Query refinement: type aware approach[*]

Zahra Taherikhonakdar
University of Windsor
taherik@uwindsor.ca

Aaditya Pradipbhai Parekh
University of Windsor
parekh23@uwindsor.ca

ABSTRACT

The goal of this proposed method is to investigate whether considering query types in query refinement (QR) improves the performance of information retrieval (IR) performance. Knowing what refinement methods perform better on what query types could lead search engines to select the best refinement methods based on given query types. In this paper, we define the problem that we will solve and formalize it. The proposed method is also presented.

KEYWORDS

Query Refinement, Query Types, Information retrieval

1 PROBLEM DEFINITION

Considering the query types, the goal is to specify the query refinement methods that improve the information retrieval performance. Here, we provide a formal definition of the problem, after which we propose our approach in detail in the next section. We define the input as a set of queries as $Q = q_1, q_2, .., q_n$. Each $q_n \in Q$ is categorized into n query types as $\mathcal{T} = t_1, t_2, ..., t_n, \forall q_n \in Q, \exists t_n \in \mathbb{T} : q_n \in t_n, t_n \subset Q$. Each query refinement methods $\mathcal{M} = m_1, m_2, .., m_n$ is performed on each input $q_n \in Q$. The refined query $q' \in Q'$ is used to retrieved documents . Then, the evaluation metrics (e.g., MAP) is used to evaluate the performance of each query refinement methods. As an outputs, we select $m_n \in \mathcal{M}$ which ranks high based on evaluation metrics for each $t_n \in \mathcal{T}$.

2 PROPOSED APPROACH

Our proposed approach seeks to find the best query refinement method for each query type. This would help search engines to select the query refinement methods that enhance the IR performance for each query considering its query type. The approach works through three phases: preprocess the data set, get evaluation results of QR methods, specify a suitable QR method in each query type. Figure 1 shows our proposed method diagram. Moreover, in the following, we describe the details of each phase.
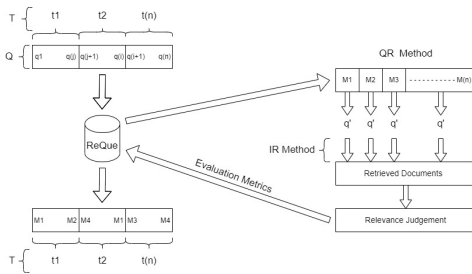

Figure 1: The diagram of our proposed method.

2.1 Preprocess The Data Set

We use the TREC 2009 Million Query data set [1] as the input of our method. This data set consists of 40,000 queries with their relevance judgments. Only the query type of the 420 of the queries are identified. In this step, we delete unnecessary columns from the original data set and only keep query and query id as inputs. Table 1 shows the example of our input $Q = q_1, q_2, .., q_n$. We used the preprocessed form of the original data set to make it appropriate as an input for our next step, which we explain in the following section.

20051	cuttle fish bone
20054	50plus
20055	usairway
20065	rock and gem shows

Table 1: The input queries and their id

2.2 Evaluation Results of QR Methods

We use the ReQue [2] as a toolkit to pass our queries as an input and return the results of the evaluation of QR methods. ReQue consists of 3 main steps. First, use different query expander techniques as query refinements methods to generate new query q', given the $q \in Q$. Table 2 illustrates the new queries for the initial queries based on QR methods. we also show the procedure of the ReQue in Figure 1.

QR Method	Initial Query	New Query
conceptnet	rock and gem shows	music stone gem things
conceptnet	cuttle fish bone	cuttlefish animal A bass skeleton body part
conceptnet	50plus	plus
stem.krovetz	rock and gem shows	rock and gem show
stem.krovetz	cuttle fish bone	cuttle fish bone
stem.krovetz	50plus	plus

Table 2: The new queries are generated from initial query by using the different QR Method

As Table 2 shows, some QR methods remove words from the original query or add new words to the original one, and some of them remain unchanged. Then, the new queries are used to retrieve documents using IR methods (e.g., bm25). Finally, evaluation metrics (e.g., MAP) are used to identify the relatedness of retrieved documents based on the given new query.

2.3 Specify Suitable QR Method For Given Query Types

So far, we have the evaluation results of QR methods for each input query. For our final step, we divided queries into different query types. Therefore, in each specific query type, we have queries with their QR methods evaluation. We ranked queries' evaluation results in each query type. The K-top query method will be selected as a nominated QR method for a given query type.

In this research, we consider queries which are divided into the following categories: informational-open, which indicates the unspecific research need; informational-close refers to the specific research needs; the intention of navigational queries are finding a specific URL and resource which also called transactional queries are trying to find web-based resources like download the music[1]. $Algorithm$1 shows the Pseudo-code for our proposed method.

Algorithm 1 Finding suitable QR method for each query type

    Inputs:
        set of $\mathbb{Q} = q_1, q_2...q_n$, and its relevance judgement
    Initialization:
        dividing queries $\mathbb{Q} = q_1, q_2...q_n$ into different query types
    $\mathbb{T} = t_1, t_2...t_n$
        QR methods $\mathbb{M} = m_1, m_2...m_n$
    Output: $m_n \in \mathcal{M}$ for each $t_n \in \mathcal{T}$

1: **procedure** FIND SUITABLE m_n FOR EACH t_n
2:     **for all** $q_n \in Q$ **do**
3:         **for all** $m_n \in \mathcal{M}$ **do**
4:             $q^{'} \leftarrow m(q)$
5:     **for all** $q^{'} \in Q^{'}$ **do**
6:         retrieved documents $d \leftarrow IRMethod(q^{'})$

        evaluation- results$\leftarrow evaluation - metric(for all q^{'} and d)$
        ranked $m \in \mathcal{M}$ for each $t \in \mathcal{M}$

REFERENCES

[1] Ben Carterette, Virgiliu Pavlu, Hui Fang, and Evangelos Kanoulas. 2009. Million Query Track 2009 Overview.. In *TREC*. Citeseer.
[2] Mahtab Tamannaee, Hossein Fani, Fattane Zarrinkalam, Jamil Samouh, Samad Paydar, and Ebrahim Bagheri. 2020. Reque: a configurable workflow and dataset collection for query refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3165–3172.