Building a Test Collection for Evaluating Search Result Diversity: A Preliminary Study

Hua Liu[1]*, Ruihua Song[2,3], Jian-Yun Nie[4], Ji-Rong Wen[3]
[1]Xi'an Jiao Tong University, [2]Shanghai Jiaotong University
[3]Microsoft Research Asia, [4]University of Montreal
doudoulh@gmail.com, {rsong,jrwen}@microsoft.com, nie@iro.umontreal.ca

ABSTRACT

Users often issue vague queries. When we cannot predict users' intentions, a natural solution is to improve user satisfaction by diversifying search results. Such an area, usually called "result diversification", lacks a systematic approach to construct a test collection, by which we can evaluate how search systems perform. In this paper, we propose leveraging the user contributed data in Wikipedia[1] to build up a test collection for ambiguous queries. A preliminary experiment shows promising results.

1. INTRODUCTION

Queries issued by Web users often have multiple meanings or intentions. For such queries, it is important for search engines to retrieve documents covering different requirements. Sanderson [2] has surveyed previous research work on ambiguity and the effort taken to diversify search results. Although there is a long history of research on addressing ranking problems for ambiguous queries, little work done to build test collections has hampered research of this type. This motivates us to construct a test collection that has ambiguous topics and a range of relevance judgments with regard to more than one interpretation.

It is challenging to sample representative ambiguous queries and enumerate their different intentions. First, a set of ambiguous queries proposed by a few people tend to be biased by individual experiences. Second, it is costly to sample ambiguous queries from query logs manually because it is difficult for humans to judge whether a query is ambiguous. Third, even if we have ambiguous queries sampled, there are still difficulties in listing all major intentions of a query.

Fortunately, thousands of people contribute a huge amount of knowledge to Wikipedia. For an ambiguous entry, Wikipedia provides a disambiguation page to allow users to choose their interested interpretations. We propose the idea of leveraging such data to sample queries, pool documents, and labeling the intentions that a document is relevant to. In a preliminary experiment, we build a test collection containing 50 representative queries for evaluating result diversification.

*Work was done when the author was visiting Microsoft Research Asia

[1]www.wikipedia.org

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

2. BUILDING A TEST COLLECTION

In general, an IR test collection is comprised of queries, documents, and judgments for query document pairs. For ambiguous queries, the intentions that a document is relevant to are also required for evaluating diversity. In this section, we describe how we leverage Wikipedia to achieve these goals.

2.1 Sampling Queries

We make use of disambiguation pages to identify ambiguous entries as Sanderson does in [2]. Then we filter the ambiguous entries from Wikipedia by checking whether it is in a half-a-year query log from a commercial search engine. This is to make it sure that our sampled ambiguous entries are real web queries. Finally, we obtain 38,606 candidate queries.

By observing the candidate queries, we find some ambiguous queries have more diverse meanings than others. For example, "TREC"[2] refers to Text Retrieval Conference, Texas Real Estate Commission, the Trans-Mediterranean Renewable Energy Cooperation, etc., which are quite different from each other. In contrast, "A Beautiful Mind"[3] tends to have more similar meanings, such as A Beautiful Mind (book), A Beautiful Mind (film), and A Beautiful Mind (soundtrack).

Therefore, to compose a set of representative ambiguous queries, we propose sampling the queries with different levels of Similarity of Intentions (SI). For each distinct meaning of an ambiguous query Q, denoted as QM_1, QM_2, ..., QM_n, we use their corresponding Wikipedia entry pages Wiki(QM_1), Wiki(QM_2), ..., Wiki(QM_n) to calculate SI as the average of cosine similarities between pairs of pages:

$$SI(Q) = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} \cos\_sim(\text{Wiki}(QM_i), \text{Wiki}(QM_j))}{n \cdot (n-1)/2}$$

where, $SI(Q)$ is in the range of 0 to 1. The larger $SI(Q)$ is, the less diverse meanings the ambiguous query Q covers.

We calculate SI for all the candidate queries and show the distribution in Figure 1. Among 38,606 ambiguous queries, 7,454 queries have SI values less than 1.0×10^{-8}, which means these ambiguous queries have quite distinct intentions. For example, "TREC" is in this group. Different from "TREC", "A Beautiful Mind" gets a medium SI value because its interpretations are related to each other. Furthermore, some examples of the queries with high SI values are "dream" (0.0835), "Hercules" (0.0509), "David Copperfield" (0.0441) and "Saint Mary's" (0.0295).

[2]http://en.wikipedia.org/wiki/TREC
[3]http://en.wikipedia.org/wiki/A_beautiful_mind

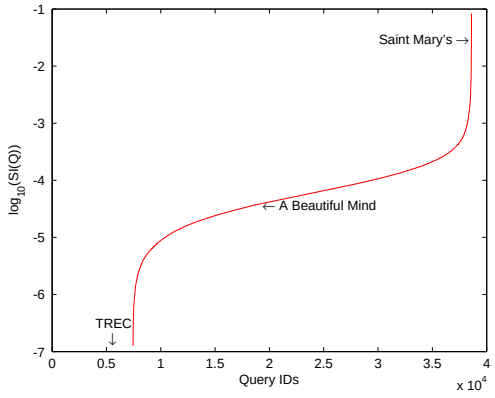

Figure 1: Distribution of $\log_{10}(SI(Q))$ among the candidate ambiguous queries

In our test collection, we randomly select 30 ambiguous queries with low SI values, 10 queries with medium SI values, and 10 queries with high SI values.

2.2 Pooling Documents

An ambiguous query alone may be not enough to retrieve the documents that are relevant to its main intentions, because some unpopular meanings may be overwhelmed by the documents on popular meanings. Thus, we create additional queries that are related to the different meanings in Wikipedia. For example, in terms of the meanings at the disambiguation page of "A Beautiful Mind", we create three additional queries: "A Beautiful Mind book", "A Beautiful Mind film", and "A Beautiful Mind soundtrack". Then we submit the query and its additional queries respectively to two commercial search engines and retrieve the top 20 returned documents for each query. Finally, by merging the retrieved documents and removing duplicates, we make a pool of documents for each sampled query.

2.3 Labeling Relevance and Topics

To evaluate result diversification, we develop a labeling tool to judge whether a document is relevant to a query as well as which main intentions the page covers. The frame on the right displays the page with keywords highlighted. On the left questionnaire frame, an annotator can mark a page as "Not Found", if the page fails to be loaded; or "Irrelevant", which means the page content is not relevant to the query at all; or "Relevant", which means the page content is relevant to the query. If "Relevant" is clicked, the annotator is also asked to choose one or more relevant intentions from a list of "candidate intentions" that are extracted from the Wikipedia disambiguation page. In addition, the annotator is allowed to input other intentions that are not covered by the list if necessary.

3. EXPERIMENTS

We set up a test collection of 50 queries in a preliminary experiment. On average, there are 5.98 intentions provided and about 213 pages judged per query. In the labeled data, annotators input new interpretations for only about 3.45% of pages. This indicates that the candidate intentions from Wikipedia can cover the meanings of ambiguous queries well. In addition, annotators select multiple intentions for 7.1 pages per query on average. Most of the pages come from dictionary-type websites, such as thefreedictionary.com and britannica.com. These websites usually have a page that shows all the meanings of an ambiguous query.

Table 1: Evaluating two search engines by using a test collection containing 50 ambiguous queries

	MAP-IA@3		MAP-IA@10	
	SE1	SE2	SE1	SE2
Low	0.401	**0.422**	0.427	**0.448**
Medium	**0.335**	0.296	**0.383**	0.337
High	**0.471**	0.437	**0.484**	0.463
All	0.402	0.400	0.429	0.429

We evaluate the performance of result diversification of two commercial search engines by using the test collection. To preserve anonymity, we refer to them as SE1 and SE2. MAP-IA proposed in [1] is used as the measure. Results are shown in Table 1.

We can see that there is no significant difference between two search engines in terms of the overall MAP-IA. However, when looking closely into different types of queries, we find the two engines are obviously different: SE2 outperforms SE1 on the ambiguous queries with clearly different intentions, whereas it performs worse than SE1 on the queries with medium and high SI values. This verifies that query sampling strategies do affect evaluation results.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we present a simple approach to build a test collection by leveraging disambiguation pages from Wikipedia. First, Similarity of Intentions (SI) is proposed to measure how different the meanings of an ambiguous query are. Based on SI, we can sample the representative queries with different properties. Second, in pooling documents, we expand an ambiguous query by additional queries from the disambiguation page. Third, we design a labeling tool that allows annotators to judge both relevance and the topics that a document is relevant to. A preliminary experiment shows that our proposed approach is feasible to construct a test collection for evaluating search result diversity.

In this preliminary study, we use Similarity of Intentions to measure how diverse the intentions of an ambiguous query are. However, there are some alternative measures, such as the number of intentions and the number of categories. Our next step is to investigate the methods and compare their performance in sampling representative queries. In addition, the set of 50 queries is too small to infer statistically sound conclusions. Is it possible to construct a large-scale dataset with minimal human effort? For example, can we label only a few documents and then employ supervised learning approaches to learn classifiers and get more labeled documents further? These interesting research problems await our future research work.

5. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM*, pages 5–14, 2009.

[2] M. Sanderson. Ambiguous queries: test collections need more sense. In *SIGIR*, pages 499–506, 2008.