PAPER

A Bipartite Graph-based Ranking Approach to Query Subtopics Diversification Focused on Word Embedding Features

MD ZIA ULLAH[†a)], *Nonmember and* **MASAKI AONO**[†b)], *Member*

SUMMARY

Web search queries are usually vague, ambiguous, or tend to have multiple intents. Users have different search intents while issuing the same query. Understanding the intents through mining subtopics underlying a query has gained much interest in recent years. Query suggestions provided by search engines hold some intents of the original query, however, suggested queries are often noisy and contain a group of alternative queries with similar meaning. Therefore, identifying the subtopics covering possible intents behind a query is a formidable task. Moreover, both the query and subtopics are short in length, it is challenging to estimate the similarity between a pair of short texts and rank them accordingly. In this paper, we propose a method for mining and ranking subtopics where we introduce multiple semantic and content-aware features, a bipartite graph-based ranking (*BGR*) method, and a similarity function for short texts. Given a query, we aggregate the suggested queries from search engines as candidate subtopics and estimate the relevance of them with the given query based on word embedding and content-aware features by modeling a bipartite graph. To estimate the similarity between two short texts, we propose a *Jensen-Shannon divergence* based similarity function through the probability distributions of the terms in the top retrieved documents from a search engine. A diversified ranked list of subtopics covering possible intents of a query is assembled by balancing the relevance and novelty. We experimented and evaluated our method on the NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 subtopic mining test collections. Our proposed method outperforms the baselines, known related methods, and the official participants of the INTENT-2 and IMINE-2 competitions.

key words: *Subtopic Mining, Query Intent, Diversification, Word Embedding, Bipartite Graph*

1. Introduction

When an information need is being formulated in a user's mind, the user issues a query as a sequence of words and submits it to the search engine. The search engine responds with a ranked list of snippet results to meet the request of users. According to user search behaviour analysis, query is usually vague, ambiguous, or tends to have multiple intents [1], [2]. Issuing the same query, different users may have different search intents, which correspond to different subtopics [3].

With an ambiguous query such as "eclipse," users may seek different interpretations including "eclipse IDE," "eclipse lunar," and "eclipse movie." With a broad query such as "programming languages," users may be interested in different subtopics including "programming languages java," "programming languages python," and "programming languages tutorial." However, it is not clear which subtopic of a broad query is actually desirable for a user [4]. In some cases, subtopics underlying a query can be temporally ambiguous; for example, the query "US Open" is more likely to be targeting the tennis open in September and the golf tournament in June [5].

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 diversified intents underlying a query, information retrieval models might result in top ranked documents possibly containing too much relevant information on a particular aspect[†] of a query. As these documents cover a few subtopics or interpretations, the user is eventually unsatisfied. In order to satisfy the user, a sensible approach is to diversify the documents considering the possible subtopics of the query [6]. The diversified retrieval models should result in a ranked list of documents that provides the maximum coverage and minimum redundancy with respect to the possible subtopics.

Identifying the subtopics underlying a query has gained much interest in recent years [7], [8]. Several methods were proposed for mining subtopics from different resources including the top retrieved documents, anchor text, query logs, Wikipedia, Freebase [9], and the related search services provided by the commercial search engines [4], [10], [11]. Query suggestions provided by commercial search engines hold some intents [10], [12]; however, suggested queries are often noisy and contain a group of similar suggestions covering a single aspect of the original query. Since both query and subtopics are short in length, it is challenging to efficiently estimate the similarity between a pair of short texts and rank them accordingly. Therefore, identifying the subtopics covering possible intents underlying a query is a formidable task.

In this paper, we address the problem of *query subtopic mining* [13], which is defined as: "given a query, list up its possible subtopics which specialize or disambiguate the search intent of the original query." In our approach, we make use of the query suggestions provided by search engines to mine and rank the subtopics covering the possible diversified intents of the query. To estimate the relevance scores of the candidate subtopics with the query, we extract multiple word embedding and content-aware features

Manuscript received May 6, 2016.
Manuscript revised August 2, 2016.

[†]Department of Computer Science and Engineering, Toyohashi University of Technology
a) E-mail: arif@kde.cs.tut.ac.jp
b) E-mail: aono@tut.jp

DOI: 10.1587/transinf.E0.D.1

[†]Intent and Aspect are interchangeably used

Copyright © 200x The Institute of Electronics, Information and Communication Engineers

followed by a supervised feature selection, and introduce a bipartite graph-based ranking (*BGR*) method. Then, we diversify the candidate subtopics with maximal marginal relevance (*MMR*) [14] model by balancing the relevance with the query and the novelty with other candidate subtopics. Here, novelty is derived from *Jensen-Shannon divergence*, tuned for short texts through the probability distributions of terms in the top retrieved documents from a search engine. Experimental results on the publicly available test collections including NTCIR-10 INTENT-2 [7] and NTCIR-12 IMINE-2 [15] demonstrate that our proposed method outperforms the baselines, known previous works, and the official participants of the INTENT-2 and IMINE-2 competitions.

To summarise, our main contributions are threefold: (1) some new features based on word embedding (in Section 3.2), (2) a bipartite graph-based ranking (*BGR*) method (in Section 3.3.2), and (3) a new similarity function derived from *Jensen-Shannon divergence*, tuned for short texts through the probability distributions of terms in the top retrieved documents from a search engine (in Section 3.3.3).

The rest of the paper is organized as follows. **Section** 2 overviews related work on query subtopic mining. We introduce our proposed method in **Section** 3. **Section** 4 discusses the overall experiments and results that we obtained. Finally, concluding remarks and some future directions of our work are described in **Section** 5.

2. Related Work

Web queries are usually short, ambiguous, and/or underspecified [1], [2], [16]. To understand the meaning of queries, researchers define taxonomies and classify queries into predefined categories. Song et al. [1] classified queries into three categories: ambiguous queries, which have more than one meaning; board queries, which cover a variety of subtopics; and clear queries, which have a specific meaning or narrow topics.

At the query level, Broder [17] divided query intent into navigational, informational, and transactional types. Nguyen and Kan [18] classified queries into four general facets including ambiguity, authority, temporal sensitivity, and spatial sensitivity. Boldi et al. [19] created a query-flow graph with query phrase nodes and used them for query suggestion. Query suggestion is a key technique for generating alternative queries to help users drill down to a subtopic of the original query [20], [21]. Different from query suggestion, subtopic mining focuses more on the diversity of possible subtopics of the original query rather than inferring relevant queries.

Wu et al. [22] mined query subtopic from questions in the community question answering (CQA) by proposing non-negative matrix factorization (NMF) to cluster the questions and extract keywords from the cluster. Hu et al. [23] leveraged the knowledge contained in Wikipedia to predict the possible subtopics for a given query. Radlinkshi et al. [24] proposed a method for inferring query intents from query reformulations and user click-through data. Santos et al. [10] exploited the query completions of the search engine to mine sub-queries (i.e. subtopics) for diversifying Web search results.

Wang et al. [4] proposed a method to mine subtopics of a query either directly from the query itself or indirectly from the top retrieved documents. In the direct approach, several external resources, such as Wikipedia, open directory project (*ODP*), query logs, and the related search services are investigated to mine subtopics. In the indirect approach, subtopics are extracted by clustering, topic modeling, and concept-tagging of the top retrieved documents. The surrounding text of query terms in the top retrieved documents were also utilized for mining and ranking subtopics [11].

Moreno et al. [25] proposed an algorithm called Dual C-Means to cluster search results in dual-representation spaces with query logs and represented the cluster label as subtopic. Damien et al. [26] proposed a method for subtopic mining and ranking by fusing multiple resources. Kim et al. [27] proposed a frequent pattern-based method to mine candidate subtopics from a set of implicitly relevant documents. Our proposed method has the same premise as Moreno et al. [25], Damien et al. [26], and Kim et al. [27].

Two-level hierarchical intents based search result diversification methods were also proposed [12]. The method, proposed by Kim et al. [28] was to mine a two-level subtopic hierarchy based on hierarchical search intentions. However, our approach is to mine the flat list of the subtopic.

Neural network-based method, *Word2Vec* was proposed by Mikolov et al. [29] which represents a word in semantic space as vector is called word embedding. Words that are semantically and syntactically similar tend to be close in this embedding space. Despite the fact that previous methods leveraged many resources to mine candidate subtopics, however, their ranking methods caused some noisy and redundant subtopics in the top rank. In compare to previous methods, we introduce word embedding to extract semantic features and effectively diversify the candidate subtopics covering possible intents of the query by balancing the relevance and novelty.

NTCIR[†] have been organizing a research competition on *query subtopic mining* in Chinese, English and Japanese languages for the last couple of years, including NTCIR-10 INTENT-2[††], NTCIR-11 IMINE-1[†††], and NTCIR-12 IMINE-2[††††]. Some approaches have been proposed by the participants exploiting multiple resources are discussed in [30]–[34].

3. Diversified Subtopic Mining

In this section, we describe our approach to subtopic mining, which is composed of candidate generation, features extraction, and ranking as depicted in Fig. 1. Given a query,

[†]http://research.nii.ac.jp/ntcir/index-en.html
[††]http://research.microsoft.com/en-us/projects/intent/
[†††]http://www.thuir.org/IMine/
[††††]http://www.dl.kuis.kyoto-u.ac.jp/imine2/

first, we aggregate the query suggestions provided by different search engines as candidate subtopics. Second, to estimate the relevance of the candidate subtopics, we extract multiple semantic and content-aware features followed by a supervised feature selection, and introduce a bipartite graph-based ranking (*BGR*) method. Third, to cover the possible search intents of the query, we produce a diversified ranked list of subtopics by balancing the relevance and the novelty, where novelty is derived from *Jensen-Shannon divergence* tuned for short texts through the probability distributions of terms in the top retrieved documents from a search engine. Our detail method is articulated as follows:

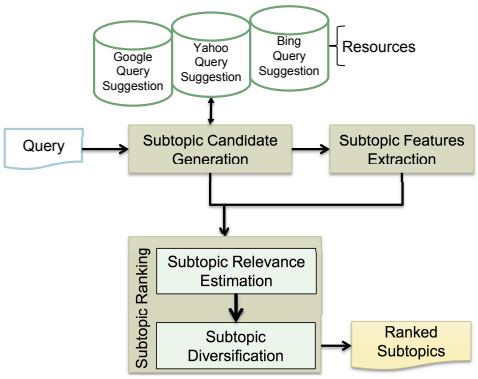

Fig. 1 Diversified Subtopic Mining Flow

3.1 Subtopic Candidate Generation

Inspired by the work of Santos et al. [10], we hypothesize that *suggested queries* in across search engines hold some intents of the query. Given a query, we collect all the suggested queries from search engines, aggregate them by filtering out the duplicates or wrongly represented ones, and consider them as candidate subtopics. To filter out the duplicates or wrongly represented ones, we apply canonicalization on the suggested queries using Krovetz [35] stemmer and remove those ones which are part of the query or exactly similar to the query. For example, given a query “old coins," we generate a list of candidate subtopics including “old coins sell," “old gold coins," “old coins for sale," and “old coins prices."

3.2 Subtopic Features Extraction

Let $q \in Q$ represents a query and $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ represents a set of candidate subtopics generated in Section 3.1. Both query and candidate subtopics are short in length and it is a challenging task to estimate the similarity between a pair of short texts. We extract multiple local and global features, which are broadly organized as word embedding and content-aware features. While the content-aware features are standard features commonly used in the literature for learning to rank for Web search [36], the word embedding based features are specifically proposed to estimate the relevance of candidate subtopics for a query.

We propose three semantic features based on word embedding and make use of the *word2vec*[†] model [29]. In order to capture the importance of the semantic matching of a query with a subtopic, we first propose a new feature, the maximum word similarity (*MWS*) as follows:

$$f_{MWS}\ (q, s) = \frac{1}{|q|} \sum_{t \in q} sem(t, s) \tag{1}$$

where

$$sem(t, s) = \max_{w \in s} f_{sem}(\vec{t}, \vec{w})$$

where $\vec{t}$ and $\vec{w}$ are the word vector representations from the *word2vec* model, corresponding to two words t and w, respectively. The function f_{sem} returns the cosine similarity between two word vectors.

To measure the global importance of a query with a subtopic, we propose our second feature, the mean vector similarity (*MVS*) as follows:

$$f_{MVS}\ (q, s) = f_{sem}(\frac{1}{|q|} \sum_{t \in q} \vec{t}, \frac{1}{|s|} \sum_{w \in s} \vec{w}) \tag{2}$$

Our third proposed feature, the uncommon word similarity (*UWS*) is defined through the similarity of the uncommon word of query and subtopic as follows:

$$f_{UWS}\ (q, s) = f_{sem}(\frac{1}{|q_u|} \sum_{t \in q_u} \vec{t}, \frac{1}{|s_u|} \sum_{w \in s_u} \vec{w}) \tag{3}$$

where query q and subtopic s represent two sets of words, respectively. Then, q_u is defined as $q - (q \cap s)$ and s_u is defined as $s - (q \cap s)$. These three semantic features in Eqs. (1), (2), and (3) are complementary to each other.

Among content-aware features, we extract term frequency, language modeling, term dependency, lexical, and Web hit-count based features. Term frequency (*TF*) based features are directly computed by scoring the occurrences of the terms of query q in a subtopic s. Term frequency based features include *DPH* [37], *PL2* [37], and *BM25* [38]. Language modeling (*LM*) based features include Kullback-Leibler (*KL*) [39], query likelihood with Jelinek-Mercer (*QLM-JM*) [40], subtopic likelihood with Jelinek-Mercer (*SLM-JM*) [40], query likelihood with Dirichlet smoothing (*QLM-DS*) [40], and subtopic likelihood with Dirichlet smoothing (*SLM-DS*) [40]. Term dependency (*TD*) based features include term-dependency Markov random field (*MRF*) [41] and Tri-gram dependency.

To measure the lexical similarity (*LS*) between a query and a subtopic, we extract features based on edit distance (*EDS*), sub-string match (*SSM*) [42], term overlap (*TO*) [42], term synonym overlap (*TSO*) [42], vector space model (*VSM*), and coordinate level matching (*CLM*) [43].

[†] *word2vec* (https://code.google.com/p/word2vec/)

If a subtopic is frequently mentioned in the Web pages, then that subtopic might be important than others. According to this intuition, we make use of search engine hit count (*HC*) to estimate features including normalized hit count (*NHC*), point-wise mutual information (*PMI*), and word co-occurrence (*WC*). To encode a prior knowledge (*PK*) about individual subtopic, we also extract simple query independent features including voting, reciprocal rank (*RR*), average term length (*ATL*), topic cohesiveness (*TC*) [44], and subtopic length (*SL*). For each query-subtopic pair, 27 features are extracted based on word embedding and content-aware relevance as stated in Table 1.

Table 1 Word embedding and content-aware based features in this work.

Features	Type	Total
MWS, MVS, UWS	*Word2Vec*	3
DPH, PL2, BM25	*TF*	3
KL, QLM-JM, SLM-JM, QLM-DS, SLM-DS	*LM*	5
MRF, Tri-Gram	*TD*	2
EDS, SSM, TO, TSO, VSM, CLM	*LS*	6
NHC, PMI, WC	*HC*	3
Voting, RR, ATL, TC, SL	*PK*	5
		27

3.3 Subtopic Ranking

For a pair of query q and candidate subtopic s, we extract all the features described in Section 3.2 and represent those in a feature vector, $\mathcal{F}_{q,s} = \{f_{DPH}(q,s), f_{PL2}(q,s), \cdots, f_{UWS}(q,s)\}$. Therefore, for a query q, we have a feature matrix, $\mathcal{MF} = \{\mathcal{F}_{s_1}, \mathcal{F}_{s_2}, \cdots, \mathcal{F}_{s_k}\}$, corresponding to a set of candidate subtopics, $\mathcal{S} = \{s_1, s_2, ..., s_k\}$. We normalize each feature vector using *MinMax* normalization technique. To estimate the diversified rank of a subtopic, first, we employ a supervised feature selection method to remove noisy and redundant features. Second, we introduce a bipartite graph-based ranking approach for estimating the relevance of the candidate subtopic. Finally, we make use of *MMR* model to produce a diversified ranked list of subtopics covering the possible intents of the query by balancing the relevance of the candidate subtopic with the query and novelty with other subtopics.

3.3.1 Supervised Feature Selection

Supervised feature selection is an important technique to determine the best set of features by reducing the noisy, redundant or highly correlated features in a large feature set. We make use of elastic-net regularized regression method due to its better performance over Lasso and Ridge regression [45]. Given a parameter α strictly between 0 and 1, and a nonnegative λ, elastic-net solves the following optimization problem:

$$\min_{\beta_0,\beta}\left(\frac{1}{2M}\sum_{i=1}^{M}(y_i-\beta_0-\mathcal{F}_i^T\beta)^2+\lambda\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2+\alpha\|\beta_j\|\right)\right) \tag{4}$$

where M is the number of samples, $\mathcal{F}_i^T$ is the transpose of feature vector of the i-th sample, and $y_i \in \{0,1\}$ is the label of the i-th sample. In our case, each sample is a query-subtopic pair. We train elastic-net on query-subtopic pairs' feature vectors and choose those features whose coefficients β are positive.

3.3.2 Subtopic Relevance Estimation

To estimate the relevance of the candidate subtopics for a query, we introduce a bipartite graph-based ranking (*BGR*) approach with considering the features selected by elastic-net in section 3.3.1.

(1) Bipartite Graph based Ranking

Many real applications can be modeled as a bipartite graph, including Video shots and Tags [46], Queries and URLs, Entities and Co-List [47] in a Web page, Phenotypes and Diseases [48].

We hypothesize that a relevant subtopic should be ranked at the higher position by multiple effective features, and intuitively, an effective feature should be weighted higher by multiple relevant subtopics. Large weight should be given to a subtopic that tends to be ranked highly by a group of effective features, and vice versa. Therefore, there is a weight propagation of features to subtopics and subtopics to features. On these intuitions, we represent a set of features and a set of candidate subtopics as a bipartite graph and introduce weight propagation from both sides of the bipartite graph. Given a set of features and a set of candidate subtopics, we propose a bipartite graph-based ranking (*BGR*) method to estimate the global importances of candidate subtopics by aggregating the local importance of the individual feature.

Let $\mathcal{R} = \{r_1, r_2, \cdots, r_m\}$ be a set of features and $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$ be a set of candidate subtopics. Consider $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a bipartite graph where vertex set $\mathcal{V}$ composed of two disjoint sets $\mathcal{R}$ and $\mathcal{S}$ such that each edge in $\mathcal{E}$ connects a vertex in $\mathcal{R}$ to a vertex in $\mathcal{S}$; that is, there is no edge between two vertices in the same set. For each feature $r_i \in \mathcal{R}$, we obtain a ranked list of subtopics L_{r_i}. The weight w_{ij} of the edge between two vertices $r_i \in \mathcal{R}$ and $s_j \in \mathcal{S}$ is defined as follows:

$$w_{ij} = \frac{1}{\sqrt{\log_2(rank(L_{r_i}, s_j) + 2.0)}} \tag{5}$$

where the function, rank(L_{r_i}, s_j) returns the position of the subtopic s_j in the ranked list L_{r_i} for the feature r_i. The reciprocal rank of the subtopic is considered to assign high importance to the subtopic in the higher rank position.

The bipartite graph $\mathcal{G}$ is represented as a bi-adjacency $m \times n$ matrix, $\mathcal{M}$. We assign an initial weight to each vertex in the bipartite graph $\mathcal{G}$. Therefore, the weight vector corresponding to the vertices in the set of features $\mathcal{R}$ is R, which is initialized as uniform values (i.e. $1/m$ where m is the number of features). The weight vector corresponding to

the vertices in the set of subtopics $\mathcal{S}$ is S, which is initialized by applying either the Eq. (1) or (2) between a query and a candidate subtopic.

For a bipartite graph, there is a natural random walk on the graph with the transition probability from both sides [49]. The transition matrix from $\mathcal{R}$ to $\mathcal{S}$ is defined as $\mathcal{W}_1 = \mathcal{D}_{\mathcal{R}}^{-1}\mathcal{M}$, where $\mathcal{D}_{\mathcal{R}}$ is the diagonal matrix with its (i, i)-element equal to the sum of the i-th row of $\mathcal{M}$. Similarly, the transition matrix from $\mathcal{S}$ to $\mathcal{R}$ is defined as $\mathcal{W}_2 = \mathcal{D}_{\mathcal{S}}^{-1}\mathcal{M}^T$, where $\mathcal{D}_{\mathcal{S}}$ is the diagonal matrix with its (j, j)-element equal to the sum of the j-th column of $\mathcal{M}$.

The weight propagation from the set of candidate subtopics $\mathcal{S}$ to the set of features $\mathcal{R}$ is represented as follows:

$$R_{k+1} = \lambda_1 \mathcal{W}_1 S_k + (1 - \lambda_1) R_0 \tag{6}$$

where $0 < \lambda_1 < 1$ is a parameter, which is used to combine the initial and the propagated scores of the features. R_0 is the vector of the initial scores of the features, S_k denotes the vector of the subtopics score after k-th iterations, and R_{k+1} denotes the vector of the features score after $(k + 1)$-th iterations.

Similarly, the weight propagation from the set of features to the set of candidate subtopics is represented as follows:

$$S_{k+1} = \lambda_2 \mathcal{W}_2 R_{k+1} + (1 - \lambda_2) S_0 \tag{7}$$

where $0 < \lambda_2 < 1$ is a parameter, which is used to combine the initial and the propagated scores of the candidate subtopics. S_0 is the vector of the initial scores of the subtopics and S_{k+1} denotes the vector of the subtopics score after $(k + 1)$-th iterations.

Considering the vertices in $\mathcal{S}$, by substituting Eq. (6) for R_{k+1} in Eq. (7), the weight equation of subtopics is represented as follows:

$$\begin{aligned} S_{k+1} &= \lambda_2 \mathcal{W}_2 [\lambda_1 \mathcal{W}_1 S_k + (1 - \lambda_1) R_0] + (1 - \lambda_2) S_0 \\ &= \lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1 S_k + \lambda_2 (1 - \lambda_1) \mathcal{W}_2 R_0 + (1 - \lambda_2) S_0 \end{aligned} \tag{8}$$

The closed form of Eq. (8) is defined as follows:

$$\begin{aligned} S_{k+1} = (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{k+1} S_0 + \sum_{t=0}^{k} (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^t \\ [\lambda_2 (1 - \lambda_1) \mathcal{W}_2 R_0 + (1 - \lambda_2) S_0] \end{aligned} \tag{9}$$

Since Eigen values of the stochastic matrices $\mathcal{W}_1$ and $\mathcal{W}_2$ are in [-1, 1], the Eq. (9) is converged. Therefore,

$$\begin{aligned} \lim_{k \to \infty} (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{k+1} = 0 \\ \lim_{k \to \infty} \sum_{t=0}^{k} (\lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^t = (I - \lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{-1} \end{aligned} \tag{10}$$

When $k \to \infty$ and $S_{k+1} \to S^*$, we have

$$S^* = (I - \lambda_1 \lambda_2 \mathcal{W}_2 \mathcal{W}_1)^{-1} [(1 - \lambda_1) \lambda_2 \mathcal{W}_2 R_0 + (1 - \lambda_2) S_0] \tag{11}$$

Given λ_1, λ_2, $\mathcal{W}_1$, $\mathcal{W}_2$, R_0, and S_0, we estimate the scores S^* directly by applying Eq. (11). These scores S^* are considered as the relevance scores, rel(q, $\mathcal{S}$) of a set of candidate subtopics $\mathcal{S}$ with the query q and eventually utilized in subtopic diversification.

3.3.3 Subtopic Diversification

To produce a diversified ranked list of subtopics by balancing the relevance and novelty, we make use of *MMR* model. The *MMR* regards the ranking problem as a procedure of successively selecting the “best" unranked subtopic. When searching for the next best subtopic, the *MMR* model chooses not the most relevant one, however, the one that best balances the relevance and novelty. Novelty means that a subtopic is new compared to those already selected and ranked.

Given a relevance function *rel* (., .) and a novelty function *novelty* (., .), the *MMR* model can be defined as follows:

$$s_i^* = \underset{s_i \in D \setminus C_i}{\operatorname{argmax}} \ \gamma \ rel\ (q, s_i) + (1 - \gamma)\ novelty\ (s_i, C_i) \tag{12}$$

where γ is a combining parameter and $\gamma \in [0, 1]$. D is the relevance oriented ranked list of subtopics retrieved by Eq. (11). C_i is the collection of subtopics that have already been selected at the i-th iteration and initially empty. Then,

$$C_{i+1} = C_i \cup \{s_i^*\}$$

where s_i^* is the subtopic ranked at the i-th position. The function, *novelty* (s_i, C_i) tries to measure the novelty of subtopic s_i given the collection C_i.

We find the maximum similarity value for subtopic s with all the selected subtopics $s' \in C_i$, and flip the sign as the novelty score as follows:

$$novelty\ (s_i, C_i) = -\max_{s' \in C_i} \ sim\ (s_i, s') \tag{13}$$

Both subtopics s and s' are short in length and they might not be lexically similar. Our observation is that if two subtopics represent the same meaning, even though they are not lexically similar, they may retrieve similar kinds of documents from a search engine. Mutual information between two probability distributions of words may represent the similarity between two subtopics. Therefore, we propose a function based on *Jensen-Shannon divergence* to estimate the similarity between two subtopics s and s' as follows:

$$sim\ (s, s') = 1.0 - JSD(s, s') \tag{14}$$

where $JSD(s, s')$ is estimated through the word probability distributions P_s and $Q_{s'}$ of the top-K retrieved documents from the Web search engine for subtopics s and s', respectively. $JSD(s, s')$ is defined as follows:

$$JSD(P_s, Q_{s'}) = \frac{1}{2} \sum_{i=1}^{|V|} (P(i) \log \frac{P(i)}{T(i)} + Q(i) \log \frac{Q(i)}{T(i)})$$

(15)

where $T = \frac{1}{2}(P + Q)$. V is the set of words in the vocabulary, collected from the titles and snippets of the documents corresponding to two subtopics s and s'. We choose *Jensen-Shannon divergence* over *Kullback-Leibler divergence*, because of its symmetric similarity estimate.

4. Experiments and Discussion

In this section, we evaluate our proposed method with different settings on the NTCIR-10 INTENT-2 [13] and NTCIR-12 IMINE-2 [15] English Subtopic Mining test collections and compare the performance with the previous works.

4.1 Dataset

The NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 English Subtopic Mining test collections include a set of 50 and 100 topics (i.e. queries), respectively. Each topic is labelled by a set of intents with probabilities and for each intent, there is a set of subtopics as relevance judgement. The statistics of the topics, intents, and subtopics of the INTENT-2 and IMINE-2 datasets are stated in Table 2. For example, Figure 2 shows that the topic “grilling" has several intents with probabilities and there is a set of subtopics as examples for each intent.

Table 2 Statistics of Topics, Intents, and Subtopics of NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 Datasets

	INTENT-2	**IMINE-2**
Topics	50	100
Intents	392	533
Subtopics	5,410	1652

Both INTENT-2 and IMINE-2 organizers collected query suggestions and query completions from *Google*, *Yahoo*, and *Bing* search engines corresponding to the set of topics and included in the datasets as resources for fairly comparing the methods using these datasets. We made use of the query suggestions that were included in INTENT-2 and IMINE-2 datasets. For estimating semantic features, specifically for Eqs. (1), (2), and (3), we made use of the pre-trained distributed word representation of *word2vec* trained on the *Google news corpus*[†]. To estimate some global features including inverse document frequency (*IDF*) and corpus frequency (*CF*), we indexed the *Clueweb09 CatB* [50] corpus employing Indri Search Engine [51] and utilized accordingly. For estimating the novelty function in Eq. (13), we query the *Bing Search API*[††] and collect the top-50 documents corresponding to all the topics and the candidate subtopics.

4.2 Evaluation Metrics

We evaluated our proposed method by estimating I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10. I-rec@10 measures the diversity of the returned subtopics, which shows how many percentages of intents can be found. D-nDCG@10 measures the overall relevance across all intents considering the subtopic ranking. D#-nDCG is a combination of I-rec@10 (50%) and D-nDCG@10 (50%). It is used as the primary evaluation metric by the INTENT-2 [7] and IMINE-2 [15] task organizers. The advantages of D#-nDCG@10 over other diversity metrics (e.g. a-nDCG and Intent-Aware metrics) are discussed in [7]. In our experiments, we utilized NTCIREVAL [52], the tool provided by the NTCIR organizers to compute the above three metrics, in which the default setting was used. Moreover, we made use of two-tailed paired t-test for statistical significance analysis where the significance level is 0.05 [53].

[†]https://code.google.com/word2vec/

[††]https://datamarket.azure.com/dataset/bing/search

```
<topic number="0410">
    <query>grilling</query>
    <intent number="1" probability="0.175000">
      <description>grilling recipes</description>
        <examples>summer grilling recipes,…</examples>
    </intent>
    <intent number="2" probability="0.165000">
      <description> grilling barbecue</description>
        <examples>meat for grilling,…</examples>
    </intent>
</topic>
```

Fig. 2 A topic “grilling" from INTENT-2 dataset which is labelled by its intents with probabilities and a set of subtopics under each intent.

4.3 Experimental Settings

We designed three experiments to evaluate the usefulness of our proposed method. In experiment 1, we discriminatively evaluated our proposed method by highlighting three contributions on INTENT-2 and IMINE-2 datasets in different settings including *Baseline*, *W2V-LRM-Cosine*, *W2V-BGR-Cosine*, and *W2V-BGR-JSD*. *Baseline* included standard content-aware features, linear ranking method (*LRM*), and cosine similarity-based novelty function. *W2V-LRM-Cosine* extended *Baseline* by including word embedding based features. *W2V-BGR-Cosine* included bipartite graph-based ranking (*BGR*) method in place of *LRM* in *W2V-LRM-Cosine*. Our proposed *W2V-BGR-JSD* included *Jensen-Shannon divergence* based novelty function in place of cosine similarity based novelty function in *W2V-BGR-Cosine*. Moreover, to show the effectiveness of the feature selection, we also evaluated our proposed method with or without considering feature selection. In experiment 2, we compared the performance of our proposed method *W2V-BGR-JSD* on INTENT-2 dataset with the known related methods including Kim et al. [27], Moreno et al. [25], Damien et al. [26], and the baselines including query completions (BingC, GoogleC and YahooC), query suggestion (BingS), and a simple merging strategy (Merge). In experiment 3, we compared the performance of our proposed method *W2V-BGR-JSD* with the official participants of INTENT-2 [7] and IMINE-2 [15] competitions.

4.4 Important Features and Parameter Tuning

To select the important features, we prepared the training samples by choosing 10 queries including the corresponding subtopics from the relevance judgement of INTENT-2 and IMINE-2 datasets, respectively. We extracted in total 27 features and employed elastic-net for feature selection. The selected features for INTENT-2 dataset were as follows: *MWS*, *MVS*, *UWS*, *DPH*, *QLM-JM*, *SLM-DS*, *MRF*, *SSM*, *TSO*, *NHC*, *WC*, *RR*, and *ATL*. Similarly, the selected features for IMINE-2 dataset were as follows: *MWS*, *MVS*, *UWS*, *DPH*, *BM25*, *MRF*, *TO*, *TSO*, *CLM*, *NHC*, *WC*, *Voting*, and *TC*. It turns out that our proposed features *MWS*, *MVS*, and *UWS* were selected for both of the datasets and are important in subtopic ranking.

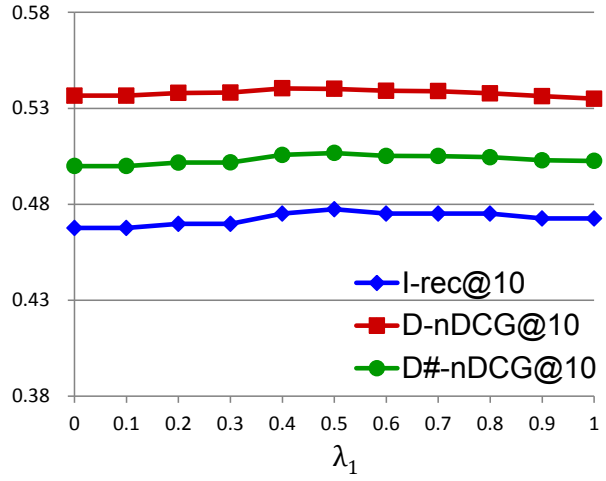

Fig. 3 An empirical analysis of the parameter λ_1 in Eq. (6) for our proposed method bipartite graph based *BGR* ranking on INTENT-2 dataset. The X-axis indicates the parameter λ_1 and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.

There were some optimization parameters in our proposed method, namely λ_1, λ_2, and γ in Eqs. (6), (7), and (12), respectively. With empirical evaluation, we found the optimal values of these parameters for both INTENT-2 and IMINE-2 datasets. To find out the optimal values of λ_1, λ_2, and γ for INTENT-2 dataset, we fixed the λ_2 to 0.5 and changed the value of λ_1 at a rate of 0.1 from 0 to 1. The evaluation result is depicted in Figure 3. It turns out that the curve of diversity measure I-rec is smooth from 0.6 to 0.8 values of λ_1. Therefore, the optimal insensitive value of λ_1 is 0.8, which reflects a high importance of the propagated weights of features than the initial weights in Eq. (6).

Then, we fixed the λ_1 to 0.8 and changed the value of λ_2 at a rate of 0.1 from 0 to 1. The evaluation result is depicted in Figure 4. It demonstrates that the curve of diversity measure I-rec is smooth from 0.3 to 0.5 for values of λ_2. Therefore, the optimal value of λ_2 is 0.4, which reveals a higher importance of the initial weight than propagated weight for candidate

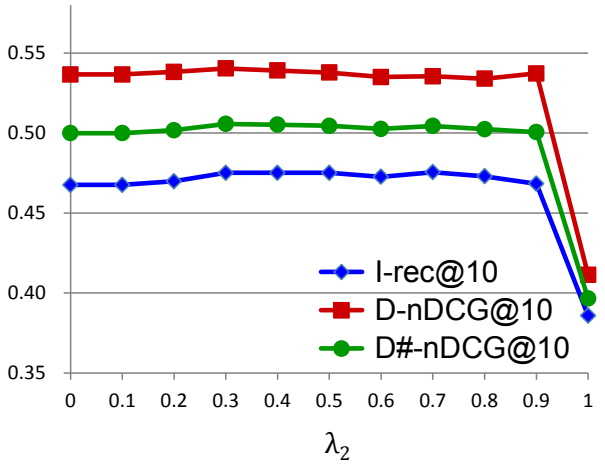

Fig. 4 An empirical study of the parameter λ_2 in Eq. (7). The X-axis indicates the parameter λ_2 and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.

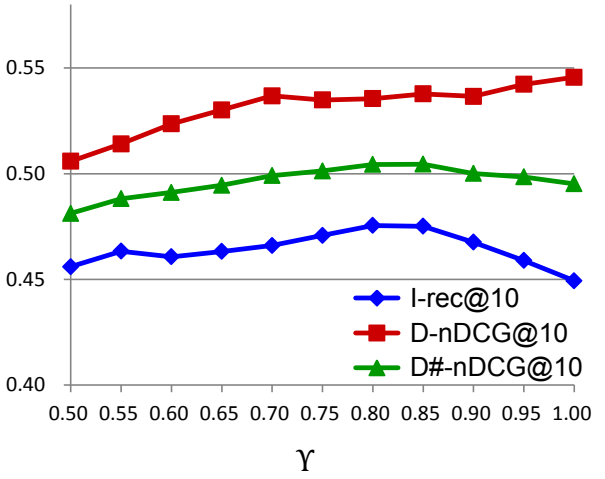

Fig. 5 Sensitivity analysis of the diversification parameter γ. The X-axis indicates the different values of the parameter γ and the Y-axis indicates the corresponding scores of I-rec, D-nDCG, and D#-nDCG at the cutoff rank 10.

subtopics in Eq. (7).

In Eq. (12), the diversification parameter γ balances the relevance and novelty of candidate subtopics. With the optimal values of λ_1 to 0.8 and λ_2 to 0.4, we changed the values of γ at a rate of 0.05 from 0.5 to 1. The evaluation result is depicted in Figure 5. It shows that if we increase the value of γ, both diversity measures I-rec and relevance measure D-nDCG increase. However, around a value of 0.80 of γ, I-rec decreases and D-nDCG increases. We found the highest value of D#-nDCG metric at 0.85 for the parameter γ, which reflected that *MMR* model assigned high scores to relevance than novelty for subtopics diversification. Therefore, the optimal values of the parameters λ_1, λ_2, and γ for INTENT-2 dataset are 0.80, 0.40, and 0.85, respectively. Similarly, we found the optimal values of the parameters λ_1, λ_2, and γ for IMINE-2 dataset are 0.60, 0.70, and 0.80, respectively.

4.5 Experimental Results

To fairly compare with the baselines and previous works, we evaluated our proposed method on INTENT-2 topics by utilizing the parameters learned for IMINE-2 topics. Similarly, we evaluated our proposed method on IMINE-2 topics by utilizing the parameters learned for INTENT-2 topics. In addition, we excluded the topics which were used for feature selection.

Experiment 1: we evaluated our proposed method in different settings including *Baseline*, *W2V-LRM-Cosine*, *W2V-BGR-Cosine*, and *W2V-BGR-JSD* on INTENT-2 and IMINE-2 datasets. For INTENT-2 dataset, the comparative performances are reported in Table 3. It turns out that with two-tailed paired t-test (p<0.05), *W2V-BGR-JSD* significantly outperforms *W2V-BGR-Cosine*, *W2V-LRM-Cosine*, and *Baseline* in terms of D-nDCG and D#-nDCG, however, *W2V-BGR-Cosine* and *W2V-LRM-Cosine* are indistinguishable from *W2V-BGR-JSD* in terms of I-rec.

Table 3 Performance comparison of our proposed method in different settings on NTCIR-10 INTENT-2 dataset at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ⋄ indicates statistically indistinguishable from the best.

Method	I-rec	D-nDCG	D#-nDCG
Our proposed (***W2V-BGR-JSD***)	**0.4745**†	**0.5394**†	**0.5048**†
W2V-BGR-Cosine	0.4616$^{\diamond}$	0.5089	0.4850
W2V-LRM-Cosine	0.4466$^{\diamond}$	0.4780	0.4626
Baseline	0.4152	0.4555	0.4360

For IMINE-2 dataset, the comparative performances are shown in Table 4. With two-tailed paired t-test (p<0.05), it demonstrates that *W2V-BGR-JSD* significantly outperforms *W2V-BGR-Cosine*, *W2V-LRM-Cosine*, and *Baseline* in terms of I-rec, D-nDCG, and D#-nDCG metrics.

Table 4 Performance comparison of our proposed method in different settings on NTCIR-12 IMINE-2 dataset at the cutoff rank 10. The best result is in bold. † indicates statistically significant.

Method	I-rec	D-nDCG	D#-nDCG
Our Proposed (***W2V-BGR-JSD***)	**0.8235**†	**0.6911**†	**0.7581**†
W2V-BGR-Cosine	0.7319	0.6312	0.6837
W2V-LRM-Cosine	0.7216	0.6217	0.6704
Baseline	0.6928	0.5762	0.6377

To show the effectiveness of the feature selection, we described the comparative performances of our proposed method with or without employing feature selection on INTENT-2 and IMINE-2 datasets in Table 5. It reveals that *W2V-BGR-JSD* significantly outperforms its variants without including feature selection for all metrics on IMINE-2 datasets. However, *W2V-BGR-JSD* is statistically indistinguishable from its variants for D-nDCG and D#-nDCG on INTENT-2 dataset.

Experiment 2: we compared our proposed method (*W2V-BGR-JSD*) with the known related methods including Kim et al. [27], Moreno et al. [25], Damien et al. [26], and the baselines including BingC, GoogleC, YahooC, BingS, and Merge for INTENT-2 dataset. The comparative performances are shown in Table 6. Overall, *W2V-BGR-JSD* outperforms the baselines, Kim et al. [27], Moreno et al. [25], and Damien et al. [26] in terms of I-rec, D-nDCG, and D#-nDCG metrics.

Table 5 Performance comparison of our proposed method with/without feature selection (FS) on NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 datasets at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ⋄ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
NTCIR-10 INTENT-2	***W2V-BGR-JSD***	**0.4745**†	**0.5394**	**0.5048**
	Without FS	0.4390	0.5279$^{\diamond}$	0.4794$^{\diamond}$
NTCIR-12 IMINE-2	***W2V-BGR-JSD***	**0.8235**†	**0.6911**†	**0.7581**†
	Without FS	0.7305	0.6245	0.6801

Table 6 Comparative performance of our proposed method with baselines and known previous methods on NTCIR-10 INTENT-2 dataset at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ⋄ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
Our Proposed	***W2V-BGR-JSD***	**0.4745**†	**0.5394**†	**0.5048**†
Baselines	BingS	0.3068	0.2787	0.2928
	BingC	0.3231	0.3268	0.3250
	GoogleC	0.3735	0.3841	0.3788
	YahooC	0.3829	0.3815	0.3822
	Merge	0.3365	0.3181	0.3273
Previous Methods	Kim et al. [27]	0.4032	0.3681	0.3788
	Moreno et al. [25]	0.4249	0.4221	0.4225
	Damien et al. [26]	0.4587$^{\diamond}$	0.3625	0.4106

With two-tailed paired t-test (p<0.05), in terms of diversity (i.e. I-rec), *W2V-BGR-JSD* significantly outperforms the baselines, Kim et al. [27], and Moreno et al. [25], however, Damien et al. [26] is statistically indistinguishable from *W2V-BGR-JSD*. Despite the fact that Kim et al. [27] proposed simple rules and popularity-based ranking, Moreno et al. [25] applied Dual-C means clustering, and Damien et al. [26] employed *Jaccard similarity* based hierarchical clustering to mine candidate subtopics, which often cause irrelevant and redundant candidates, however, our proposed novelty function in Eq. (13), which is derived from *Jensen-Shannon divergence* tuned for short texts eliminates redundant candidate subtopics and benefits for diverse relevant subtopics at the top rank.

In terms of relevance (i.e. D-nDCG), *W2V-BGR-JSD* shows statistically significant improvement over the baselines, Kim et al. [27], Moreno et al. [25], and Damien et al. [26]. To estimate the relevance of the candidate subtopics, our proposed bipartite graph-based ranking method efficiently and effectively approximates the global importances of the subtopics by exploiting the word embedding and content-aware based features. In terms of D#-nDCG, *W2V-BGR-JSD* also significantly outperforms the baselines, Kim et al. [27], Moreno et al. [25], and Damien et al. [26].

Experiment 3: we evaluated our proposed method *W2V-BGR-JSD* by comparing the performance with the official participants' methods of INTENT-2 and IMINE-2 competitions. The comparative performances of *W2V-BGR-JSD* with the participants of INTENT-2 are demonstrated in Table 7. Overall, *W2V-BGR-JSD* outperforms all the official participants' methods in terms of I-rec, D-nDCG, and D#-nDCG metrics.

Table 7 Performance comparison of our proposed method with the official participants of NTCIR-10 INTENT-2 competition at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ⋄ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
Our Proposed	***W2V-BGR-JSD***	**0.4745**†	**0.5394**†	**0.5048**†
NTCIR-10 INTENT-2	THUIR-S-E-4A [30]	0.4364	0.5062⋄	0.4713⋄
	THCIB-S-E-1A [32]	0.4431	0.4657	0.4544
	THUIR-S-E-1A [30]	0.4512⋄	0.4775⋄	0.4644⋄
	hultech-S-E-1A [31]	0.3680	0.5368⋄	0.4524⋄
	KLE-S-E-4A [33]	0.4457⋄	0.4401	0.4429
	SEM12-S-E-2A [34]	0.3777	0.4290	0.4014
	ORG-S-E-4A	0.3815	0.3829	0.3822
	TUTA1-S-E-1A	0.2181	0.2577	0.2379
	LIA-S-E-4A	0.2000	0.2753	0.2376

With two-tailed paired t-test ($p<0.05$), in terms of diversity (i.e. I-rec), *W2V-BGR-JSD* significantly outperforms all the participants' methods except THUIR-S-E-1A and KLE-S-E-4A, which are statistically indistinguishable from *W2V-BGR-JSD*. In terms of relevance (i.e. D-nDCG), *W2V-BGR-JSD* significantly outperforms all the participants' methods except hultech-S-E-1A and THUIR-S-E-4A with two-tailed paired t-test ($p<0.05$). However, hultech-S-E-1A, THUIR-S-E-1A, and THUIR-S-E-4A are statistically identical with *W2V-BGR-JSD* in terms of D-nDCG. In terms of D#-nDCG, *W2V-BGR-JSD* significantly outperforms the participants' methods, however, hultech-S-E-1A, THUIR-S-E-1A, and THUIR-S-E-4A are indistinguishable.

Table 8 Performance comparison of our proposed method with the official participants of NTCIR-12 IMINE-2 competition, baseline, and Kim et al. [27] at the cutoff rank 10. The best result is in bold. † indicates statistically significant and ⋄ indicates statistically indistinguishable from the best.

	Method	I-rec	D-nDCG	D#-nDCG
Our Proposed	***W2V-BGR-JSD***	**0.8235**†	**0.6911**†	**0.7581**†
NTCIR-12 IMINE-2	KDEIM-Q-E-1S	0.7557	0.6644	0.7101
	HULTECH-Q-E-1Q	0.7280	0.6787⋄	0.7033
	RUCIR-Q-E-4Q	0.7601	0.5097	0.6349
Baseline	Merge	0.6144	0.3884	0.5044
Related work	Kim et al. [27]	0.5233	0.3210	0.4242

In addition, the comparative performances of *W2V-BGR-JSD* with the official participants of IMINE-2 competitions†, baseline, and Kim et al. [27] are reported in Table 8. Overall, *W2V-BGR-JSD* outperforms all the official participants' methods, baseline, and Kim et al [27] in terms of I-rec, D-nDCG, and D#-nDCG metrics. With two-tailed paired t-test ($p<0.05$), in terms of diversity (i.e. I-rec), *W2V-BGR-JSD* shows statistically significant performance over all related methods. In terms of relevance (i.e. D-nDCG), *W2V-BGR-JSD* significantly outperforms KDEIM-Q-E-1S, RUCIR-Q-E-4Q, baseline, and Kim et al. [27], however, the difference with HULTECH-Q-E-1Q is insignificant. In terms of D#-nDCG, *W2V-BGR-JSD* also significantly outperforms the related methods.

Experimental results demonstrate that our proposed bipartite graph based ranking (*BGR*) method with semantic features and *Jensen-Shannon divergence* based novelty function consistently performs better than previous works on both INTENT-2 and IMINE-2 datasets.

4.6 Discussion

Taking query "#09 (porteville)" from the INTENT-2 dataset as an example, we listed the top-10 subtopics returned by W2V-BGR-JSD, Baseline (Merge), Kim et al. [27], and Moreno et al. [25] in Table 9. Note that relevant subtopic is labeled by its intent number. It shows that our proposed W2V-BGR-JSD returns 7 relevant subtopics covering 6 intents (out of 7) in the top-10 ranks. Both Baseline (Merge) and Kim et al. [27] return 3 relevant subtopics covering 3 intents. Though Moreno et al. [25] returns 5 relevant subtopics, however, those subtopics are redundant and cover only 3 search intents.

The computation bottleneck in our approach is the feature extraction, bipartite graph-based ranking, and diversification. However, using hash tables, features are extracted in linear time. Though Eq. (11) requires matrix inversion which takes $O(n^3)$ time, however, computation time is negligible for a few number of candidate subtopics (i.e. small n). Since the diversification problem is NP-hard, the greedy algorithm can achieve the optimal ranking in $O(n^2)$, which is still negligible for small n [54]. To satisfy the diverse users, a traditional search engine can be augmented by extending two components: subtopic mining and search diversification. Given a query, a search engine can utilize our method to mine the possible subtopics and diversify the initially retrieved top ranked documents based on the mined subtopics.

The limitations of our proposed method are carefully choosing the features, learning the parameters of bipartite graph and diversification, and estimating the novelty of the subtopic. Moreover, optimally combining subtopics from heterogeneous sources might improve the performance.

5. Conclusion

In this paper, we proposed a method for mining and ranking subtopics of the query. We introduced new features based on word embedding and utilized content-aware features that were selected by a supervised method. The relevance of the candidate subtopic with the query was estimated by introducing a bipartite graph-based ranking (*BGR*) method. For

†http://www.dl.kuis.kyoto-u.ac.jp/imine2/dataset/#results

Table 9 Results of subtopic ranking for the topic "#09 (porterville)". '-' indicates irrelevant subtopic

Rank	W2V-BGR-JSD	Baseline (Merge)	Kim et al. [27]	Moreno et al. [25]
1	porterville california, 5	porterville recorder, -	porterville college, 1	porterville unified school district, 2
2	porterville college, 1	porterville college, 1	porterville unified school district, 2	porterville community college, 1
3	Map of Porterville CA, 7	porterville unified school district, 2	city of porterville, 5	porterville, -
4	porterville recorder, -	porterville high school, -	porterville recorder, -	porterville developmental center, -
5	city of porterville, 5	city of porterville, 5	porterville, -	porterville homes for sale, -
6	Porterville Jobs, -	porterville police department, -	porterville high school, -	porterville high school, -
7	porterville weather, 4	academy trainings in porterville, -	resident of porterville, -	porterville police department, -
8	Porterville Hotels, 3	resident of porterville, -	porterville city council, -	map of porterville ca, 7
9	porterville mls, -	porterville city council, -	copyright 2005 12 by porterville unified school district, -	porterville adult school, 2
10	porterville adult school, 2	porterville breakfast lions, -	porterville police department, -	porterville college, 1

diversifying the candidate subtopics, we introduced a novelty function to estimate the similarity between two candidate subtopics (i.e. short texts) by means of *Jensen-Shannon divergence* through the probability distributions of terms in the retrieved documents from a search engine. We experimented and evaluated our proposed method on NTCIR-10 INTENT-2 and NTCIR-12 IMINE-2 datasets in terms of I-rec, D-nDCG, and D#-nDCG metrics at the cutoff rank 10. We demonstrated that our proposed method significantly outperforms the baselines, the previously known subtopic mining methods [25]–[27], and the official participants of INTENT-2 and IMINE-2 competitions. In future work, we will incorporate subtopics from different resources in our subtopic mining framework and enhance search result diversification to satisfy the users' information needs.

6. Acknowledgments

This research was supported by JSPS Grant-in-Aid for Scientific Research (B) 26280038. We thank the anonymous reviewers for their constructive and insightful comments.

References

[1] R. Song, Z. Luo, J.Y. Nie, Y. Yu, and H.W. Hon, "Identification of ambiguous queries in web search," Information Processing & Management, vol.45, no.2, pp.216–229, 2009.

[2] K. Spärck-Jones, S.E. Robertson, and M. Sanderson, "Ambiguous requests: implications for retrieval tests, systems and theories," ACM SIGIR Forum, vol.41, no.2, pp.8–17, 2007.

[3] P. Ren, Z. Chen, J. Ma, S. Wang, Z. Zhang, and Z. Ren, "Mining and ranking users intents behind queries," Information Retrieval Journal, vol.18, no.6, pp.504–529, 2015.

[4] C.J. Wang, Y.W. Lin, M.F. Tsai, and H.H. Chen, "Mining subtopics from different aspects for diversifying search results," Information retrieval, vol.16, no.4, pp.452–483, 2013.

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

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

[7] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, M.P. Kato, R. Song, and M. Iwata, "Summary of the ntcir-10 intent-2 task: Subtopic mining and search result diversification," Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp.761–764, ACM, 2013.

[8] Y. Liu, R. Song, M. Zhang, Z. Dou, T. Yamamoto, M.P. Kato, H. Ohshima, and K. Zhou, "Overview of the ntcir-11 imine task.," NTCIR, 2014.

[9] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp.1247–1250, ACM, 2008.

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

[11] 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, vol.16, no.4, pp.484–503, 2013.

[12] S. Hu, Z. Dou, X. Wang, T. Sakai, and J.R. Wen, "Search result diversification based on hierarchical intents," Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp.63–72, ACM, 2015.

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

[14] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp.335–336, ACM, 1998.

[15] T. Yamamoto, Y. Liu, M. Zhang, Z. Dou, K. Zhou, I. Markov, M.P. Kato, H. Ohshima, and S. Fujita, "Overview of the ntcir-12 imine-2 task," Proceedings of the 12th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access, 2016.

[16] C.L. Clarke, N. Craswell, and I. Soboroff, "Overview of the trec 2009 web track," tech. rep., DTIC Document, 2009.

[17] A. Broder, "A taxonomy of web search," ACM Sigir forum, vol.36, no.2, pp.3–10, 2002.

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

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

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

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

[22] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Mining query subtopics from questions in community question answering.," AAAI, pp.339–345, 2015.

[23] J. Hu, G. Wang, F. Lochovsky, J.t. Sun, and Z. Chen, "Understanding

user's query intent with wikipedia," Proceedings of the 18th international conference on World wide web, pp.471–480, ACM, 2009.
[24] F. Radlinski, M. Szummer, and N. Craswell, "Inferring query intent from reformulations and clicks," Proceedings of the 19th international conference on World wide web, pp.1171–1172, ACM, 2010.
[25] J.G. Moreno, G. Dias, and G. Cleuziou, "Query log driven web search results clustering," Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp.777–786, ACM, 2014.
[26] A. Damien, M. Zhang, Y. Liu, and S. Ma, "Improve web search diversification with intent subtopic mining," in Natural Language Processing and Chinese Computing, pp.322–333, Springer, 2013.
[27] S.J. Kim and J.H. Lee, "Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents," Inf. Process. Manage., vol.51, no.6, pp.773–785, Nov. 2015.
[28] S.J. Kim, J. Shin, and J.H. Lee, "Subtopic mining based on three-level hierarchical search intentions," European Conference on Information Retrieval, pp.741–747, Springer, 2016.
[29] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Advances in neural information processing systems, pp.3111–3119, 2013.
[30] Y. Xue, F. Chen, A. Damien, C. Luo, X. Li, S. Huo, M. Zhang, Y. Liu, and S. Ma, "Thuir at ntcir-10 intent-2 task.," NTCIR, NTCIR, 2013.
[31] J.G. Moreno and G. Dias, "Hultech at the ntcir-10 intent-2 task: Discovering user intents through search results clustering," Proceedings of NTCIR, NTCIR, 2013.
[32] J. Wang, G. Tang, Y. Xia, Q. Zhou, F. Zheng, Q. Hu, S. Na, and Y. Huang, "Understanding the query: Thcib and thuis at ntcir-10 intent task," Proceedings of NTCIR, NTCIR, 2013.
[33] S.J. Kim and J.H. Lee, "The kle's subtopic mining system for the ntcir-10 intent-2 task," Proceedings of NTCIR, NTCIR, 2013.
[34] M.Z. Ullah, M. Aono, and M.H. Seddiqui, "Sem12 at the ntcir-10 intent-2 english subtopic mining subtask," Proceedings of NTCIR, NTCIR, 2013.
[35] R. Krovetz, "Viewing morphology as an inference process," Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp.191–202, ACM, 1993.
[36] T.Y. Liu, "Learning to rank for information retrieval," Foundations and Trends in Information Retrieval, vol.3, no.3, pp.225–331, 2009.
[37] G. Amati, Probability models for information retrieval based on divergence from randomness, Ph.D. thesis, University of Glasgow, 2003.
[38] S. Robertson and H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Now Publishers Inc, 2009.
[39] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp.111–119, ACM, 2001.
[40] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp.334–342, ACM, 2001.
[41] D. Metzler and W.B. Croft, "A markov random field model for term dependencies," Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp.472–479, ACM, 2005.
[42] D. Metzler and T. Kanungo, "Machine learned sentence selection strategies for query-biased summarization," SIGIR Learning to Rank Workshop, pp.40–47, 2008.
[43] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol.24, no.5, pp.513–523, 1988.
[44] M. Bendersky, W.B. Croft, and Y. Diao, "Quality-biased ranking of web documents," Proceedings of the fourth ACM international conference on Web search and data mining, pp.95–104, ACM, 2011.
[45] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol.67, no.2, pp.301–320, 2005.
[46] K. YANAI *et al.*, "Visualtextualrank: An extension of visualrank to large-scale video shot extraction exploiting tag co-occurrence," IEICE TRANSACTIONS on Information and Systems, vol.98, no.1, pp.166–172, 2015.
[47] L. Cao, J. Guo, and X. Cheng, "Bipartite graph based entity ranking for related entity finding," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on, pp.130–137, IEEE, 2011.
[48] M.Z. Ullah, M. Aono, and M.H. Seddiqui, "Estimating a ranked list of human genetic diseases by associating phenotype-gene with gene-disease bipartite graphs," ACM Trans. Intell. Syst. Technol., vol.6, no.4, pp.56:1–56:21, July 2015.
[49] H. Deng, M.R. Lyu, and I. King, "A generalized co-hits algorithm and its application to bipartite graphs," Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09, New York, NY, USA, pp.239–248, ACM, 2009.
[50] J. Callan, M. Hoy, C. Yoo, and L. Zhao, "Clueweb09 data set," 2009.
[51] T. Strohman, D. Metzler, H. Turtle, and W.B. Croft, "Indri: A language model-based search engine for complex queries," Proceedings of the International Conference on Intelligent Analysis, pp.2–6, Citeseer, 2005.
[52] T. Sakai, "Ntcireval: A generic toolkit for information access evaluation," Proceedings of the Forum on Information Technology, pp.23–30, 2011.
[53] T. Sakai, "Statistical reform in information retrieval?," ACM SIGIR Forum, vol.48, no.1, pp.3–12, 2014.
[54] R.L. Santos, C. Macdonald, and I. Ounis, "Search result diversification," Foundations and Trends in Information Retrieval, vol.9, no.1, pp.1–90, 2015.

Md Zia Ullah received his B.Sc. (Hons.) degree from the University of Chittagong, Chittagong, Bangladesh in 2010, and his M.Eng. degree from the Toyohashi University of Technology (TUT), Toyohashi, Aichi, Japan in 2013. He is currently a PhD candidate at TUT. His research interests include information retrieval, intent mining, diversification, object recognition, and bioinformatics.

Masaki Aono received the BS and MS degrees from the Department of Information Science from the University of Tokyo, Tokyo, Japan, the PhD degree from the Department of Computer Science at Rensselaer Polytechnic Institute, New York. He was with the IBM Tokyo Research Laboratory from 1984 to 2003. He is currently a professor at the Graduate School of Computer Science and Engineering Department, Toyohashi University of Technology. His research interests include text and data mining for massive streaming data, and information retrieval for multimedia including 2D images, videos, and 3D shape models. He is a member of the ACM and IEEE Computer Society. He has been a Japanese delegate of the ISO/IEC JTC1 SC24 Standard Committee since 1996.