SEM13 at the NTCIR-11 IMINE Task: Subtopic Mining and Document Ranking Subtasks

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

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

ABSTRACT

In this paper, we describe our participation in the English Subtopic Mining and Document Ranking subtasks of the NTCIR-11 IMINE Task. In the Subtopic Mining subtask, we mine subtopics from query suggestions, query dimensions, and Freebase entities of a given query, rank them based on their importance for the given query, and finally construct a two-level hierarchy of subtopics. In the Document Ranking subtask, we diversify top search results by estimating the coverage of the mined subtopics. The best performance of our system achieves an *Hscore* of 0.1762, a *Fscore* of 0.3043, a *Sscore* of 0.3689, and an *H-measure* of 0.0634 for subtopic mining task. For document ranking run, the best performance of our system achieves a *D#-nDCG@10* of 0.6022 (coarse-grain) and 0.5291 (fine-grain), which are a comparable performance to other participants.

Team Name

SEM13

Subtasks

Subtopic Mining Subtask (English)
Document Ranking Subtask (English)

Keywords

subtopic, intent, diversification

1. INTRODUCTION

When an information need is being formulated in user's mind, query in the form of a sequence of words will be typed into the search box, ideally, the search engine should respond with a ranked list of snippet results that best meet the need of user. Web search queries are typically short, ambiguous, and contain multiple aspects or subtopics [6,29]. A query is classified into three types i.e. *ambiguous*, *broad*, and *clear* [27]. The search intent of faceted queries is usually clear; so that the search engine can report helpful results. However, information retrieval systems often fail to capture users' search intents exactly if a submitted query is ambiguous or broad. Because an ambiguous query has more than one interpretation and different users have different intents for the same query, which corresponds to different subtopics.

Some intents of a query are constantly popular; however some others intents are time-dependent. For example, the query "apple" may refer to two subtopics: (1) "apple Inc." and (2) "apple fruit". Each subtopic may also contain several second level subtopics; for example "apple iPhone 5s", "apple iPad", "apple iOS", and "apple store" with respect to the subtopic of "apple Inc.". In some cases, subtopics associated to query can be temporally ambiguous; for instance, the query US Open is more likely to be targeting the tennis open in September, and the golf tournament in June [18]. In addition, it is not clear which aspect of a multifaceted query is actually desirable for a user. For example, the faceted query "air travel information" may contain different subtopics, such as (1) information on air travel, airports, and airlines, (2) restrictions for checked baggage during air travel, and (3) websites that collect statistics and report about airports [30].

With the enormous size of the Web, a misunderstanding of the information needs underlying a user's query can misguide the search engine to produce a ranked result page that may frustrate the user, and lead the user to abandon the originally submitted query. Traditional information retrieval models, such as the Boolean model and the vector space model treat every input query as a clear, well-defined representation, and completely neglect any sort of ambiguity. Ignoring the underlying subtopics of a query, information retrieval models might produce top ranked documents possibly containing too much relevant information on a particular aspect of a query, and eventually leave the general user unsatisfied.

To maximize the satisfaction of users, a retrieval model should select a list of documents that are not only relevant to the popular subtopics, but also covers different subtopics of the query. In order to satisfy the user, a sensible approach is to diversify the documents retrieved for the query [7]. The diversified retrieval model should produce a ranked list of documents that provide the maximum coverage and minimum redundancy with respect to the possible aspects underlying a query. The solution of the aforementioned diversification problem might be composed of two parts: understanding the intent behind a query and diversifying the results with respect to the possible intents.

Recently, mining subtopics of an user query for diversifying the retrieved documents has received considerable attention [22]. Several methods are proposed for mining subtopics from different aspects, such as the retrieved documents, the query logs, Wikipedia, Freebase [10], and the related search services provided by the commercial search engines [23,30,31].

In this paper, we address our solution to the subtasks of *Subtopic Mining* and *Document Ranking* of NTCIR-11 IMINE task.

The rest of the paper is organized as follows: **Section 2** overviews related work on subtopic mining and document ranking. **Section 3** introduces our proposed framework for intent mining. **Section 4** includes the overall experiments and the results that we obtained. Finally, concluding remarks and some future directions of our work are described in **Section 5**.

2. RELATED WORK

Queries are usually short, ambiguous and/or underspecified [6, 27, 29]. To understand the meanings of queries, researchers define taxonomies and classify queries into predefined categories. Song et al. [28] classified the queries into three categories: ambiguous queries, which have more than one meaning; board queries, which covers a variety of subtopics; and clear queries, which have a specific meaning or narrow topics. At the query level, Broder [4] divided query intent into navigational, informational, and transactional types. Nguyen and Kan [17] classified queries into four general facets of ambiguity, authority, temporal sensitivity, and spatial sensitivity. Boldi et al. [3] created query-flow graph with query phrase nodes and used them for query recommendation.

Query suggestion is a key technique for generating alternative queries to help users drill down to a subtopic of the original query [14, 33]. Different from query suggestion or query completion, subtopic mining focuses more on the diversity of possible subtopics of the original query rather than inferring relevant queries. Jian Hu et al. [11] integrated the knowledge contained in Wikipedia to predict the possible intents for a given query. Filip Radlinkshi et al. [20] proposed an approach for inferring query intents from reformulations and clicks. For an input query, the click and reformulation information are combined to identify a set of possible related queries to construct an undirected graph. An edge is introduced between two queries if they were often clicked for the same documents. Finally, random walk similarity is used to find intent cluster. At the session level, Radlinski and Joachims [19] mined intent from query chains and used it for learning to rank algorithm.

Recently, Wang et al. [30] proposed a method to mine subtopics of a query either directly from the query itself or indirectly from the retrieved documents of the retrieval systems to diversify the search results. In indirect approach, subtopics are extracted by clustering, topic modeling, and concept-tagging of the retrieved documents. In direct approach, several external resources, such as Wikipedia, Open Directory Project (ODP), search query logs, and the related search services are investigated to mine subtopics. Santos et al. [23] leveraged the query reformulations of web search engines (WSEs) to mine sub-queries (i.e. subtopics) for diversifying web search results. The surrounding text of query terms in the top retrieved documents are also utilized to mine subtopics [31].

3. OUR APPROACH

In this section, we describe our proposed framework of the diversified retrieval system. In our system, we focus on two parts: *Subtopic mining* and *Document ranking*. Our proposed framework is depicted in Fig. 1.

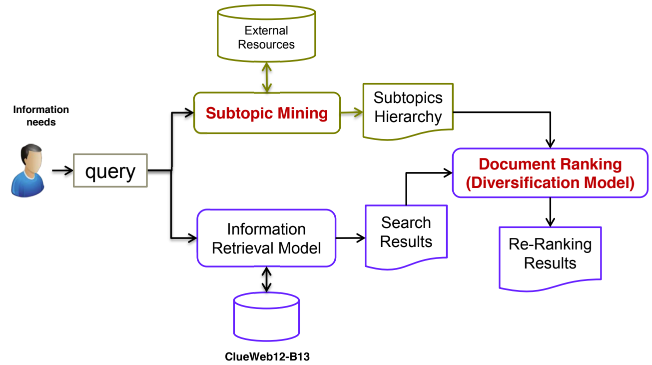

Figure 1: Framework of diversified retrieval system

3.1 Subtopic Mining

Our subtopic mining framework is depicted in Fig. 2. In this framework, we mine subtopic candidates from multiple resources including query suggestions, query dimensions, and Freebase KB. Given a query, we construct a two-level hierarchy of subtopics. In two-level hierarchy, there are some first-level subtopics, whereas under each first-level subtopic, there are several second-level subtopics. To construct a two-level hierarchy of subtopics, at first, we mine second-level subtopics. From the second-level subtopics, we mine first-level subtopics and construct the hierarchy. To mine second-level subtopics, we filter out a subtopic candidate if the query contains the subtopic terms or has similar terms; because the poorly formed subtopic candidate do not specialize or disambiguate the original query. To rank the candidate subtopics, we estimate the importance scores by extracting some query-dependent and query-independent features. Since we mine candidate subtopics from multiple sources, there might be many redundant subtopics. To select the subtopics by considering maximum relevance with minimum redundancy, we apply Maximum Marginal Relevance (MMR) based diversification model to the filtered subtopic candidates. We consider these diversified subtopics as the second-level subtopics. To extract the first-level subtopics, we apply K-means algorithm on the second-level subtopics to generate some clusters and label these generated clusters with the top frequent terms in each cluster. The label of each cluster is considered as first-level subtopic. With the first-label and second-label subtopics, we construct a two-level hierarchy of subtopics for each query.

3.1.1 Candidate Generation

Query suggestions obtained from web search engines (WSEs) are an easy and effective choice for obtaining subtopics. Santos et al. [23] evaluation's result reveals that "suggested queries" (a.k.a query auto completions) are more effective than "related queries". Query dimensions generated by Dou et al. [8] are also produced some effective subtopics of a query. Existing knowledge graph such as freebase contains some important subtopics for several queries. Given a query, we retrieve all the query suggestions, dimensions, and freebase entities, and aggregate them by filtering out the duplicates or wrongly represented ones, and consider them as subtopic candidates.

3.1.2 Feature Extraction

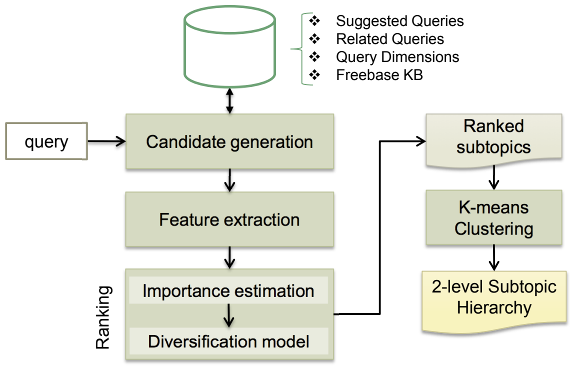

Figure 2: Subtopic mining framework

To measure the importance of the subtopic candidates, we apply the fusing technique with multiple features extracted from the subtopic candidates. In this regard, we broadly organize all subtopic candidate features used by our approach as either query-dependent or query-independent, according to whether they are computed on-the-fly at querying time or offline at indexing time, respectively.

3.1.3 Query-dependent features

Given a query Q, query-dependent features are directly computed by scoring the occurrences of the terms of query Q in each subtopic candidate of $\{S_1, S_2, S_3, ..., S_N\}$. Among the query-dependent features, we extract some term dependency, term frequency, and lexical features.

Language modeling with Dirichlet smoothing feature, f_{LMDS} is defined based on language modeling approach to information retrieval [26] and smoothed using Dirichlet smoothing [32]. It is computed as the log likelihood of the query being generated from the subtopic candidate.

Language modeling with Jelinek-Mercer smoothing [32] feature, f_{LMJM} is defined as the linear combination of the probability of the query term given the subtopic candidate and the probability of the query term in background language model.

A particularly effective approach to exploit term dependency was proposed by [15]. In this model, unigram, bigram sequential dependency, and bigram full dependency are linearly interpolated. The term dependency with Markov random field based feature, f_{MRF} is computed from the query and the subtopic candidate.

To measure the lexical similarity between the query and the subtopic candidate, edit distance based feature, f_{EDS} is computed as:

$$f_{EDS} = 1 - \frac{Edit\_distance(Q,S)}{Max(length(Q), length(S))}$$

Another simple lexical feature is Exact Match. Exact match feature, f_{EM} is a binary feature that returns 1 if there is an exact lexical match of the query within the subtopic candidate [16]. It is computed as:

$$f_{EM} = I(Q \text{ substring of } S)$$

where I is the indicator function that returns 1 if its argument is satisfied.

Overlap feature, $f_{Overlap}$ is simply defined as the fraction of query terms that occur, after stemming and stopping, in the subtopic candidate [16]. It is computed as:

$$f_{Overlap} = \frac{\sum_{w \in Q} I(w \in S)}{|Q|}$$

Overlap-syn feature, $f_{Overlap-syn}$ is the generalization of the Overlap feature by also considering synonyms of query terms. It is defined as the fraction of query terms that either match with subtopic candidate term or have a synonym that matches with subtopic candidate term. It is computed as:

$$f_{Overlap-syn}(Q,S) = \frac{\sum_{w \in Q} I(Syn(w) \in S)}{|Q|}$$

where $Syn(w)$ denotes the set of synonyms of the term w, including the term itself. We use the WordNet 3.0 [9] to get the synonyms of noun, adjective, verb, and adverb followed by singularising with Krovertz stemmer [12] and POS tagging with Standford NLP parser [25].

BM25 [21] is an effective term weighting method that incorporates the query term frequency, subtopic length, and inverse subtopic frequency. We extract BM25 feature, f_{BM25} from the query and subtopic candidate, where the parameters $k_1 = 1.80$ and $b = 0.75$ are assigned for empirical reason.

A non-parametric divergence from randomness (DFR) based models, DFH [1] has been shown to perform effectively across a variety of Web search tasks [24]. We extract the term frequency based DFH feature, f_{DFH} is computed as:

$$f_{DFH}(Q,S) = \sum_{w \in S} \frac{tf_{w,S}\ (1 - \frac{tf_{w,S}}{l_S})^2}{tf_{w,S} + 1} \log_2(tf_{w,S}\ \frac{avgl_S}{l_S\ tf_{w,C}}) + 0.5 \log_2(2\pi tf_{w,S}\ (1 - \frac{tf_{w,S}}{l_S}))$$

where $tf_{w,S}$ is the frequency of the term in the subtopic candidate S and $tf_{w,C}$ is the frequency of the term in the collections.

3.1.4 Query-independent features

The goal of query independent features is to encode a prior knowledge we have about individual subtopic candidate. Here, we extract some simple query independent features.

Longer terms in subtopic would reflect a more thoughtful and readable style. To focus on the readability of the subtopic, average term length (ATL) in a subtopic candidate is defined as:

$$f_{ATL} = \frac{1}{l_S} \sum_{w \in S} tf_{w,S}\ l_w$$

where l_w denotes the length in characters of the term w.

Additional readability features have been recently proposed is topic cohesiveness (TC) [2]. Topic cohesiveness feature, f_{TC} is computed as:

$$f_{TC} = -\sum_{w \in S} P(w|S)\ \log P(w|S)$$

where $P(w|S)$ is computed using a maximum likelihood estimation.

3.2 Subtopic ranking

In this section, we rank the subtopic candidates to optimize both relevancy and diversity of the subtopic candidates.

3.2.1 Importance Estimation

We estimate the importance score $Imp(.)$ of a subtopic candidate by considering the extracted features in the above section. We normalize each feature using max-min normalization technique. For a subtopic candidate S, we represent all the extracted features in a feature vector, $\mathbf{FV}_S = \{f_{LMDS}, f_{LMJM}, f_{MRF}, ..., f_{TC}\}$ with dimension 11, one dimension of $\mathbf{FV}$ for each feature. Thus, given a query, we have a list of subtopic candidates with the corresponding feature vectors.

Now, the mean feature vector, $\overline{\mathbf{MF}}$ is computed from the feature vectors of all the subtopic candidates as:

$$\overline{\mathbf{MF}} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{FV}_{S_j}$$

Here, N is the number of subtopic candidates.

We define the importance score $Imp(S)$ of a subtopic candidate S as the cosine similarity between the subtopic feature vector $\mathbf{FV}_S$ and the mean feature vector $\overline{\mathbf{MF}}$. Therefore, subtopic candidate importance score is estimated as:

$$Imp(S) = CosineSim(\mathbf{FV}_S, \overline{\mathbf{MF}}) = \frac{\mathbf{FV}_S \cdot \overline{\mathbf{MF}}}{\parallel \mathbf{FV}_S \parallel \parallel \overline{\mathbf{MF}} \parallel}$$

We consider the importance score $Imp(S)$ of a subtopic candidate S as the relevancy score $Rel(S)$ for subtopic diversification.

3.2.2 MMR-based Ranking

We utilize the maximum marginal relevance (MMR) framework [5] to further evaluate the diversity of mined subtopic candidates.

Given a relevance function $Rel(.)$ and a similarity function $Sim(.,.)$, the MMR model could be set up as follows:

$$S_{i+1} = argmax_{S \notin D_i}\ \alpha\ Rel(S) + (1-\alpha)\ Nov(S, D_i)$$

where $\alpha \in [0,1]$ and it is a combining parameter. Then

$$D_{i+1} = D_i \cup S_{i+1}$$

Here, S_i is the subtopic ranked at the i^{th} position and D_i is the collection containing the top i diversified subtopics. The function $Nov(S, D_i)$ tries to measure the novelty of S given D_i has already been chosen and ranked. In our approach, we implement the novelty function as follows:

$$Nov(S, D_i) = -max_{S' \in D_i} Sim(S, S')$$

We apply Jaccard Similarity between two sets to calculate the similarity between S and S'. First, a subtopic can be represented by a set of terms; after that, we apply Jaccard Similarity to calculate the similarity between S and S':

$$Sim(S, S') = \frac{S \cap S'}{S \cup S'}$$

Finally, we find the maximum among the similarity values between S and all $S' \in D_i$, and take its opposite number as the novelty score. The top ranked diversified subtopics in D are considered as second-level subtopics.

3.2.3 Hierarchy Construction

We apply k-means clustering algorithm on the second-level subtopic feature vectors by setting $k = 5$. Each cluster is labelled by the frequently occurring terms of subtopics, belonging to it. The weight of each cluster is computed as the average weight of all subtopics underlying it. The label of each cluster is considered as the first-label subtopic. We take at most five first-level subtopics, whereas, under each first-level subtopic, we choose at most top ten second-level subtopics. Therefore, we construct a two-level hierarchy of subtopics.

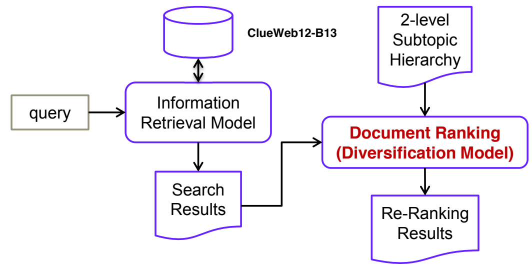

Figure 3: Diversified document ranking framework

3.3 Document Ranking

This section describes our approach for diversifying the original retrieval results. As a diversification model for document ranking, we use the Explicit Web Aspect Diversification (xQuAD) [23] framework. Our document ranking framework is depicted in Fig. 3.

Given a query, we retrieve top 1000 document using the Clueweb12-B13 search interface [1]. For each document, we extract the document rank and snippet.

Let R_q be the initial ranking of document retrieved for the given query q. The diversification model, xQuAD is defined as follows:

$$f_{xQuAD}(q, d, D_q) = (1-\lambda)\ p(d|q) + \lambda \sum_{s \in S_q} p(s|q)\ p(d|q,s) \prod_{d_j \in D_q} (1 - p(d_j|q,s)) \tag{1}$$

Here, $d \in R_q$ is a document, s is a subtopic, S_q is the set of subtopics explicitly mined for the given query q, D_q is the set of diversified documents.

In the Equation 1, there are three main components, that are, document relevance given the query: $p(d|q)$, subtopic importance given the query: $p(s|q)$, and document coverage given the query and subtopic: $p(d|q,s)$.

In our approach, we estimate the document relevance, $p(d|q)$ from the rank in the original retrieved list as $1/\sqrt{(}rank(d))$. For subtopic importance, $p(s|q)$, we use the score produced in the subtopic mining subtask. To compute the document coverage, $p(d|q,s)$, we extract the unigram and bigram features from the document snippet and subtopic separately in two feature vectors. We compute the cosine similarity between these two feature vectors, which is assigned as the document coverage.

4. EXPERIMENTS

We submitted five runs to the English Subtopic Mining subtask. The configurations of the Subtopic Mining runs are stated in the Table 1. We selectively combined the different resources and apply different methods to generate a subtopic mining run. For instance, a run "SEM13-S-E-1A", which was

[1] http://lemurproject.org/clueweb12/services.php

Table 1: Subtopic mining subtask run description

Run	Resources	Methods
SEM13-S-E-1A	Query suggestion, dimension, freebase	MMR, Stemming
SEM13-S-E-2A	Query suggestion, dimension, freebase	Stemming
SEM13-S-E-3A	Query suggestion, dimension, freebase	MMR
SEM13-S-E-4A	Query suggestion, dimension	MMR
SEM13-S-E-5A	Query suggestion, dimension	Cluster labeling

Table 2: Document ranking subtask run description

Run	Subtopics	description
SEM13-D-E-1A	SEM13-S-E-1A	Indri adhoc, 30 subtopics, xQuAD
SEM13-D-E-2A	SEM13-S-E-2A	Indri adhoc, 40 subtopics, xQuAD
SEM13-D-E-3A	SEM13-S-E-3A	Indri adhoc, 20 subtopics, xQuAD
SEM13-D-E-4A	SEM13-S-E-4A	Indri adhoc, 35 subtopics, xQuAD
SEM13-D-E-5A	SEM13-S-E-5A	Indri adhoc, 25 subtopics, xQuAD

produced by extracting the subtopic candidates from the query suggestions, query dimensions, and Freebase entities, and ranked by estimating the multiple features as described in the section 3.1.2, followed by diversifying the subtopics using the maximum marginal relevance (*MMR*). In this run, we also stemmed the subtopics with Krovertz stemmer [12].

We also submitted five runs to the English Document Ranking subtask. The configurations of the Document Ranking runs are stated in the Table 2. We selectively used the different number of subtopics to a document ranking run. For example, a run "SEM13-D-E-1A", which was produced by diversifying the initial ranked documents based on the xQuAD framework by considering the 30 subtopics of the query.

4.1 Evaluation Metric

Subtopic Mining runs are evaluated by estimating the $Hscore$, $Fscore$, $Sscore$, and $H-measure$ metrics. The detail description of these metrics are introduced here [13]. $Hscore$ measures the quality of the subtopic hierarchy, $Fscore$ measures the quality of the first-level subtopics, and $Sscore$ measures the quality of the second-level subtopics. On the other hand, Document ranking runs are evaluated by estimating the $D\#-nDCG$@10 metric.

4.2 Experimental Results

The official evaluation results of our submitted subtopic mining runs are stated in the Table 3. The $Hscore$ of our submitted runs are worse than other participants' runs [13]. In our view, the clusterring approach is not a good idea to generate the subtopic hierarchy. The evaluation results of our submitted document ranking runs are stated in the Table 4. It shows that our system produces a comparable D#-nDCG10 with other participants' runs.

Table 3: Results of subtopic mining run

Run	Hscore	Sscore	Sscore	H-measure
SEM13-S-E-1A	0.1762	0.3043	0.3689	0.0634
SEM13-S-E-2A	0.1844	0.3174	0.3566	0.0610
SEM13-S-E-3A	0.1860	0.2882	0.3333	0.0606
SEM13-S-E-4A	0.1672	0.2056	0.3039	0.0501
SEM13-S-E-5A	0.1580	0.2511	0.3285	0.0470

Table 4: Results of document ranking run, D#-nDCG@10

Run	Coarse-Gain (1st level subtopics)	Fine-grain (2nd level subtopics)
SEM13-D-E-1A	0.6022	0.5291
SEM13-D-E-2A	0.4495	0.3806
SEM13-D-E-3A	0.4735	0.3985
SEM13-D-E-4A	0.3227	0.2505
SEM13-D-E-5A	0.3081	0.2414

5. CONCLUSIONS AND FUTURE WORK

This paper proposed the methods for Subtopic Mining and Document Ranking subtasks in the NTCIR-11 IMINE Task. In Subtopic Mining Subtask, multiple resources are exploited to mine diversified subtopics using clustering based methods for a given query. In Document Ranking Subtask, explicit search result diversification based method are applied to re-rank the results retrieved by Indri search interface using the subtopics mined in the subtopic mining subtask. A set of experiments is carried out to verify the effectiveness of our proposed system. The performace of our subtopic mining methods are not good in compare to other participants' methods. The performce of our document ranking method is comparable to other participants' methods. In future, we would like to utilize other resources and efficient methods to effectively construct a subtopic hierarchy. We also would like to extend the xQuAD diversification model to boost the performance of the diversified document ranking.

6. REFERENCES

[1] G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi. Fub, iasi-cnr and university of tor vergata at trec 2008 blog track. Technical report, DTIC Document, 2008.

[2] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 95–104. ACM, 2011.

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

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

[5] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on*

Research and development in information retrieval, pages 335–336. ACM, 1998.

[6] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the trec 2009 web track. Technical report, DTIC Document, 2009.

[7] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2008.

[8] Z. Dou, S. Hu, Y. Luo, R. Song, and J.-R. Wen. Finding dimensions for queries. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1311–1320. ACM, 2011.

[9] C. Fellbaum. *WordNet*. Wiley Online Library, 1998.

[10] Google. Freebase data dumps, 2014.

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

[12] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202. ACM, 1993.

[13] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M. Kato, H. Ohshima, and K. Zhou. Overview of the ntcir-11 imine task. *In Proc. of NTCIR*, pages 1–16, 2014.

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

[15] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM, 2005.

[16] D. Metzler and T. Kanungo. Machine learned sentence selection strategies for query-biased summarization. In *SIGIR Learning to Rank Workshop*, pages 40–47, 2008.

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

[18] T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Advances in Information Retrieval*, pages 222–234. Springer, 2014.

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

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

[21] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

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

[23] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890. ACM, 2010.

[24] R. L. Santos, R. M. McCreadie, C. Macdonald, and I. Ounis. University of glasgow at trec 2010: Experiments with terrier in blog and web tracks. In *TREC*, 2010.

[25] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing with compositional vector grammars. In *In Proceedings of the ACL conference*. Citeseer, 2013.

[26] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM, 1999.

[27] R. Song, Z. Luo, J.-Y. Nie, Y. Yu, and H.-W. Hon. Identification of ambiguous queries in web search. *Information Processing & Management*, 45(2):216–229, 2009.

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

[29] K. Spärck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: implications for retrieval tests, systems and theories. In *ACM SIGIR Forum*, volume 41, pages 8–17. ACM, 2007.

[30] C.-J. Wang, Y.-W. Lin, M.-F. Tsai, and H.-H. Chen. Mining subtopics from different aspects for diversifying search results. *Information retrieval*, 16(4):452–483, 2013.

[31] Q. Wang, Y. Qian, R. Song, Z. Dou, F. Zhang, T. Sakai, and Q. Zheng. Mining subtopics from text fragments for a web query. *Information retrieval*, 16(4):484–503, 2013.

[32] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.

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