QRU-1: A Public Dataset for Promoting Query Representation and Understanding Research

Hang Li
Microsoft Research Asia
Danling Street, Haidian
Beijing, China
hangli@microsoft.com

Gu Xu
Microsoft
One Microsoft Way
Redmond, WA
guxu@microsoft.com

W. Bruce Croft
Dept. of Computer Science
University of Massachusetts
Amherst, MA
croft@cs.umass.edu

Michael Bendersky
Dept. of Computer Science
University of Massachusetts
Amherst, MA
bemike@cs.umass.edu

Ziqi Wang
Dept. of Computer Science
Peking University
Beijing, China
zikkiwang@gmail.com

Evelyne Viegas
Microsoft
One Microsoft Way
Redmond, WA
evelynev@microsoft.com

ABSTRACT

A new public dataset for promoting query representation and understanding research, referred to as QRU-1, was recently released by Microsoft Research. The QRU-1 dataset contains reformulations of Web TREC topics that are automatically generated using a large-scale proprietary web search log, without compromising user privacy. In this paper, we describe the content of this dataset and the process of its creation. We also discuss the potential uses of the dataset, including a detailed description of a query reformulation experiment.

1. INTRODUCTION

Understanding the user's intent or the information need that underlies a query has long been recognized as a crucial part of effective information retrieval. With the recent availability of large amounts of data about user behavior and queries in web search logs, there has been an upsurge of interest in new approaches to query understanding and representing user intent.

In order to promote discussion of these approaches and to identify challenges and long term research goals, we organized a series of two workshops on Query Representation and Understanding at SIGIR 2010 and SIGIR 2011 [2]. The readers may refer to the workshop website[1] for more detailed information about these workshops.

A key issue, raised by many workshop participants, was the importance of creating public datasets for query representation and understanding research. There are two main obstacles that hinder the availability of such datasets. First, most academic researchers do not have access to web search logs. Second, commercial search engines restrict access to their proprietary data due to privacy concerns. Accordingly, there is a need in a creative approach that can leverage the wealth of data in web search logs without compromising user privacy.

A new public dataset that is described in this paper is an example of such an approach. This dataset is named QRU-1 (short for *Query Representation and Understanding*), and is available for download on the Microsoft Research website[2], as well as on the workshop website.

The QRU-1 dataset is constructed based on the topics that were developed during the TREC Web Track [1]. For each of the hundred topics used in Web Track in TREC 2009 and TREC 2010, we assign approximately twenty similar queries. The similar queries assigned to the original TREC topic represent the same intent, but are expressed in different forms, including synonyms, stemming variations, spelling errors and abbreviations.

The similar queries in the QRU-1 dataset are automatically generated from a model trained from Bing search log data with the title of the TREC topic as an input. In addition, a manual cleaning of the generated queries is also performed and unlikely similar queries are discarded, based on a predetermined guideline. It is observed that 70% of the remaining similar queries actually occur in another Bing search log. In this way, the similar queries are generated from a model instead of being directly collected from the search log, and thus we can more effectively protect the privacy of the search engine users.

The QRU-1 dataset can be used in a variety of tasks, including query rewriting, query suggestion, query segmentation and query expansion (for precise definition of these tasks, please refer to Croft et al. [2]). As a case study, in this paper, we investigate the potential of the proposed dataset for

[1] http://ciir.cs.umass.edu/sigir2011/qru/

[2] http://research.microsoft.com/en-us/downloads/d6e8c8f2-721f-4222-81fa-4251b6c33752/default.aspx

Workshop on Web Search Click Data, February 12, 2012, Seattle, Washington, USA.

```
Topic #1: obama family tree
--------
barack obama family
obama family
obama s family
barack obama family tree
the obama family
barack obama s family
obamas
obama genealogy
barack obama s family tree
barack obama ancestry
president obama s family
obamas family
obama family history
obama s family tree
barack obama genealogy
barack obama family history
barack obama geneology
president obama and family
obama s ancestry
barak obama family tree
barak obama family
obama family tre
obama and family tree
```

Figure 1: Example of the TREC topic "obama family tree" and its similar queries.

improving the relevance of web search results using query reformulation.

The remainder of this paper is organized as follows. Section 2 introduces the content of the QRU-1 dataset. Section 3 explains the creation process of the QRU-1 dataset, and Section 4 describes the use of this dataset for query reformulation. A detailed explanation of the model for generating similar queries is deferred to the appendix.

2. CONTENT OF THE QRU-1 DATASET

The QRU-1 dataset is based on the topics developed for the TREC Web Tracks in 2009 in 2010. QRU-1 contains 100 TREC topics, and each topic is represented by a short title, which is commonly used as a search query in Web Track runs [1]. For example, "obama family tree" and "earn money at home" are examples of the TREC topics, included in the dataset.

There are approximately 20 similar queries generated for each of these 100 topics in the QRU-1 dataset. Figure 1 and Figure 2 show the similar queries generated for topics "obama family tree" and "earn money at home", respectively. As can be seen from these examples, similar queries represent the same intents, but may take different forms including synonyms, stemming variations, and spelling errors. In total, there are 2,036 similar queries in the QRU-1 dataset.

3. CREATION OF THE QRU-1 DATASET

3.1 Overview

```
Topic #95: earn money at home
---------
earn money from home
earn money at home
how to earn money at home
earn money on the internet
ways to earn money at home
how to earn money from home
earn extra money at home
earning money from home
earn extra cash at home
earning money at home
earn at home
earn money working from home
earn money from home free
how to earn money on the internet
earn cash at home
earn currency at home
earn money at hom
earn money at hoem
```

Figure 2: Example of the TREC topic "earn money at home" and its similar queries.

Table 1: Examples of retained and discarded queries for the TREC topic "obama family tree".

Similar Query	Status
barack obama family	retained
barak obama family tree	retained
obama family	retained
pictures of the obama family	discarded
michelle obama family tree	discarded
obama family plant	discarded

Microsoft has strict rules for protecting users' privacy. To reduce the risk of privacy infringement, we chose to release a query dataset through filtering by a model, rather than to directly release a dataset containing real user queries. The resultant set of queries has both the minimum privacy risk and the maximum utility, as the generated queries are very close to real user queries.

The process of similar query generation is as follows. At the first stage, we take all the TREC topic titles and generate about 30 similar queries (e.g., "ny times" for the query "new york times") for each topic using the query generation method described in Section 3.2. These similar queries are likely to represent the same or similar search intent as the original TREC topic, and may contain synonyms, typos, stemming variations and abbreviations.

At the second stage, we manually clean the generated query set by removing unlikely queries that do not convey the same query intent or can endanger user privacy. The guideline and the principles of the data cleaning process are detailed in Section 3.3.

3.2 Similar Query Generation

In our query generation method, given an input query (in this case, a title of a TREC topic), we first attempt to find all the high-frequency head queries that are similar to it in a click-through bipartite graph. In addition, we employ a

Table 2: Retrieval performance using the QRU-1 dataset.

BM25	**MAP**	**NDCG@20**	**ERR@20**
Baseline metric	16.84	20.97	10.19
Best metric	21.98 *(+30.5%)*	30.03 *(+43.2%)*	16.05 *(+57.5%)*
% outperforming queries	12%	16%	18%
% topics improved	59%	60%	64%

SD	**MAP**	**NDCG@20**	**ERR@20**
Baseline metric	19.13	20.19	8.34
Best metric	25.00 *(+30.7%)*	32.88 *(+62.9%)*	15.09 *(+80.9%)*
% outperforming queries	12%	16%	18%
% topics improved	51%	63%	67%

string transformation model to generate similar queries from the input query.

The similar head queries are collected from a click-through bipartite graph. Intuitively, if two queries share many clicked URLs, they are viewed as similar queries. We use Pearson Correlation Coefficient as a similarity measure, as proposed by Xu et al. [7]. Only high frequency queries are extracted to protect user privacy.

Other type of similar queries is generated from a string transformation model. The string transformation model is trained on head queries, which can learn substitution rules such as 'ny' to 'new york'. The model is then applied to compose new similar queries. A detailed description of this string transformation model, first proposed by Wang et al. [6], can be found in the appendix.

After this process, there are about 30 queries, generated for each of the original TREC topics. In order to make the resulting dataset more effective for research purposes, we perform manual cleaning of the data, based on the guideline described in the next section.

3.3 Data Cleaning

In the data cleaning process, we manually clean the generated similar queries. About 23% of generated queries are removed. For the remaining queries, around 70% of them could be found in another search log of Bing.

Below is the guideline for data cleaning. If a generated query meets the following conditions, it is retained in the dataset; otherwise it is discarded.

1) The generated query represents the same intent as the original query. The original TREC topics are often ambiguous, and may contain more than one subtopic. The generated query will be retained if it represents any of the subtopics, or it is judged by the annotator as representing a likely subtopic.
2) It is likely to be input by users, including typos.

The examples of similar queries for 'obama family tree' are shown in Table. 1. The generated queries such as 'pictures of the obama family', 'obama family plant', 'michelle obama family tree' were discarded, because they do not represent the same intent as the original query.

4. EXPERIMENTAL EVALUATION

The QRU-1 dataset can be useful in a variety of scenarios. In this paper, we investigate its potential for improving the relevance of results in web search. However, the QRU-1 dataset can be beneficial for many other tasks, including query suggestion, query expansion and result diversification. We leave the exploration of these tasks for future work.

Our primary goal in this section is to establish the *potential* of the dataset to benefit future research. Therefore, instead of focusing on specific techniques for integrating the generated similar queries into the retrieval process, we explore the *upper bound* of their potential influence on the relevance of retrieval results.

Following a previous study on query reformulation by Dang and Croft [3], we record the retrieval performance of the most effective formulation of the query (including the original topic title) for a given TREC topic. That is, we effectively simulate the actions of an oracle user who always selects the best-performing query among all the proposed candidates. In this fashion, we calculate the upper bound on the retrieval effectiveness that can be achieved using our dataset.

All the retrieval experiments are conducted using the web corpus ClueWeb-B, which is a set of approximately 50 million pages with the highest crawl priority derived from a large web corpus. We use two retrieval models. The first retrieval model is the standard BM25 retrieval model. The second retrieval model is the sequential dependence model (SD), first proposed by Metzler and Croft [4]. Sequential dependence is a state-of-the-art retrieval model which linearly combines terms, phrases and proximity matches. All the retrieval experiments are implemented using Indri, an open-source search engine [5].

We use MAP, NDCG@20 and ERR@20 (the standard retrieval metrics used at the TREC Web Track [1]) to measure the retrieval performance. For each of these metrics, we report the following statistics for both BM25 and SD: (a) the baseline metric (achieved by the original query), (b) the best metric (achievable by either the original query or one of the generated queries), (c) the percentage of generated queries that outperform the original query, and (d) the percentage of topics, which performance is improved by using at least one of the generated queries. The results are reported in Table 2.

Table 3: Examples of effective query reformulations using the QRU-1 dataset.

Topic Title #1	*obama family tree*	**ERR@20**	*13.42*
Reformulations	barack obama ancestry		32.93
	obama s family		32.40
	barack obama s family		32.05
Topic Title #5	*mitchell college*	**ERR@20**	*1.2*
Reformulations	mitchell college new london		19.6
	mitchell college new london ct		19.2
	www mitchell edu		5.7
Topic Title #30	*diabetes education*	**ERR@20**	*9.77*
Reformulations	national diabetes education program		15.30
	diabetes education program		15.13
	national diabetes education		14.79
Topic Title #40	*michworks*	**ERR@20**	*7.9*
Reformulations	michworks talent bank		22.3
	mi works talent bank		12.1
	www michworks org		11.4
Topic Title #44	*map united states*	**ERR@20**	*3.42*
Reformulations	map usa states		13.92
	map usa		10.34
	united states america map		7.33
Topic Title #91	*er tv show*	**ERR@20**	*7.1*
Reformulations	er tv series		23.5
	er television series		20.1
	er tv		19.9

Table 2 demonstrates that the generated queries can significantly improve the retrieval performance both for the BM25 and the SD retrieval models. By using the QRU-1 dataset, the retrieval effectiveness can be potentially improved for up to two thirds of the TREC topics (in terms of ERR@20). These improvements are consistent across all metrics, and are especially visible for metrics that measure early precision. For instance, in terms of ERR@20, the original retrieval effectiveness can be almost doubled (in the case of the SD retrieval model) by selecting the best performing query among the generated similar queries and the original TREC topic title.

It is important to note, however, that not all of the generated queries are equally helpful. As Table 2 shows, less than 20% of the generated queries are better than the original topic title. This is consistent with query reformulation results using query logs and anchor text, as reported by Dang and Croft [3]. Therefore, an automatic query selection for effective reformulation is an important research topic that could be advanced by the availability of the QRU-1 dataset.

Finally, it is interesting to examine the topics for which the QRU-1 dataset provides effective query reformulations. Table 3 shows examples of such topics, including the original topic titles and three of the top-performing (when the SD retrieval model is used) generated queries. These examples illustrate several effective query reformulation strategies that can be accomplished using the QRU-1 dataset.

1. *Abbreviation induction* strategy can insert relevant abbreviations into the original query (e.g., replacing "united states" with the abbreviation "usa" in query "map united states")
2. *Query expansion* strategy can add relevant terms to the original query (e.g., expanding the query "mitchell college" with the location "new london").
3. *Query substitution* strategy can overcome problems of vocabulary mismatch, and replace non-discriminatory terms with terms that better represent the user intent (e.g., substituting a common term "show" with a more specific term "series" in the query "er tv show").
4. *Query reduction* strategy can improve retrieval effectiveness by removing unhelpful terms from the query (e.g., removing the term "tree" in the query "obama family tree").
5. *URL suggestion* strategy can provide the most relevant website for navigational queries (e.g., replacing the original query "michworks" with the URL `www.michworks.org`).
6. *Source suggestion* strategy can help in answering general informational queries by providing an authoritative information source (e.g., replacing the original query "diabetes education" with "national diabetes education program").

We note that these strategies are complementary, and can be combined to further improve retrieval effectiveness, which is an interesting venue for future research.

5. SUMMARY

In this paper, we have described the development of the QRU-1 dataset. The purpose of the QRU-1 dataset is to provide a resource for query representation and understanding research, which is often hindered by the inaccessibility of proprietary web search logs. As an example of such research, we demonstrate the potential of the QRU-1 dataset

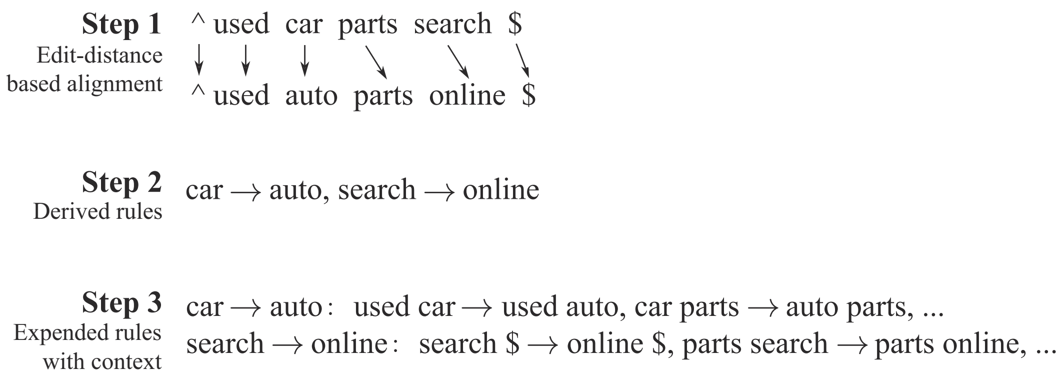
Step 1
Edit-distance based alignment
^ used car parts search $
^ used auto parts online $
Step 2
Derived rules
car → auto, search → online
Step 3
Expended rules with context
car → auto : used car → used auto, car parts → auto parts, ...
search → online : search $ → online $, parts search → parts online, ...

Figure 3: Rule extraction example.

for significantly improving the relevance of search results. We hope that the QRU-1 dataset will benefit the research community, and will prove to be a useful resource for a variety of research tasks involving query representation and understanding.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by ARRA NSF IIS-9014442. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

7. REFERENCES

[1] C. L. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web Track. In *Proceedings of TREC 2010*, 2010.

[2] W. B. Croft, M. Bendersky, H. Li, and G. Xu. Query representation and understanding workshop. *SIGIR Forum*, 44(2):48–53, 2010.

[3] V. Dang and W. B. Croft. Query reformulation using anchor text. In *Proceedings of WSDM*, pages 41–50, 2010.

[4] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of SIGIR*, pages 472–479, 2005.

[5] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proc. of IA*, 2004.

[6] Z. Wang, G. Xu, H. Li, and M. Zhang. A fast and accurate method for approximate string search. In *Proceedings of ACL-HLT*, pages 52–61, 2011.

[7] J. Xu and G. Xu. Learning similarity function for rare queries. In *Proceedings of WSDM*, pages 615–624, 2011.

APPENDIX

A. SIMILAR QUERY GENERATION

We employed a probabilistic string transformation method to generate similar queries. We refer the interested readers to Wang et al. [6] for details. Here we give a brief introduction to the method.

There are two processes in this method, learning and generation. In the learning process, rules for string transformation (e.g., 'ny' to 'new york') are first extracted from training string pairs. Then the model of string transformation is constructed by the learning system, consisting of rules and their weights. In the generation process, given a new input string (e.g., 'ny times'), the generation system produces the top k candidates of output strings by referring to the model (e.g., 'new york time').

The model consists of rules and weights. A rule is formally represented as $\alpha \rightarrow \beta$ which denotes an operation of replacing substring α in the input string with substring β. All the possible rules are derived from the training data based on string alignment. Fig. 3 shows the steps of rule extraction. First we align the words in the input string and the output string based on word-level edit-distance, and then derive rules from the alignment. Next we expand the derived rules with surrounding contexts. Without loss of generality, we only consider using $+2, +1, 0, -1, -2$ words as contexts in this paper.

Let (s_i, s_o) denote a string pair. If a set of rules, $R(s_i, s_o)$, can be utilized to transform the input string s_i to an output target s_o, then the rule set is said to form a "transformation" for the string pair s_i and s_o. Note that for a given string pair, there might be multiple possible transformations for it. We consider a conditional probability distribution of s_o and $R(s_i, s_o)$ given s_i and take it as model for string transformation. We specifically define the model as the following log linear model:

$$\begin{aligned} & P(s_o, R(s_i, s_o)|s_i) \\ = & \frac{\exp\left(\sum_{r \in R(s_i,s_o)} \lambda_r\right)}{\sum_{(s'_t, R(s_i,s'_t)) \in \mathcal{Z}(s_i)} \exp\left(\sum_{o \in R(s_i,s'_t)} \lambda_o\right)} \end{aligned} \tag{1}$$

where r and o denote rules, λ_r and λ_o denote weights, and the normalization is carried over $\mathcal{Z}(s_i)$, all pairs of string s'_t and transformation $R(s_i, s'_t)$, such that s_i can be transformed to s'_t by $R(s_i, s'_t)$. The log linear model actually uses binary features to indicate whether or not rules are applied.

In string generation, given an input string s_i, we aim to generate the k output string candidates s_o that can be transformed from s_i and have the largest probabilities $P(s_o, R(s_i, s_o)|s_i)$ assigned by the learned model. We only need to utilize the following scoring function to rank candidates of output strings s_o given an input string s_i.

$$\text{rank}(s_o|s_i) = \max_{R(s_i,s_o)} \left(\sum_{r \in R(s_i,s_o)} \lambda_r \right) \tag{2}$$

For each possible transformation, we simply take summation of the weights of the rules used in the transformation.