a new rank for document d (denoted by $rerank(c,d)$), and let $rel_{inf}(c,d) = 1/\sqrt{rerank(c,d)}$. The reranking is intended to prioritize documents that cover many intents compared to those that are highly relevant to one particular intent. Thus, for each document d in the original ranked list for c, we first count the number of intents that also retrieved d. Using the number of covered intents as the first key (larger the better) and the original rank as the second key (smaller the better), we sort the original ranked list.

7 Experiments

This section reports on a component-by-component evaluation of our proposed framework using the NTCIR-9 Document Ranking test collections.

7.1 Data

Our experiments utilize the NTCIR-9 INTENT Japanese Subtopic Mining and Document Ranking test collections [18]. These test collections were constructed as follows:

1. In the Subtpic Mining subtask, 100 topics were released to participating teams, who returned a ranked list of subtopics for each topic;
2. The INTENT task organisers pooled the submitted subtopics and let assessors manually cluster them to form intents, and to provide a name for each intent;
3. The organisers then estimated intent probabilities based on assessor voting;
4. In the Document Ranking subtask, the same 100 topics were released to participating teams, who returned a diversified list of search results for each topic;
5. The organisers pooled the submitted documents and let assessors conduct per-intent graded relevance assessments, using the set of intents identified through the Subtopic Mining subtask.

Table 1. Intent type classification accuracy for the 481 intents

	true navigational	true informational	total
estimated as navigational	83	117	200
estimated as informational	54	227	281
total	137	344	481

As for the Document Ranking subtask, the document collection used in the Document Ranking task is the ClueWeb09-JA collection, which is the Japanese portion of ClueWeb09[6]. Per-topic graded relevance assessments are provided on a five-point scale: from $L0$ (judged nonrelevant) to $L4$ (highly relevant), based on assessments by two assessors for every topic.

The organisers released *query suggestion data*, which were scraped from Google, Bing and Yahoo, for the NTCIR-9 INTENT topics to its participants, in order to enhance the repeatability of the participants' experiments and to enable fair comparison. In our subtopic mining method, we also utilise this data set.

In addition to the above official data from the INTENT tasks, we obtained the *intent type labels* for the INTENT-1 Japanese topics from Sakai and Song [15], so that we can conduct intent type-aware evaluation. According to the intent type labels, only 56 topics of the 100 Japanese INTENT-1 topics contains at least one navigational and informational intents. For this reason, hereafter we use these 56 topics only. On average, each topic has 2.32 navigational intents (21%) and 8.89 informational intents (79%).

Evaluating search result diversification using an existing diversity test collection, however, is problematic. This is because existing diversity test collections are highly unlikely to be reusable, as their relevance assessments are obtained through shallow pooling [13]. For example, TREC 2010 and NTCIR-9 diversity test collections all used the pool depth of 20. Therefore, if a new system is evaluated using the official relevance assessments, the system is underestimated, as it returns many unjudged documents, some of which might be relevant. In light of this, we conducted some additional relevance assessments of our own to obtain more reliable results, following the relevance assessment procedure used at the INTENT task. We shall discuss this in Section 7.3.

7.2 Results of Intent Types Estimation

In this section, we discuss the accuracy of our intent type estimation component. As was described in Section 5, we use an SVM classifier to determine whether each given intent is likely to be navigational or informational. As SVM requires training data, we conducted the evaluation as follows. The 56 Japanese topics from the INTENT task had 1,902 (539 navigational and 1,363 informational) intents in total, but our subtopic mining and intent importance estimation components managed to identify only 481 of them (137 navigational and 344 informational). Since the remaining 1,421 (402 navigational and 1,019 informational) intents are never used in any part of our evaluation, these unused intents were utilized for training the SVM classifier. Furthermore, in order to avoid including extremely rare intents in the training data, only those that have at least 50 hits in our clickthrough data were used. This gave us 819 intents (231 navigational and 588

[6] `http://lemurproject.org/clueweb09/`

informational). Finally, to balance the amount of training data, we randomly sampled 231 informational intents.

Table 1 shows the classification results for the aforementioned 481 intents. The overall classification accuracy was $(83 + 227)/481 = 0.644$. It can be observed that navigational intents are more difficult to classify than the informational ones. From our classification results, we found that our approach that relies on tail terms has some clear limitations. In particular, it is often difficult to determine whether an intent is navigational or informational from its tail term alone. For example, "beijing image" (user wants pictures of Beijing) may be labelled as informational, as the information need is vague and it is not clear if any one particular image will completely satisfy the user. On the other hand, "dutch flag image" (user wants an image of the Dutch national flag) may be labelled as navigational, as returning one item may suffice. The gold standard data set itself contains some gray area: Sakai and Song [15] report that the kappa agreement of intent type labels between two assessors was .713 for TREC diversity topics. In short, our intent type classification task itself is a difficult one.

7.3 Results of Search Result Diversification

Evaluation Metrics. To finally evaluate the diversified search results, we use five evaluation metrics, namely, I-rec, *D-nDCG*, *D♯-nDCG* [14], *DIN-nDCG* and *DIN♯-nDCG* [13][7]. The first three measures are the official metrics used at the NTCIR-9 INTENT task: D♯-nDCG is a linear combination of I-rec (a pure diversity measure) and D-nDCG (an overall relevance measure). We evaluate the top 10 documents as our objective is to diversify the first search engine result page.

In contrast, the recently proposed DIN-nDCG and DIN♯-nDCG are more suitable for the purpose of intent type-aware diversity evaluation. DIN(♯)-nDCG is a simple modification of D(♯)-nDCG: the only difference is that, whenever multiple relevant documents are retrieved for a navigational intent, DIN(♯)-nDCG treats only the highest ranked relevant document as relevant to that intent. These intent type-aware metrics were used at the NTCIR-10 INTENT-2 Document Ranking subtasks.

More details on the evaluation metrics can be found elsewhere [13].

Evaluation with the Intent Data. We evaluate the overall performance of our diversified search system using the intent sets from the INTENT task. In this experiment, we compared three methods: the framework by Dou *et al.* [5] (**Dou**), the relevance oriented model proposed in Section 6.2 (**REL**), and the diversity oriented model proposed in Section 6.2 (**DIV**). In addition, we obtained top performing runs from the NTCIR-9 INTENT Japanese Document Ranking tasks: MSINT-D-J-3 and MSINT-D-J-2, which were the top two performers in terms of both I-rec@10 and D♯-nDCG@10; and uogTr-D-J-1 and uogTr-D-J-2, which were the top two performers in term of D-nDCG@10. (These official results suggest that the MSINT runs are diversity oriented while the uog runs are relevance oriented [18].)

As we briefly mentioned in Section 7.1, we conducted some additional relevance assessments for this experiment as some of the documents returned by our systems

[7] nDCG stands for normalized Discounted Cumulative Gain; D- stands for Diversification; DIN- stands for Diversification with Informational and Navigational intents.

Table 2. Diversification performances with the intents, importance and types obtained by the system (56 topics, each with all intents). The highest score is shown in bold. A two-sided t-test was used for significance testing. Significant differences with MSINT-D-J-2 is indicated by a $*$ ($\alpha = 0.05$) or a $**$ ($\alpha = 0.01$). Similarly, a $\star$, a † and a ‡ indicate significant differences with MSINT-D-J-3, uogTr-D-J-1 and uogTr-D-J-2, respectively.

	I-rec@10	D-nDCG@10	D♯-nDCG@10	DIN-nDCG@10	DIN♯-nDCG@10
uogTr-D-J-2	0.6843	0.4500	0.5671	0.3481	0.5162
uogTr-D-J-1	0.6832	0.4540	0.5686	0.3505	0.5169
MSINT-D-J-2	0.7626	0.4326	0.5976	0.3574	0.5600
MSINT-D-J-3	0.7649	0.4328	0.5988	0.3574	0.5611
Dou	0.7733 †† ‡‡	0.4557	0.6145 † ‡	0.3762	0.5748 †† ‡‡
DIV	0.7798 †† ‡‡	0.4551	0.6174 † ‡	0.3755	0.5777 †† ‡‡
REL	**0.7935** †† ‡‡	**0.4810** $**$ $\star\star$	**0.6373** $**$ $\star\star$ †† ‡‡	**0.3861** $*$ $\star$	**0.5898** $*$ $\star$ †† ‡‡

Table 3. Comparison of different diversification methods in terms of significant difference. A two-sided t-test was used for significance testing. Significant differences between two methods are indicated by a $*$ ($\alpha = 0.05$) or a $**$ ($\alpha = 0.01$). A method name written with a $*$ or a $**$ is a winner. A symbol "-" represents there is no significant difference between two methods.

	I-rec@10	D-nDCG@10	D♯-nDCG@10	DIN-nDCG@10	DIN♯-nDCG@10
Dou vs. **REL**	-	**REL**$*$	**REL**$*$	-	-
DIV vs. **REL**	-	**REL**$**$	**REL**$*$	-	-

are not covered by the official relevance assessments. The first two authors of this paper used the official relevance assessment tool from the INTENT task [18] to independently conduct relevance assessments for 97 unjudged documents, and the relevance assessments were merged with the official ones. The inter-assessor kappa agreement for this additional document set was 0.581, which is statistically significant at $\alpha = 0.01$.

Table 2 shows the performances of our seven runs (three proposed systems plus four official runs from NTCIR-9). The runs have been sorted by DIN♯-nDCG. It can be observed that **REL** significantly outperforms all top performing runs from the NTCIR-9 INTENT Japanese Document Ranking task in terms of D♯-nDCG and DIN♯-nDCG. Table 3 summarize the significant test results when different diversification methods are compared. Table 3 shows that **REL** is the best diversification method.

8 Conclusion

We proposed a new intent type-aware search result diversification framework, and conducted evaluation using the NTCIR-9 INTENT Japanese Subtopic Mining and Document Ranking test collections. Except for the character set-based feature used for intent type estimation, our proposed framework is basically language-independent.

Our main findings are as follows: (a) Our intent type estimation method for Japanese achieved 64.4% accuracy. Moreover, navigational intents were more difficult to classify than informational ones; and (b) For search result diversification, methods using the relevance oriented model significantly outperformed our diversity oriented model and

the original model by Dou *et al.* [5]. Our best method achieved 0.6373 in D♯-nDCG and 0.5898 in DIN♯-nDCG over 56 topics, which are statistically significant gains over the top performers of the NTCIR-9 INTENT Japanese Document Ranking runs.

Our future work includes evaluation with English diversity test collections (i.e. TREC diversity data), and exploration of more sophisticated diversification methods. For example, our current models do not consider the contents of the documents already selected: some document features may be useful for estimating whether a document is likely to be relevant to a navigational intent or to an informational intent, or even both.

Acknowledgements. This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013) from MEXT of Japan and JSPS KAKENHI Grant Number 243993.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proc. of ACM WSDM 2009, pp. 5–14 (2009)
2. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: Proc. of ACM SIGKDD 2000, pp. 407–416 (2000)
3. Broder, A.: A taxonomy of web search. SIGIR Forum 36(2), 3–10 (2002)
4. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. of ACM SIGIR 1998, pp. 335–336 (1998)
5. Dou, Z., Hu, S., Chen, K., Song, R., Wen, J.-R.: Multi-dimensional search result diversification. In: Proc. of ACM WSDM 2011, pp. 475–484 (2011)
6. Dou, Z., Song, R., Wen, J.-R.: A large-scale evaluation and analysis of personalized search strategies. In: Proc. of WWW 2007, pp. 581–590 (2007)
7. Hosseini, M., Abolhassani, H., Harikandeh, M.S.: Content free clustering for search engine query log. In: Proc. of SMO 2007, pp. 201–206 (2007)
8. Kang, I.-H., Kim, G.: Query type classification for web document retrieval. In: Proc. of ACM SIGIR 2003, pp. 64–71 (2003)
9. Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: Proc. of WWW 2005, pp. 391–400 (2005)
10. Li, X., Wang, Y.-Y., Acero, A.: Learning query intent from regularized click graphs. In: Proc. of ACM SIGIR 2008, pp. 339–346 (2008)
11. Orii, N., Song, Y.-I., Sakai, T.: Microsoft Research Asia at the NTCIR-9 1CLICK Task. In: Proc. of NTCIR-9, pp. 216–222 (2011)
12. Sakai, T.: Evaluation with informational and navigational intents. In: Proc. of WWW 2012, pp. 499–508 (2012)
13. Sakai, T.: Web search evaluation with informational and navigational intents. Journal of Information Processing 21(1), 145–155 (2013)
14. Sakai, T., Song, R.: Evaluating diversified search results using per-intent graded relevance. In: Proc. of ACM SIGIR 2011, pp. 1043–1052 (2011)
15. Sakai, T., Song, Y.-I.: On labelling intent types for evaluating search result diversification. In: Banchs, R.E., Silvestri, F., Liu, T.-Y. (eds.) AIRS 2013. LNCS, vol. 8281, pp. 38–49. Springer, Heidelberg (2013)
16. Santos, R.L., Macdonald, C., Ounis, I.: Exploiting query reformulations for web search result diversification. In: Proc. of WWW 2010, pp. 881–890 (2010)

17. Santos, R.L., Macdonald, C., Ounis, I.: Intent-aware search result diversification. In: Proc. of ACM SIGIR 2011, pp. 595–604 (2011)
18. Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M.: Overview of the NTCIR-9 INTENT Task. In: Proc. of NTCIR-9, pp. 82–105 (2011)
19. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)
20. Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., Ma, J.: Learning to cluster web search results. In: Proc. of ACM SIGIR 2004, pp. 210–217 (2004)