Query Subtopic Mining Exploiting Word Embedding for Search Result Diversification

Md Zia Ullah*, Md Shajalal, Abu Nowshed Chy, and Masaki Aono

Department of Computer Science and Engineering,
Toyohashi University of Technology, Toyohashi, Aichi, Japan
{arif*,shajalal,nowshed}@kde.cs.tut.ac.jp, aono@tut.jp

Abstract. Understanding the users' search intents through mining query subtopic is a challenging task and a prerequisite step for search diversification. This paper proposes mining query subtopic by exploiting the word embedding and short-text similarity measure. We extract candidate subtopic from multiple sources and introduce a new way of ranking based on a new novelty estimation that faithfully represents the possible search intents of the query. To estimate the subtopic relevance, we introduce new semantic features based on word embedding and bipartite graph based ranking. To estimate the novelty of a subtopic, we propose a method by combining the contextual and categorical similarities. Experimental results on NTCIR subtopic mining datasets turn out that our proposed approach outperforms the baselines, known previous methods, and the official participants of the subtopic mining tasks.

Keywords: Subtopic Mining, Word Embedding, Diversification, Novelty

1 Introduction

According to user search behavior analysis, query is usually unclear, ambiguous, or board [10]. Issuing the same query, different users may have different search intents, which correspond to different subtopic [8]. For example, 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 subtopic, 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 [11]. Search engine often fails to capture the diversified search intents of a user if the issued query is ambiguous or broad and results in a list of redundant documents. As these documents may cover a few subtopic or interpretations, the user is usually unsatisfied.

In this paper, we address the problem of *query subtopic mining*, which is defined as: "given a query, list up its possible subtopic which specialises or disambiguates the search intent of the original query." In this regard, our contributions are threefold: (1) some new features based on word embedding, (2) a

bipartite graph based ranking for estimating the relevance of the subtopic, and (3) estimating the novelty of the subtopic by combining a mutual information based similarity and categorical similarity.

The rest of the paper is organized as follows. **Section** 2 introduces our proposed subtopic mining approach. **Section** 3 discusses the overall experiments and results that we obtained. Finally, concluding remarks and some future directions of our work are described in **Section** 4.

2 Mining Query Subtopic

In this section, we describe our approach to query subtopic mining, which is composed of subtopic extraction, features extraction, and ranking. Given a query, first we extract candidate subtopics from multiple resources. Second, we extract multiple semantic and content-aware features to estimate the relevance of the candidate subtopics, followed by a supervised feature selection and a bipartite graph based ranking. Third, to cover the possible search intents, we introduce a novelty measure to diversify the subtopic.

2.1 Subtopic Extraction

Inspired by the work of Santos et al. [9], our hypothesis is that *suggested queries* in across search engines hold some intents of the query. For a query, we utilize the suggested queries, provided by the search engines. If a query is matched with the title of a Wikipedia disambiguation page, we extract the different meanings from that page. Then, we aggregate the subtopic by filtering out the candidates, which is the part of the query or exactly similar.

2.2 Features Extraction

Let $q \in \mathcal{Q}$ represents a query and $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ represents a set of candidate subtopics extracted in Section 2.1. We extract multiple local and global features, which are broadly organized as word embedding and content-aware features. We propose two semantic features based on locally trained word embedding and make use of *word2vec*[1] model [6].

In order to capture of the semantic matching of a query with a subtopic, we first propose a new feature, the maximum word similarity (MWS) as follows:

$$\begin{aligned} f_{MWS}(q, s) &= \frac{1}{|q|} \sum_{t \in q} \ sem(t, s) \\ sem(t, s) &= \max_{w \in s} \ f_{sem}(\boldsymbol{t}, \boldsymbol{w}) \end{aligned} \tag{1}$$

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

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

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

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

Among content-aware features, we extract features based on term frequency, including *DPH* [9], *PL2* [9], and *BM25* [9]; language modeling, including Kullback-Leibler (*KL*) [9], Query Likelihood with Jelinek-Mercer (*QLM-JM*) [9], Query Likelihood with Dirichlet smoothing (*QLM-DS*) [9], and Term-dependency Markov random field (*MRF*) [9]; lexical, including edit distance, sub-string match (*SSM*) [5], term overlap [5], and term synonym overlap (*TSO*) [5]; web hit-count, including normalized hit count (*NHC*), point-wise mutual information (*PMI*), and word co-occurrence (*WC*); and some query independent features, including average term length (*ATL*), topic cohesiveness (*TC*) [9], and subtopic length (*SL*).

2.3 Subtopic Ranking

To remove noisy and redundant features, we normalize the features using *Min-Max* and employ elastic-net regularized regression.

Bipartite Graph based Ranking Many real applications can be modeled as a bipartite graph, such as Entities and Co-List [1] in a Web page. We hypothesize that a relevant subtopic should be ranked at the higher position by multiple effective features and intuitively, an effective feature should be ranked at higher position by multiple relevant subtopics. On this intuition, we represent a set of features $\mathcal{F} = \{f_1, f_2, \cdots, f_m\}$ and a set of candidate subtopics $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$ as a bipartite graph, $\mathcal{G} = (\mathcal{F} \cup \mathcal{S}, \mathcal{E})$, and introduce weight propagations from both sides. The weight $w_{i,j} = 1/\sqrt{\log_2(rank(L_{f_i}, s_j) + 2.0)}$ of an edge between a feature f_i and a subtopic s_j, where rank(L_{f_i}, s_j) returns the position of the subtopic s_j in the ranked list L_{f_i} for the feature f_i.

Let $\mathcal{M}$ be a bi-adjacency matrix of $\mathcal{G}$, $\mathcal{W}_1 = \mathcal{D}_{\mathcal{F}}^{-1}\mathcal{M}$, and $\mathcal{W}_2 = \mathcal{D}_{\mathcal{S}}^{-1}\mathcal{M}^T$, where $\mathcal{D}_{\mathcal{F}}$ and $\mathcal{D}_{\mathcal{S}}$ are the row-diagonal and the column-diagonal matrices of $\mathcal{M}$.

The weight propagations from the set $\mathcal{S}$ to the set $\mathcal{F}$ and vice versa are represented as follows:

$$\begin{aligned} F_{k+1} &= \lambda_1 \mathcal{W}_1 S_k + (1-\lambda_1)\ F_0 \\ S_{k+1} &= \lambda_2 \mathcal{W}_2 F_{k+1} + (1-\lambda_2)\ S_0 \end{aligned} \tag{3}$$

where $0 < \lambda_1, \lambda_2 < 1$, F_0 and S_0 are the initial weight vectors, F_k and S_k denotes the weight vectors after the k-th iterations.

From the iterative solution of the Equations (3), we have

$$S^* = (I - \lambda_1\lambda_2\mathcal{W}_2\mathcal{W}_1)^{-1}\ [(1-\lambda_1)\lambda_2\mathcal{W}_2 F_0 + (1-\lambda_2)\ S_0] \tag{4}$$

Given λ_1, λ_2, $\mathcal{W}_1$, $\mathcal{W}_2$, F_0, and S_0, we estimate the scores S^* directly by applying Eq. (4). These scores S^* are considered as the relevance scores, rel(q, $\mathcal{S}$), which is utilized in diversification.

Subtopic Diversification To select the maximum relevant and the minimum redundant subtopic, we diversify the subtopic using the MMR [2] framework, which can be defined as follows:

$$s_i^* = \underset{s_i \in R \setminus C_i}{\operatorname{argmax}} \ \gamma \ rel(q, s_i) + (1 - \gamma) \ novelty(s_i, C_i) \tag{5}$$

where $\gamma \in [0, 1]$, $rel(q, s_i)$ is the relevance score, and $novelty(s_i, C_i)$ is the novelty score of the subtopic s_i. R is the ranked list of subtopic retrieved by Equation (4). C_i is the collection of subtopic that have already been selected at the i-th iteration and initially empty.

Since subtopics are short in length and they might not be lexically similar. We hypothesize that if two subtopics represent the similar meaning, they may belong to the similar categories and retrieve similar kinds of documents from a search engine. Therefore, we propose to estimate the novelty of a subtopic by combining the contextual and categorical similarities as follows:

$$novelty\ (s_i, C_i) = - \max_{s' \in C_i} \left(1.0 - \sqrt{JSD(s_i, s')} + \sum_{x \in X} \frac{[(s_i, s') \in x]}{|x|} \right) \tag{6}$$

where $JSD(s_i, s')$ is estimated through the Jensen-Shannon divergence of the word probability distributions of the top-k documents refer to the subtopics s_i and s'. X is the set of clusters obtained by applying the frequent phrase based soft clustering on the candidate subtopics, $|x|$ is the number of subtopics belong to the cluster x, and $[(s_i, s') \in x] = 1$ if true, zero, otherwise.

3 Experiments and Evaluations

In this section, we evaluate our proposed method (*W2V-BGR-Nov*) and compare the performance with previous methods, including [3, 4, 7], the diversification methods MMR[2], XQuAD [9], and the baseline, MergeBGY, merging of query completions from Bing, Google, and Yahoo. For relevance estimation, linear ranking is used in the MergeBGY, whereas Equation (4) is used for MMR and XQuAD. Moreover, the cluster label of the frequent phrase based soft clustering of candidates is considered as the sub-topics for XQuAD. For estimating novelty, cosine similarity is utilized for MergeBGY, MMR, and XQuAD. We estimate evaluation metrics, including I-rec@10, D-nDCG@10, and D#-nDCG@10; and use the two-tailed paired t-test for statistical significance testing ($p<0.05$).

3.1 Dataset

The INTENT-2 and IMINE-2 [12] test collections include 50 and 100 topics, respectively. As resources, query completions from Bing, Google, and Yahoo were collected and included in the datasets. To estimate the features, including Equations (1) and (2), we retrieved the top-1000 documents from the clueweb12-b13 corpus based on language model for each topic and locally trained *word2vec*. The parameters in the *word2vec* tool are Skip-gram architecture, window width of 10, dimensionality of 200, and the sampling threshold of 10^{-3}.

3.2 Important Features and Parameter Tuning

We trained elastic-net on INTENT-2 dataset, however, we employed on IMINE-2 dataset, and vice versa. We extracted in total 27 features and the selected features were as follows: *MWS*, *MVS*, *DPH*, *QLM-JM*, *MRF*, *SSM*, *TSO*, *NHC*, *WC*, and *ATL*. It turned out that our proposed features *MWS* and *MVS* are important and were chosen during feature selection. Through empirical evaluation, we found the optimal insensitive range of values of λ_1 and λ_2 in Equation (4) as [0.6 – 0.8] and [0.4 – 0.6], respectively. We found the optimal value of γ in Equation (5) as 0.85, which reflects that *MMR* rewards relevance than diversity in mining subtopic.

Table 1. Comparative performance of our proposed W2V-BGR-NOV with previous methods for INTENT-2 topics. The best result is in bold. † indicates statistically significant difference and ⋄ indicates statistically indistinguishable from the best

	Method	I-rec@10	D-nDCG@10	D#-nDCG@10
Our Proposed	***W2V-BGR-NOV***	**0.4774**†	**0.5401**†	**0.5069**†
Baseline	MergeBGY	0.3365	0.3181	0.3273
Previous Methods	Se-jong et al.[4]	0.4457	0.4401	0.4429
	Moreno et al.[7]	0.4249	0.4221	0.4225
	Damien et al.[3]	0.4587$^{\diamond}$	0.3625	0.4106
Diversification methods	XQuAD[9]	0.3637	0.5055	0.4346
	MMR[2]	0.3945	0.4079	0.4048

Table 2. Comparative performance of our proposed W2V-BGR-NOV with the known previous methods for IMINE-2 topics. Notation conventions are the same as in Table 1.

	Method	I-rec@10	D-nDCG@10	D#-nDCG@10
Our Proposed	***W2V-BGR-NOV***	**0.8349**†	**0.6836**†	**0.7602**†
IMINE-2 Participants	KDEIM-Q-E-1S	0.7557	0.6644	0.7101
	HULTECH-Q-E-1Q	0.7280	0.6787$^{\diamond}$	0.7033
	RUCIR-Q-E-4Q	0.7601	0.5097	0.6349
Diversification methods	XQuAD[9]	0.6422	0.6571	0.6510
	MMR[2]	0.7572	0.6112	0.6908

3.3 Experimental Results

The comparative performances are reported in Tables 1 and 2 for INTENT-2 and IMINE-2 topics. The results show that overall *W2V-BGR-NOV* is the best. In terms of diversity (i.e. I-rec@10), *W2V-BGR-NOV* significantly outperforms all baselines except [3] for INTENT-2 topics. Though previous methods utilize multiple resources which often cause noisy subtopics, however, our proposed estimation of subtopic novelty in Equation (6) eliminates redundant subtopics and benefits more diverse subtopics. In terms of relevance (i.e. D-nDCG@10), *W2V-BGR-NOV* outperforms all baselines except HULTECH-Q-E-1Q for IMINE-2 topics. Our proposed word embedding based features, followed by bipartite graph based ranking capture better semantics to estimate the relevance of the

subtopics. In terms of D#-nDCG@10, which is a combination of I-rec@10 (0.5) and D-nDCG@10 (0.5), *W2V-BGR-NOV* significantly outperforms all the baselines. The overall result demonstrates that our proposed *W2V-BGR-NOV* is effective in query subtopic diversification.

4 Conclusion

In this paper, we proposed mining and ranking query subtopic by exploiting word embedding and short-text similarity measure. We introduced new features based on word embedding and bipartite graph based ranking to estimate the relevance of the subtopic. To diversify the subtopic covering multiple intents, we proposed to estimate the novelty of a subtopic by combining the contextual and categorical similarities. Experimental results demonstrate that our proposed approach outperforms the baseline and known previous methods. In the future, we will evaluate the effectiveness of the mined subtopics by employing search diversification.

Acknowledgement This research was supported by JSPS Grant-in-Aid for Scientific Research (B) 26280038.

References

1. Cao, L., Guo, J., Cheng, X.: Bipartite graph based entity ranking for related entity finding. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on. pp. 130–137. IEEE (2011)
2. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. pp. 335–336. ACM (1998)
3. Damien, A., Zhang, M., Liu, Y., Ma, S.: Improve web search diversification with intent subtopic mining. In: Natural Language Processing and Chinese Computing, pp. 322–333. Springer (2013)
4. Kim, S.J., Lee, J.H.: Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents. Inf. Process. Manage. 51(6), 773–785 (Nov 2015)
5. Metzler, D., Kanungo, T.: Machine learned sentence selection strategies for query-biased summarization. In: SIGIR Learning to Rank Workshop. pp. 40–47 (2008)
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)
7. Moreno, J.G., Dias, G., Cleuziou, G.: Query log driven web search results clustering. In: SIGIR. pp. 777–786. ACM (2014)
8. Ren, P., Chen, Z., Ma, J., Wang, S., Zhang, Z., Ren, Z.: Mining and ranking users intents behind queries. Information Retrieval Journal 18(6), 504–529 (2015)
9. Santos, R.L.T.: Explicit web search result diversification. Ph.D. thesis, University of Glasgow (2013)
10. Song, R., Luo, Z., Nie, J.Y., Yu, Y., Hon, H.W.: Identification of ambiguous queries in web search. Information Processing & Management 45(2), 216–229 (2009)
11. Wang, C.J., Lin, Y.W., Tsai, M.F., Chen, H.H.: Mining subtopics from different aspects for diversifying search results. Information retrieval 16(4), 452–483 (2013)
12. Yamamoto, T., Liu, Y., Zhang, M., Dou, Z., Zhou, K., Markov, I., Kato, M.P., Ohshima, H., Fujita, S.: Overview of the ntcir-12 imine-2 task. In: NTCIR (2016)