Inferring Query Intent from Reformulations and Clicks

Filip Radlinski
Microsoft Research
Cambridge, UK
filiprad@microsoft.com

Martin Szummer
Microsoft Research
Cambridge, UK
szummer@microsoft.com

Nick Craswell
Microsoft
Redmond, WA, USA
nickcr@microsoft.com

ABSTRACT

Many researchers have noted that web search queries are often ambiguous or unclear. We present an approach for identifying the popular meanings of queries using web search logs and user click behavior. We show our approach to produce more complete and user-centric intents than expert judges by evaluating on TREC queries. This approach was also used by the TREC 2009 Web Track judges to obtain more representative topic descriptions from real queries.

Categories and Subject Descriptors

H.3.3. [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Algorithms
Keywords: Diversity, Intents, Subtopics

1. INTRODUCTION

Information retrieval evaluation (such as in TREC) increasingly focuses on relevance judgments collected for real queries observed in search engine logs. In particular, given a query and document, judges must first infer the *intent* (i.e. the meaning) of the query before being able to judge how *relevant* a document is to that intent. Previous work (e.g. [3]) has shown that different users often have different intents for the same query. We present a method to assist judges to discover the actual intents of queries.

For example, consider the query *ai*. As an acronym, it may stand for *artificial intelligence*, *american idol* or *art institute* among many other things. What fraction of users who type this query are interested in each? Real users also often mistype queries – a user who typed *ai* may have also meant *aim*, the instant messenger software. Moreover, ambiguity is not limited to acronyms and single word homonyms such as *jaguar*, *flash* and *mercury*. Queries such as *machine learning* also conceal multiple intents: users may be looking for an overview of modern techniques, downloadable tools or theoretical results.

Unfortunately, the possible intents and the relative importance of each cannot be simply obtained by clustering retrieved documents. What matters is the popularity of user *needs* rather than the documents the search engine retrieves, and some cases (such as that mapping *ai* to *aim*) cannot be captured by clustering documents.

WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

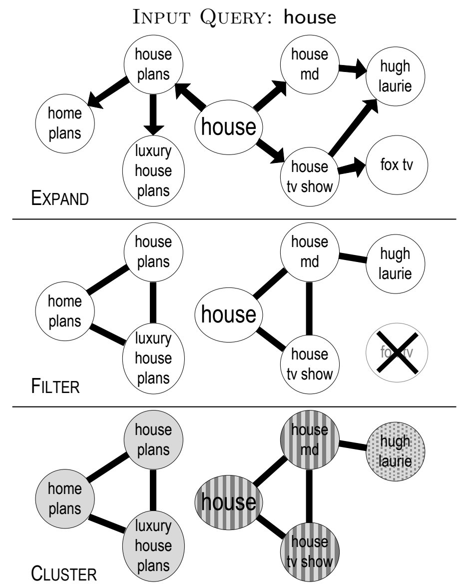

Figure 1: Three Steps of Intent Identification

2. CLICKS AND REFORMULATIONS

After entering a query and being presented with results, web search users often click on the results and/or follow up with other queries. Many researchers have showed that clicks and reformulations can be used for a variety of tasks.

Our key contribution is the observation that by *combining* evidence from clicks and reformulations, using logs of many millions of queries from a commercial search engine, we are able to identify the most probable intents for a query. The combination is important because clicks and reformulations capture different types of information. Reformulations often provide alternative phrasings of the user's intent, for instance when the user was not presented with (or did not notice) sufficiently satisfying results, or would like more results. On the other hand, clicks indicate the user was presented with apparently worthwhile documents.

3. IDENTIFYING INFORMATION NEEDS

Given an input query q, we combine click and reformulation information to find likely user intents using three steps: EXPAND, FILTER, CLUSTER. Each step is described next.

The goal of the EXPAND step is to identify a set of possibly related queries to q. Recall is key here: We wish to find all possible intents, including those missing from the results. EXPAND finds the $k = 10$ most frequent valid reformulations of q, then the k most frequent valid reformulations of those. We

Table 1: Representative queries for each top cluster on the first TREC 2003 Web track topic queries. w_c is the estimated relative popularity of the information need represented by each cluster.

TD2: *juvenile delinquency*	**TD6**: *physical therapists*
Clusters: juvenile delinquency (w_c=1.16); causes juvenile delinquency (w_c=0.50); delinquency prevention (w_c=0.25); definition of juvenile delinquency (w_c=0.20); articles on juvenile delinquency (w_c=0.18); reasons for juvenile delinquency (w_c=0.15)	**Clusters:** physical therapist (w_c=1.22); physical therapists salary (w_c=0.80); how to become a physical therapist (w_c=0.21); physical therapy schools in california (w_c=0.15); physical therapist school of california (w_c=0.11); physical therapist assistant programs(w_c=0.10)
Editors: What are rates of juvenile crime in various jurisdictions, what is the nature of the offenses, how are they punished and what measures are taken for prevention?	**Editors:** How can I obtain information about training, licensing, and skills needed for physical therapists?
TD3: *Lewis and Clark expedition*	**TD4**: *wireless communications*
Clusters: lewis and clark (w_c=1.63) lewis and clark expedition facts (w_c=0.34); lewis clark map (w_c=0.28); pictures of lewis and clark (w_c=0.19); sacagawea (w_c=0.14); lewis and clark timeline (w_c=0.11);	**Clusters:** wireless communications (w_c=1.07); what is wireless comm. (w_c=0.56); wireless comm. systems (w_c=0.19); history wireless technology (w_c=0.13); wireless cell phone companies(w_c=0.13); wireless broadband providers(w_c=0.10)
Editors: What are some useful sites containing information about the historic Lewis and Clark expedition?	**Editors:** Information on existing and planned uses, research/technology, regulations and legislative interest

say q' is a *valid* reformulation of q if (1) q was followed by q' within ten minutes by at least 2 distinct users, and (2) of all pairs of queries (q_i, q') issued by any user within 10 minutes, (q, q') occurred at least a fraction δ of the time (we set $\delta = 0.001$). This last constraint avoids very frequent queries (such as *myspace* and *hotmail*) appearing as valid reformulations for all queries. Thus we obtain a query neighborhood, as illustrated for a toy example in the top plot in Figure 1.

Next, the FILTER step reduces the query neighborhood to more closely related queries, improving precision. Illustrated in the middle of Figure 1, we connect two queries if they were often clicked for the same documents, using a two step random walk on the bipartite query-document click graph [2]. All pairs of queries with a random walk similarity above a fixed threshold are connected (this may add links not present in the reformulation graph, and usually removes many others). Additionally, all components of size less than t are removed completely (we use $t = 2$).

Finally, the CLUSTER step uses the random walk similarities to find intent clusters. Although any clustering algorithm can be used, we use [1] as it is efficient for large graphs and automatically selects the number of clusters. An illustrative clustering is shown at the bottom of Figure 1.

Estimating Information Need Popularity

To assign an importance to each intent, we use the sum of a zero, one and two step random walk on the reformulation graph. Specifically, the weight w_q of the input query q is $w_q = 1$. The weight of every valid reformulation q' of q is $w_{q'} = w_q \cdot N(q{\rightarrow}q')/\sum_{r\in R(q)} N(q{\rightarrow}r)$, where $N(q{\rightarrow}q')$ is the number of times the reformulation q to q' was observed, and $R(q)$ is the set of all q's. This assignment is repeated for a second step. Finally, the probabilities of arriving at each query on multiple paths are summed. The importance of each intent cluster w_c is the sum of the weights of the queries in that cluster. Moreover, the query with highest weight w_q in each cluster is taken as a cluster representative.

4. EVALUATION AND DISCUSSION

We illustrate the effectiveness of this approach using the first six TREC 2003 Web Track topic distillation topics (taking the first six to avoid cherry-picking queries for which our method works best). The TREC topics are real queries, selected by editors from a search engine log. The editors considered the queries and guessed information needs. For each query, Table 1 shows the editor-inferred intent as well as the representative query for each of the top six intent clusters identified by our method. Topics TD1 (mining gold silver coal) and TD5 (pest control safety) are missing, as they are among 7 of the 50 TREC topic queries that did not have any valid reformulations in our log. Also, note that due to space constraints we only present the representative query from each cluster. Some of the information needs become clearer from the entire set of queries in the clusters found.

Our approach finds intents with a different emphasis than those inferred by the editors: For topic TD2, it appears that users are more interested in causes and prevention of delinquency than punishment measures used. For topic TD3, some users are particularly interested in maps, pictures and the related American Indian Sacagawea. In topic TD4, *wireless communications* often refers to cell phone and broadband communication, as well as history, but rarely legislation. For topic TD6, many users are interested in salary information, not specifically required by the editors.

Having established that our method produces results that are aligned with TREC topics but also provide plausible additional information about real user intents, it was employed during topic development for the TREC 2009 Web Track. Specifically, topic developers for the new Diversity Task used our results as their initial indication of user intent. This was augmented by further interaction with search engines.

Finally, for many other queries (including popular single-word examples), the information need clusters also appear realistic. For instance, the results for *columbia* identify clusters referring to the country, the clothing brand, information about the country in Spanish, the record company, and the related brand North Face.

5. REFERENCES

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 2008.

[2] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR*, 2007.

[3] J. Teevan, S. Dumais, and E. Horvitz. Characterizing the value of personalizing search. In *SIGIR*, 2007.