KDEIM at NTCIR-12 IMine-2 Search Intent Mining Task: Query Understanding Through Diversified Ranking of Subtopics

Md Zia Ullah
arif@kde.cs.tut.ac.jp

Md Shajalal
shajalal@kde.cs.tut.ac.jp

Masaki Aono
aono@tut.jp

Department of Computer Science and Engineering
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, 441-8580, Aichi, Japan

ABSTRACT

In this paper, we describe our participation in the Query Understanding subtask of the NTCIR-12 IMINE Task. We propose a method that extracts subtopics by leveraging the query suggestions from search engines. The importance of the subtopics with the query is estimated by exploiting multiple query-dependent and query-independent features with supervised feature selection. To diversify the subtopics, we employ maximum marginal relevance (*MMR*) framework based diversification technique by balancing the relevance and novelty. The best performance of our method achieves an *I-rec* of 0.7557, a *D-nDCG* of 0.6644, a *D#-nDCG* of 0.7100, and a *QU-score* of 0.5057 at the cutoff rank 10 for query understanding task.

Team Name

KDEIM

Subtasks

Query Understanding 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; the search engine responds with a ranked list of snippet results to meet the information needs of an user. Web search queries are typically short, ambiguous, and contain multiple subtopics [3,16,17]. Search engine often fails to capture users' intents if an issued query is ambiguous. The reason is that an ambiguous query has more than one interpretation and different users may have different intents underlying the same query.

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 [11]. Ignoring the underlying subtopics of a query, search engine results in top ranked documents possibly containing too much relevant information on a particular subtopic of a query that may leave the user unsatisfied. In order to satisfy the user, a sensible approach is to diversify the documents retrieved for an ambiguous or board query [4]. 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 intents behind a query and diversifying the results with respect to the possible intents.

The NTCIR-12 IMine-2 Search Intent Mining [19] have dealed with the above problem via two subtasks. The first subtask is how to mine and rank the underlying subtopics and the second subtask is how to selectively diversify search results. We participated in the former subtask. In this regard, we propose a method to extract and rank possible subtopics underlying a query. Given a query, we consider the query suggestions of the search engines as the candidate subtopics and rank the subtopics by balancing the relevance with the query and novelty with other candidate subtopics.

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

2. RELATED WORK

Search queries are usually short, ambiguous, and tend to have several intents [3, 16, 17]. Several methods have been proposed to mine the possible subtopics behind a query. Wang et al. [18] proposed a method to mine subtopics of query either indirectly from the search results or directly from external resources, such as Wikipedia, open directory project (*ODP*), query logs, and the related search services provided by search engines. Hu et al. [6] identified the intents of query utilizing the knowledge contained in the Wikipedia. To mine candidate subtopics, Jiyin et al. [5] proposed a random-walk based approach to estimate the similarities of the explicit subtopics mined from a number of heterogeneous resources: click logs, anchor text, and web n-gram. Filip Radlinkshi et al. [12] proposed an approach for identifying query intents from query reformulations and click-through data. The click and reformulation information are combined to identify a set of possible related queries to construct an undirected graph for a query. An edge is introduced between two queries if they were often clicked for

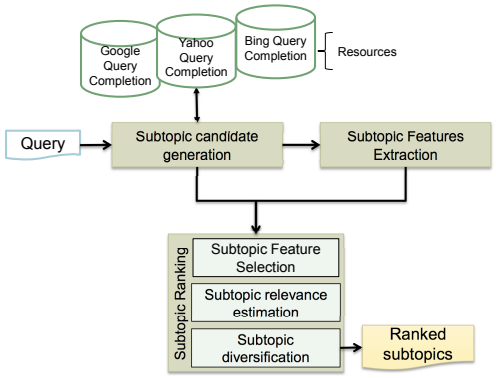

Figure 1: Diversified subtopic mining flow

the same documents. Then, random walk similarity is used to find intent cluster. Recently, Kim et al. [7] proposed a method to mine subtopics using simple patterns and hierarchical structure of subtopic candidates. They extracted relevant phrase as subtopic candidates using simple patterns and constructed a hierarchical structure using sets of relevant documents from web document collection. Zheng et al. [20] proposed a maximal frequent pattern mining algorithm to extract the pattern from retrieval results of a query as candidate subtopics.

3. OUR APPROACH

In this section, we describe our proposed method of the query understanding, which is depicted in Fig. 1. Query suggestions provided by web search engines (*WSE*) are an easy and effective choice for obtaining candidate subtopics. Inspired by Santos et al. [14], we collect the query suggestions from multiple search engines. The suggested queries are aggregated by filtering out the duplicates or wrongly represented ones, and consider them as candidate subtopics. To rank the candidate subtopics, we estimate the relevance scores of the candidate subtopics by multiple extracting query-dependent and query-independent features with supervised feature selection. To diversify subtopics by considering maximum relevance with minimum redundancy, we employ maximal marginal relevance (*MMR*) [2] based diversification model.

3.1 Subtopic Feature Extraction

To estimate the relevance of the candidate subtopics with the query, we extract multiple features which are organized as query-dependent and query-independent, according to whether they are computed on-the-fly at querying time or offline at indexing time, respectively.

Query-dependent features are directly computed by scoring the occurrences of the terms of a query in a subtopic. We extract some term frequency (*TF*), language modeling (*LM*), term dependency (*TD*), lexical, and web hit count (*HC*) based features. Term frequency based features include DPH [1], PL2 [1], and BM25 [13]. Language modeling features include Kullback-Leibler (*KL*) [15], query likelihood with Jelinek-Mercer (QLM-JM) [15], subtopic likelihood with Jelinek-Mercer (SLM-JM) [15], query likelihood with Dirichlet smoothing (QLM.DS) [15], and subtopic likelihood with Dirichlet smoothing (SLM.DS) [15]. Term dependency features include term dependency Markov random field (MRF) [8] and Tri-gram dependency. To measure the lexical similarity between the query and the candidate, we extract edit distance based feature, sub-string match, term overlap [9], term synonym overlap, vector space model (VSM), and coordinate level matching (CLM) [15] features. 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 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) [15], and subtopic length (SL).

3.2 Subtopic Ranking

For a pair of query q and subtopic s, we extract all the features described in Section 3.1 and represent those in a feature vector, $\mathcal{F}_{q,s} = \{f_{DPH}(q,s), f_{PL2}(q,s), \cdots, f_{SL}(q,s)\}$. 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 rank of a subtopic, first, we employ a supervised feature selection to remove noisy and redundant features. Second, we apply a linear ranking function for estimating the relevance of subtopic. Finally, we employ maximum marginal relevance (*MMR*) [2] framework to assemble a ranked list of subtopics by balancing the relevance of subtopic with query and novelty with other subtopics.

3.2.1 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. Given a parameter α strictly between 0 and 1, and a nonnegative λ, elastic-net solves the following optimization problem:

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

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 which give the positive coefficient β.

3.2.2 Subtopic Relevance Estimation

To estimate the relevance of a subtopic s with a query q, we employ a linear ranking model by incorporating all the features as follows:

$$rel(q,s) = \sum_{i=1}^{N} w_i \cdot f_i(q,s) \quad (2)$$

where w_i is the learned weight corresponding to feature f_i that we get from Equation 1 after feature selection. N is the total number of features and $f_i(q,s)$ is the i^{th} feature of

subtopic s for query q. The relevance score $rel(s)$ is used in the diversification framework.

3.2.3 Subtopic Diversification

To diversify the mined subtopic candidates, we utilize MMR framework. Given a relevance function, $rel(.)$ and a novelty function, $novelty(.,.)$, the MMR model could be set up as follows:

$$s_i^* = \underset{s_i \in R \setminus C_i}{\text{argmax}} \ \gamma \ rel\ (q, s_i) + (1-\gamma)\ novelty\ (s_i, C_i) \quad (3)$$

where γ is a combining parameter and $\gamma \in [0,1]$. R is the ranked list of subtopics retrieved in Equation 2. C_i is the collection of subtopics have already been selected at the ith 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 C_i has already been selected and ranked.

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

$$novelty\ (s_i, C_i) = - \max_{s' \in C_i} \ sim\ (s_i, s') \quad (4)$$

The similarity between two subtopics s and s' is estimated as follows:

$$sim\ (s, s') = cosine(s, s') \quad (5)$$

where $cosine(s, s')$ is estimated through the vector representation of subtopics s and s'.

3.3 Vertical Selection

Many commercial Web search engines merge several types of search results and generate a search engine results page in response to a user's query. For example, the results of query "flower" now may contain image results and encyclopedia results as well as usual Web search results. We refer to such "types" of search results as verticals. For example, image, movie, audio, finance, news can be a vertical. Six verticals are selected by the Imine-2 organizers including *Web*, *Image*, *News*, *QA*, *Encyclopedia*, and *Shopping*.

For selecting vertical, we propose an adhoc classification. We select a set of terms corresponding to each of the six verticals. For example, we select a set of representative terms including *Image*, *Photo*, *Album*, *Gallery*, and *Artwork* for *Image* vertical. Each of the representative terms for a vertical are mapped to a 300-dimensional word vector by utilzing *word2vec* [10] pre-trained model which was trained on *Google News Corpus*[1]. A vertical is represented by a 300-dimension vector by summing up the vectors of the representative terms corresponding to that vertical. Therefore, we have a 300-dimensional resourceful vector corresponding to each of the vertical.

To estimate the vertical of a subtopic, first, each of the terms in the subtopic is also mapped to 300-dimension vector by utilizing *word2vec*. Summing up the vectors of all the terms in a subtopic, we may have a 300-dimensional vector for that subtopic. Then, we compute cosine similarity between a subtopic vector and a vertical vector. If the similarity is higher than 0.75, we annotate the subtopic with that vertical.

[1]https://code.google.com/archive/p/word2vec

4. EXPERIMENTS

We submitted four runs to the English Query Understanding subtask [19]. The run configurations are stated in the Table 1. Among the submitted runs, there are two types of runs, S-run and Q-run. S-Run: identifying subtopics, but not verticals underlying a query. Q-Run: identifying both subtopics and relevant verticals given a query. We submitted two S-Runs and two Q-Runs. To make S-Run, for instance, a run "KDEIM-Q-E-4S", which was produced by extracting the subtopic candidates from the query suggestions and ranked by estimating the multiple query-dependent and query-independent features as described in the Section 3.1, followed by diversifying the subtopics using *MMR*. To make Q-run, we do the same pipeline as S-Run and make use of *word2vec* [10] as described in Section 3.3 for vertical selection.

Table 1: Run configuration of query understanding subtask

Run	**Resources**	**Methods**
KDEIM-Q-E-1S	Query Suggestion	Multiple features, Elastic-net, Diversification
KDEIM-Q-E-2Q	Ditto	Multiple features, Elastic-Net, Diversification, Word2Vec
KDEIM-Q-E-3Q	Ditto	Multiple features, Elastic-Net, Word2Vec
KDEIM-Q-E-4S	Ditto	Multiple features

4.1 Evaluation Metric

Subtopic mining runs are evaluated by estimating the I-rec, D-nDCG, and D#-nDCG metrics. I-rec measures the diversity of the returned subtopics, which shows how many percentages of intents can be found. D-nDCG measures the overall relevance across all intents considering the subtopic ranking. D#-nDCG is a combination of I-rec and D-nDCG. The detail description of these metrics are stated here [19].

4.2 Experimental Results

The official evaluation results of our submitted runs are stated in the Table 2 and 3.

Table 2: Performance of subtopic mining subtask

Run	I-rec@10	D-nDCG@10	D#-nDCG@10
KDEIM-Q-E-1S	0.7556	0.6644	0.7100
KDEIM-Q-E-2Q	0.7556	0.6644	0.7100
KDEIM-Q-E-3Q	0.7458	0.6472	0.6965
KDEIM-Q-E-4S	0.7484	0.5645	0.6565

Table 3: Performance of vertical Identification subtask

Run	V-score	QU-score
KDEIM-Q-E-2Q	0.3014	0.5057
KDEIM-Q-E-3Q	0.3014	0.4948

5. CONCLUSION

This paper described our participation in the Query Understanding subtask in the NTCIR-12 IMINE-2 Task. We proposed a method for diversified subtopic mining by exploiting multiple resources and content-aware features. A set of experiments is carried out to verify the effectiveness of our proposed method. Experimental results reveal that our subtopic mining method achieved a good scores in several evaluation metrics. In future, we would like to emphasize on the vertical identification of subtopic.

6. REFERENCES

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[2] 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.

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

[4] 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.

[5] J. He, V. Hollink, and A. de Vries. Combining implicit and explicit topic representations for result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 851–860. ACM, 2012.

[6] 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.

[7] S.-J. Kim and J.-H. Lee. Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents. *Information Processing & Management*, 51(6):773–785, 2015.

[8] 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.

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

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[11] 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.

[12] 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.

[13] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[14] 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.

[15] R. L. T. Santos. *Explicit web search result diversification*. PhD thesis, University of Glasgow, 2013.

[16] 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.

[17] K. Spärck-Jones, S. E. Robertson, and M. Sanderson. Ambiguous requests: implications for retrieval tests, systems and theories. *ACM SIGIR Forum*, 41(2):8–17, 2007.

[18] 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.

[19] T. Yamamoto, Y. Liu, M. Zhang, D. Zhicheng, Z. Ke, M. IIya, M. P. kato, O. Hiroaki, and F. Sumio. Overview of the ntcir-12 imine-2 task. *Proceedings of NTCIR-12*, 2016.

[20] W. Zheng, H. Fang, H. Cheng, and X. Wang. Diversifying search results through pattern-based subtopic modeling. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 8(4):37–56, 2012.