IMC at the NTCIR-12 IMine-2 Query Understanding Subtask

Jiahui Gu
Department of Computer Science and Technology
Beijing Institute of Technology
Beijing 100081, China
gujh@bit.edu.cn

Chong Feng
Department of Computer Science and Technology
Beijing Institute of Technology
Beijing 100081, China
fengchong@bit.edu.cn

Yashen Wang
Department of Computer Scienceand Technology
Beijing Institute of Technology
Beijing 100081, China
yswang@bit.edu.cn

ABSTRACT
This paper describes the participation of IMC team in the Chinese Query Understanding Subtask in the NTCIR-12 IMine-2 Task. To identify the subtopics of a given query, we utilize several data resource and innovatively employ new words extraction theory to obtain the expansion terms for a query, which is the kernel of the proposed system. Then we generate the query subtopic based on the expansion terms obtained above. Moreover, we also attempt to leverage topic model in another way of subtopic terms generation, and use K-means algorithm for diversity clustering of query subtopics.

Team Name
IMC

Subtasks
Query Understanding Subtask (Chinese)

Keywords
new words extraction, subtopics, topic model

1. INTRODUCTION
Web search engine provides an important mechanism for users to meet their information needs, which are usually formulated by queries[1]. Users' queries to Web tend to have more than one interpretation or refer to multiple aspects due to their ambiguity and other characteristics[2].

Different representing ways of identifying the search intends of a query have been proposed which are based on subtopics, query dimensions and query aspects. For example, Clarke et al.[4] represented query intents by subtopics which denote different senses or multiple facets of queries. Dou Z et al.[6] referred to query dimensions as addressing the problem of finding multiple groups of words or phrases that explain the underlying query facets. Wang et al.[7] studied how to extract broad aspects from query reformulations, each broad aspect is represented by a set of keywords.

Many methods on intent mining task are used such as graph model, clustering algorithm and latent concept expansion. Zhang Z et al.[3] converted query subtopic mining into measure the similarity among subtopics and denoted the dependencies of subtopics as two kinds of graph that is all-connection structure and strong-connection structure. Hu Y et al.[5] conducted clustering on the clicked URLs of each query and its expanded queries. Bouchoucha A et al.[8] used embedding framework to select latent expansion terms for an original query, such as each expansion term can be mapped into one or several possible aspect(s) of the query and also be accounted as a subtopic.

In this paper, multiple interpretation and aspects associated with a query are called ***subtopics***. Inspired by the essence of concept expansion[8], we propose a novel approach of QENEW that is Query Expansion based on New-Word Extraction Algorithm to extract new words and the proposed approach explores a first step toward generating subtopics from new words extraction algorithm. The QENEW algorithm could be briefly interpreted as considering these generated new words as the query expansion terms, and then taking the linear concatenation to turn the original query and these expansion terms into phrases. In the end, there is no doubt that the phrases are the actual desired subtopics. We summarize the process of our work as follows. Firstly, we preprocess the datasets and Chinese queries, which is the basis of QENEW algorithm producing candidate new words. Secondly, through applying statistical language knowledge and information entropy theory to measure the new words inner features and taking a ranking algorithm, a list of effective new words or query expansion terms could be obtained. What's more, with the desire of subtopic diversity clustering, we design additional tests to study the performance by means of K-means algorithm. For subtopic generation, topic model is applied to the new words extraction algorithm. Furthermore, according to learning the papers relevant to intent mining, we could conclude that combining multiple resources is usually more effective than considering any resource in isolation. So we collect external resources from the representative search engines attempting to diversify the query expansion terms.

2. OUR SYSTEM
In this section, we will introduce our system in figure 1 comprising of Query Expansion based on New-Word Extraction Algorithm (QENEW), topic model and K-means algorithm. Firstly describe our QENEW algorithm to generate new words and then obtain the query subtopics. Afterwards, based on the result of new words extracted, we utilize K-means algorithm[1] and topic model[2] to make further experiments expecting to get better result.

2.1 Query Expansion based on New-Word Extraction Algorithm (QENEW)

[1] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

[2] http://radimrehurek.com/gensim/index.html

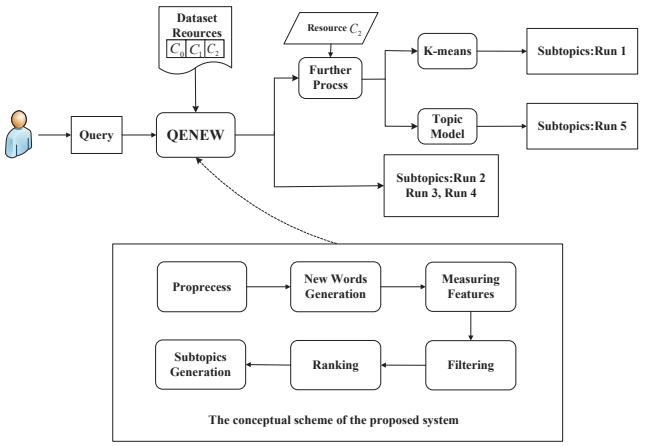

Figure 1: The flowchart of our system

Inspired by the method of latent concept extension which attends to get the extension words of a query, we apply the theory of new words extraction to obtain our query expansion words. The main procedure of QENEW algorithm is briefly introduced below.

2.1.1 Preprocessing and new words generation

As for preprocessing, firstly tokenize the Chinese Queries and corpus. Then separately for each query, we filter the same words appeared both in corpus and Chinese Queries, afterwards, take the rest part of the corpus as input of our QENEW algorithm.

Take query "*边界 (boundary)*" for example, the final preprocessed input of the model turns into "*中越问题 (the problem of China and Vietnam)*" corresponding to the original phrase "*中越边界问题 (the boundary problem between China and Vietnam)*".

For new words generation, we take the whole corpus as a long string and compute the whole length of the corpus's characters. Next specify the maximal length of a new word and generate the possible new words by the following method.

For string $S = c_1c_2\cdots c_n$, define l_{text} as the corpus's characters length and l_{seq} as the maximal length of a new word. The possible new words (P_NW) set are generated by:

$$P\_NW = c_ic_{i+1}\cdots c_{i+j}$$

Where $n \in [1, l_{text}], i \in [0, l_{text}), j \in [1, l_{seq}),\ i + j < l_{text} + l_{seq} - 2$.

2.1.2 Measuring features

After obtaining the new words set P_NW, we measure words' internal features by means of a series of measurement based on statistical language knowledge and information entropy like frequency, string cohesion and string liberalization, which is necessary for conforming whether it's an efficient word. The measurements are described as follows:

Frequency(*F*). It is intuitive that a string can be potential word if it occurs repeatedly and has high frequency. To measure the reliability of new words, it's necessary to compute the frequency of every possible new word in P_NW. In our experiments, the occurrence frequency of new word P_NW_k was calculated by the formula below:

$$F(P\_NW_k) = \frac{T(P\_NW_k)}{l_{text}} \qquad (2-1)$$

where the function T is the times of P_NW_k occurring in the corpus.

String Cohesion(*SC*). String cohesion indicates the internal feature of new words which corresponds to the correlation of different components. In order to calculate the degree of condensation of a P_NW, we need to enumerate its cohesive methods, that is, which two parts the P_NW is composed of. For new word $P\_NW_k = c_1c_2\cdots c_n (n > 1)$, the string cohesion value SC is defined as:

$$SC = min\left\{\frac{F(P\_NW_k)}{F(c_1) * F(c_{2\ldots}c_n)}, \frac{F(P\_NW_k)}{F(c_n) * F(P\_NW_{n-1})}\right\} \qquad (2-2)$$

where $F(P\_NW_k)$ and $F(c)$ respectively denote the possible new word and character frequency of P_NW_k and *c*.

Take "*电影院 (movie theatre)*" for example,

$$SC(\text{电影院}) = min\left\{\frac{F(\text{电影院})}{F(\text{电}) * F(\text{影院})}, \frac{F(\text{电影院})}{F(\text{电影}) * F(\text{院})}\right\}$$

String Liberalization(*SL*). String liberalization indicates the string's diversity of neighborhood features and measures the uncertainty of a $P\_NW's$ left and right character by using the information entropy. It's no doubt that a string can be counted as an efficient word if it can flexibly appear in different environments and have abundant left and right character sets. For new word *S*, define its left character set as $C_l = \{c_1, c_2\cdots c_l\}$ and right character set as $C_r = \{c_1, c_2\cdots c_r\}$.We compute SL_l and SL_r, the left and right information entropy with respect to string *S*, as:

$$SL_l = -\sum_{i\in C_l}[F(c_i) \times log\, F(c_i)] \qquad (2-3)$$

$$SL_r = -\sum_{j\in C_r}[F(c_j) \times log\, F(c_j)] \qquad (2-4)$$

$$SL(S) = min\{SL_l, SL_r\} \qquad (2-5)$$

where $SL(S)$ denotes the information entropy of string *S*. The larger the SL value is, the greater degree of liberalization is and the more possible the string *S* will be a potential new word.

2.1.3 Filtering and ranking

Based on assumption that a new word represents an aspect of a query, it should be human readable and express a specific meaning. Considering the aforementioned three features, we further filter and rank the P_NW set to get a ranked list of the candidate new words set (*CNW*).

Filtering. After finding P_NW and computing the features values, there are many words such as "*桌面壁* ", "*线观看* " which respectively corresponding to "*桌面壁纸 (desktop wallpaper)*"and "*在线观看 (online watching)*". So it's necessary to remove these garbage words. Then we set some standard values for occurrence times *T*, *SL*, l_{seq} and *SC*, those satisfying the criteria could be taken as an efficient new word.

Ranking. Considering the measuring features in section 2.1.2, we linearly combine string frequency, string cohesion and string liberalization together for new words ranking. Thus, the possibility of a new word *w* can be formulated as follows:

$$p(w) = \alpha_1F(w) + \alpha_2\widehat{SC}(w) + \alpha_3\widehat{SL}(w) \qquad (2-6)$$

where $w \epsilon CNW$, $\sum_{i=1}^{3} \alpha_i = 1, \alpha_i$ are coefficients $\widehat{SL}(w)$ and $\widehat{SC}(w)$ are the standardization of string cohesion and string liberalization respectively.

In addition, to meet the quantity demands of query intents, we only select the top K new words as the final expansion terms, which choose the top K new words from the ranked *CNW* set as the final query expansion terms.

2.1.4 Subtopic generation

Generate subtopics by independently combining the query *Q* and the query expansions terms together. In this way, we can generate a readable subtopic that describes an aspect of the query.

2.2 K-means Algorithm

In this section, we will show the employ of K-means algorithm for diversity clustering of query subtopics. In order to make use of the specificity feature of the "*相关搜索 (related searches)*" at bottom part of HTML pages, and avoid similar to QENEW's output, we combine the QENEW algorithm and K-means algorithm together to make diversity clustering of query subtopics. During making experiments, we take the default values of parameters in K-means algorithm and set the cluster number is between 5 and 10. As for the ranking score, the normalized document numbers of every cluster are the final score and the term from the shortest document in per cluster is the subtopic.

2.3 Topic Model

On the basis of the query expansion based on new-word extraction algorithm, we utilize topic model to make further experiments of subtopic terms generation, which is combining the output of QENEW and external crawled data resource as the input of topic model to generate another result. The Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) are the models mentioned in our experiment. In topic model, we respectively set the topic number of LDA and HDP is respectively 8 and use one topic words to describe the corresponding topic. And also, the top 10 topic words with the highest occurrence probabilities from the 16 topics are the subtopics of the query, whose score is the ranking order. It is worth noting that the 5th submitted run is in relation to the topic model while the four remaining (i.e., IMC-Q-C-1S, in section 3.3) are not.

3. EXPERIMENTS

In this section, we describe the datasets that used in our experiment, then after introducing the experimental parameters we make a detailed comparison for the use of datasets and technologies among the submitted runs.

3.1 Datasets

Commercial search engines with diversified search results have been applying to the study of information retrieval. Apart from the official corpus, using data from the commercial search engines usually contributes to select intent terms with better coverage of the query aspects and improve the diversity of search results.

Considering the top retrieved HTML documents contain the contents that could satisfy users' diversified search needs of a query, we collect top 5 HTML documents for each query in one of the three search engines such as Baidu[3], Google[4], Bing[5].

In all, our corpus($\boldsymbol{C}$) is composed of three parts, which are $\boldsymbol{C_0}, \boldsymbol{C_1}$and $\boldsymbol{C_2}$, using an alternative simple equation is expressed as: $\boldsymbol{C} = \boldsymbol{C_0} + \boldsymbol{C_1} + \boldsymbol{C_2}$. The specific meaning of the three signs is introduced below:

C_0: The corpus provided by the official are marked as $\boldsymbol{C_0}$.

$\boldsymbol{C_1}$: The data coming from these documents in HTML pages is flagged as $\boldsymbol{C_1}$.

$\boldsymbol{C_2}$: As for the label of "*相关搜索 (related searches)*" in the bottom of HTML pages, it's usually viewed as a specific interpretation of query's aspect. Use symbol $\boldsymbol{C_2}$ to represent the data of related searches denoting the second part of the external corpus collections.

3.2 Experimental parameters setting

In the proposed QENEW algorithm, it has a series of parameters, which is the maximal length of a new word, the standard conditions of filtering algorithm, the three coefficients in our ranking function and the number of top K new words. Finally, the optimal parameters used in our experiments are l_{seq} =10, K=10, *T*=10, *SC*=50, *SL*=0.4, α_1=0.45, α_2=0.2 and α_3=0.35.

3.3 Submitted runs

During the participation, our team submitted 5 runs for the query understanding subtask and each detailed description of the runs are listed as follows:

(1) *1st run*:: IMC-Q-C-1S: Datasets used in this run are $\boldsymbol{C_0}, \boldsymbol{C_1}$ and $\boldsymbol{C_2}$, and the input of QENEW algorithm is $\boldsymbol{C_1}$and $\boldsymbol{C_0}$. The QENEW's output together with dataset $\boldsymbol{C_2}$ uses K-means algorithm to generate the final run mentioned in section 2.2.

(2) *2nd run*:: IMC-Q-C-2S: Datasets used in this run are $\boldsymbol{C_0}$, $\boldsymbol{C_1}, \boldsymbol{C_2}$ and the input of QENEW algorithm is $\boldsymbol{C_0}$ $\boldsymbol{C_1}$, and$\boldsymbol{C_2}$. The algorithm's output is the final run.

(3) *3rd run*:: IMC-Q-C-3S: This run is similar to IMC-Q-C-2S, except that the datasets are $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$.

(4) *4th run*:: IMC-Q-C-4S: This run is the same as IMC-Q-C-2S, except that the measuring features in QENEW ranking method is different from the other four runs. The fourth run's ranking method is based on the sole frequency feature without the use of string cohesion and string liberalization, that is $\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 0$.

(5) *5th run*:: IMC-Q-C-5S: This run is similar to IMC-Q-C-1S, except that the further process is built on the topic model in section 2.3.

4. CONCLUSIONS

In this paper, we present our system on NTCIR-12 IMine-2 task, for Chinese query understanding subtask. We experiment with a novel strategy for generating query expansion based on new words extraction theory, which purely consider the data's inner features and employ the information entropy theory and statistical language knowledge to make experiments. However, our QENEW algorithm still needs further investigation. For example, improve ranking method and excavate more useful features to generate more efficient query expansion terms.

[3] https://www.baidu.com/

[4] https://www.google.com.hk/

[5] http://cn.bing.com/

5. REFERENCES

[1] Wang C J, Lin Y W, Tsai M F, et al. NTU Approaches to Subtopic Mining and Document Ranking at NTCIR-9 Intent Task[C]//NTCIR. 2011.

[2] Wang C J, Lin Y W, Tsai M F, et al. Mining subtopics from different aspects for diversifying search results[J]. Information Retrieval, 2013, 16(4):452-483.

[3] Zhang Z, Sun L, Han X. Learning to Mine Query Subtopics from Query Log[J]. Volume 2: Short Papers, 341.

[4] Clarke C L, Craswell N, Soboroff I. Overview of the trec 2009 web track[R]. WATERLOO UNIV (ONTARIO), 2009.

[5] Hu Y, Qian Y, Li H, et al. Mining query subtopics from search log data[C]//Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012: 305-314.

[6] Dou Z, Hu S, Luo Y, et al. Finding dimensions for queries[C]//Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011: 1311-1320.

[7] Wang, X., Chakrabarti, D., Punera, K.: Mining broad latent query aspects from search sessions. In: Proceedings of the 15th KDD, pp. 867 – 876 (2009).

[8] Bouchoucha A, Nie J Y, Liu X. Université de Montréal at the NTCIR-11 IMine Task[C]//NTCIR. 2014.

[9] Song, R., Zhang, M., Sakai, T., Kato, M. P., Liu, Y., Sugimoto, M., Wang, Q. (2011). Overview of the NTCIR-9 INTENT Task. In Proceedings of the 9th NTCIR Workshop Meeting.

[10] Yamamoto T, Liu Y, Zhang M, et al. Overview of the NTCIR-12 IMine-2 Task (early draft)[J]