Collection and Analysis of Ground Truth Data for Query Intent

Maya Sappelli
m.sappelli@cs.ru.nl

Suzan Verberne
s.verberne@cs.ru.nl

Maarten van der Heijden
m.vanderheijden@cs.ru.nl

Max Hinne
mhinne@cs.ru.nl

Wessel Kraaij
w.kraaij@cs.ru.nl

ABSTRACT

Search engines try to support people in finding information and locating services on the web. What people are looking for depends on their underlying intent and is described by the query they enter in the search engine. These queries are often short and ambiguous. This paper describes the collection of ground truth data for query intent. Participants were asked to label their own search queries according to what they hoped to find with that query. The data can be used to investigate the reliability of external human assessors and to train automatic classification models.

General Terms

Data collection

Keywords

Query intent, Ground truth

1. INTRODUCTION

All popular web search engines are designed for keyword queries. Although entering a few keywords is less natural than phrasing a full question, it is an efficient way of finding information and users have become used to formulating concise queries. For example, in the query log data set "Accelerating Search in Academic Research Spring 2006 Data Asset" released by Microsoft, 70% of the 12 Million queries (which were entered into the MSN Live search engine) consist of one or two words.

It seems unlikely that a few keywords can precisely describe what information a user desires, which we refer to as *search intent* (also known as *query intent*).[1] The exact definition of this concept is still a topic of debate [4, 8]; but we can say that, roughly, search intent is what the user implicitly hoped to find using the submitted query. This is different from Broder's definition of information need [2] in that information need can be defined as the drive to formulate a series of queries. The intent of a specific query is often part of a bigger information need. This is the case when only part of the information need is expected to be satisfied by a query, for example because the information need is too big to be expressed in a single query. If the intent behind a query is known, a search engine can improve on retrieval results by adapting the presented results based on the more specific intent instead of the (underspecified) query [10].

Several studies have proposed classification schemes for query intent. After studying a large collection of AltaVista query logs, Broder [2] suggested that the intent of a query can be either informational, navigational or transactional. Later, many expansions and alternative schemes have been proposed, which we will summarize in Section 2. Ultimately, a search engine should be able to automatically classify a query according to such a scheme, so that the search intent of the user can be taken into account in the retrieval result. However, for the implementation of automatic classification models, training data is needed: a set of queries, labeled with their underlying intent. In previous studies, annotations of query intent labeling have been created by human assessors [1, 5]. However, in those studies, the assessors are not the searchers themselves.

We asked search engine users to label their own queries according to the underlying intent. This provides a ground truth that can be used (1) to investigate the reliability of external human assessors and (2) to train automatic classification models. Our data set is an important contribution to the field of query intent classification, since many studies rely on classifications by external assessors as gold standard classifications because they do not have access to classifications by the searchers themselves. We intend to make our data set publicly available.

Our paper is structured as follows. In Section 2 we describe intent classification schemes from the literature; Section 3 presents our classification scheme. In Section 4 we describe the user study that we conducted. The results of these experiments are presented in Section 5. Lastly, Section 6 concludes our work and described future research that we plan with this data.

2. INTENT CLASSIFICATION SCHEMES IN THE LITERATURE

The early paper by Broder [2] presents a taxonomy of web search, defining three categories for the intent behind queries: navigational (the user wants to reach a particular

[1] We use the terms 'query intent' and 'search intent' interchangeably.

DIR 2012 Ghent, Belgium

website), informational (the user wants to find a piece of information on the web) and transactional (the user wants to perform a web-mediated task). Rose and Levinson [7] refine the intent classification by Broder. They define three main categories for query intent: navigational, informational (which consists of five subcategories: directed, undirected, advice, locate, list) and resource (download, entertainment, interact, obtain).

More recently, it has been argued that search intent has more dimensions than the navigational – informational – transactional classification by Broder. Baeza-Yates et al. [1] present a classification scheme with two dimensions: topic (categories taken from the Open Directory Project[2] and goal (informational, non-informational or ambiguous). Sushmita et al. [9] distinguish between "query domain" (e.g. image, video, or map) and "query genre" (e.g. news, blog, or Wikipedia).

González-Caro et al. [5, 3] present multiple dimensions of user intent. Some of these are very general, such as Genre, Topic and Task (informational or non-informational). Others are better defined, such as Specificity and Authority sensitivity. The authors manually classify 5,000 queries according to all dimensions and give a good analysis of the agreement between judges and the correlation between dimensions.

Hinne et al. [6] propose an intent classification scheme with three dimensions: topic, action type and modus.

3. OUR INTENT CLASSIFICATION SCHEME

We introduce a multi-dimensional classification scheme of query intent that is inspired by and uses aspects from Broder [2], Baeza-Yates et al. [1], González-Caro et al. [5], Sushmita et al. [9] and Hinne et al. [6] (see Section 2). Our classification scheme consists of the following dimensions of search intent.

1. Topic: categorical, fixed set of categories (from ODP).
2. Action type: categorical, consisting of:
   - informational,
   - navigational,
   - transactional.
3. Modus: categorical, consisting of:
   - image,
   - video,
   - map,
   - text,
   - other.
4. source authority sensitivity: 4-point ordinal scale (high sensitivity: relevance depends more on authority of source).
5. location sensitivity: 4-point ordinal scale (high sensitivity: relevance depends more on location).
6. time sensitivity: 4-point ordinal scale (high sensitivity: relevance depends more on time).
7. specificity: 4-point ordinal scale (high specificity: very directed goal; low specificity: explorative goal).

The topic should give a general idea of what the query is about, for which we use the well-known Open Directory Project categories. *Action type* is the Broder categorisation; Modus is based on Sushmita et al. [9]. The ordinal dimensions are inspired by González-Caro et al. [5]. While many more dimensions can be imagined, we think that these capture an important portion of query intent.

[2]Open Directory Project (ODP): http://dmoz.org

4. DATA COLLECTION

In order to obtain query labelings from search engine users, we created a plugin for the Mozilla Firefox web browser. After installation by the user, the plugin (locally) logs all queries submitted to Google and other Google domains, such as Google images. We asked colleagues (all academic scientists and PhD students) to participate in our experiment. Participants were asked to occasionally annotate the queries they submitted in the last 48 hours, using a form showing our intent classification scheme. Table 1 shows the explanations of the intent dimensions that were given to the participants. To ensure participants understood what they were asked to do, we first presented three reference queries which were the same for all participants. Other queries were displayed in chronological order.

In order to avoid privacy issues, participants were allowed to skip any query they did not want to submit. When a participant clicked the 'submit' button, he was presented with a summary of his queries, from which queries could be excluded once again. After confirmation, the queries and annotations were sent to our server. For each submitted query, we stored the query itself, a timestamp of the moment the query was issued, a participant ID (a randomly initiated number used to group queries in sessions per participant) and the annotation labels.

5. RESULTS

In total, 11 participants enrolled in the experiment. Together, they annotated 605 queries with their query intent, of which 135 were annotated more than once[3]. On average, each person annotated 55 queries. Table 2 shows the number of queries per action type as annotated by the participants. Table 3 shows the number of queries per modus as annotated by the participants.

The three topic categories that were used most frequently in the annotated queries were *computer*, *science* and *recreation*. Table 4 shows the five most frequent categories and their frequencies. Figure 1 displays the labeling distributions, from low to high, for the following ordinal dimensions: source authority sensitivity, location sensitivity, temporal sensitivity and specificity of the queries.

Table 2: Number of queries per action type

Action type	# Queries[a]
Informational	546
Transactional	30
Navigational	70

[a]The sum of the queries may be higher than 605 since multiple action types could be selected per query

[3]It is important to notice that it is possible to represent different search intents with the same query, for example, when the same query is issued by the same person at different times.

Table 1: Explanation of the intent dimensions for the participants.

Dimension	Explanation
Topic	What is the general topic of your query?
Action type	Is the goal of your query: (a) to find information (informational), (b) to perform an online task such as buying, booking or filling in a form (transactional), (c) to navigate to a specific website (navigational)?
Modus	In which form would you like the intended result to have?
Source authority sensitivity	How important is it that the intended result of your query is trustworthy?
Location sensitivity	Are you looking for something in a specific geographic location?
Time sensitivity	Are you looking for something that is related to a specific moment in time?
Specificity	Are you looking for one specific fact (high specificity) or general information (low specificity)?

Table 3: Number of queries per modus

Modus	# Queries[a]
Image	33
Video	10
Map	27
Text	512
Other	6

[a]The sum of the queries may not add up to 605 since multiple modi could be selected per query and modus could be omitted

Table 4: The most frequently selected categories

Category	# Queries
Computers	250
Science	193
Recreation	87
Health	84
Reference	76

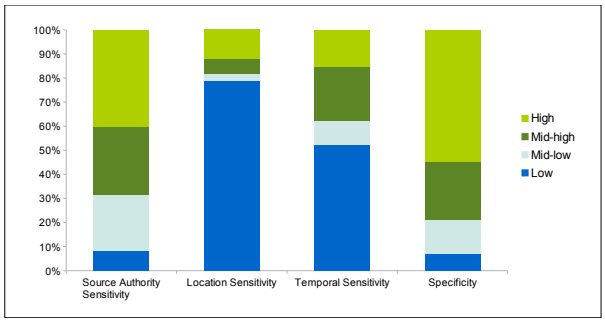

Figure 1: Distribution of source authority sensitivity, location sensitivity, temporal sensitivity and specificity, measured a scale from 1 (low) to 4 (high)

5.1 Analysis

In this section we take a closer look at the annotated queries. We first calculated correlations between the classification dimensions. The correlation between ordinal dimensions (source authority sensitivity, location sensitivity, temporal sensitivity and specificity) was estimated using Kendall's τ-b measure. We found a significant moderately strong positive correlation between source authority sensitivity and specificity ($\tau = 0.377$, $p < 0.0001$) as well as between location sensitivity and temporal sensitivity ($\tau = 0.421$, $p < 0.0001$). Additionally, we found weak positive correlations between location sensitivity and source authority sensitivity ($\tau = 0.135$, $p = 0.0002$) and between location sensitivity and specificity ($\tau = 0.106$, $p = 0.0042$).

Correlations for the categorical dimensions (topic, action type and modus) were determined using a chi-squared test. However, the outcome of this test is unreliable because there were too many zero-occurences in the crosstable of the dimensions. We do, however, see some interesting trends in the data:

- There tends to be a relation between the categories *news* and *sports* on the one hand and a high temporal sensitivity on the other hand: all *news* annotated queries (8 queries) and all but one *sports* annotated queries (15 queries) were annotated with a high temporal sensitivity.
- The category *science* and the combination of the categories *health* and *science* seem to be indicators of a high source authority sensitivity. Of the 183 queries annotated with *science*, 154 were annotated with a high source authority sensitivity, and all of the 79 queries annotated with the combination *health* and *science* were annotated with a high source authority sensitivity. The category *computer* was mostly annotated with a mid-high or mid-low source authority sensitivity (215 of 250 queries).
- There seems to be a relation between the modus of the query and the location sensitivity. Of the 26 queries that were annotated with the *map* modus, 23 were annotated with a high location sensitivity.

We also found that a number of aspects of query intent were not reflected by the textual content of the query:

- There were few query words that were specifically related to the modus or the action type of the query. For example, in the queries that were annotated with the *image* modus there were no occurrences of words such as "image" or "picture".
- Only 2 of the 90 queries that were annotated with a high temporal sensitivity contained a time-related query word.
- Of the 72 queries that included a location reference such as a city or a country, 36 were annotated with a high location sensitivity. In the remaining queries with lower location sensitivity, 11 location references occurred.

Finally, we found that 54% of the queries (331 queries) were annotated with a high specificity, 56% of which consisted of only one or two words (187 queries). We found a very weak negative correlation between query length and specificity ($\tau = -0.0968$, $p = 0.0047$). Figure 2 shows the

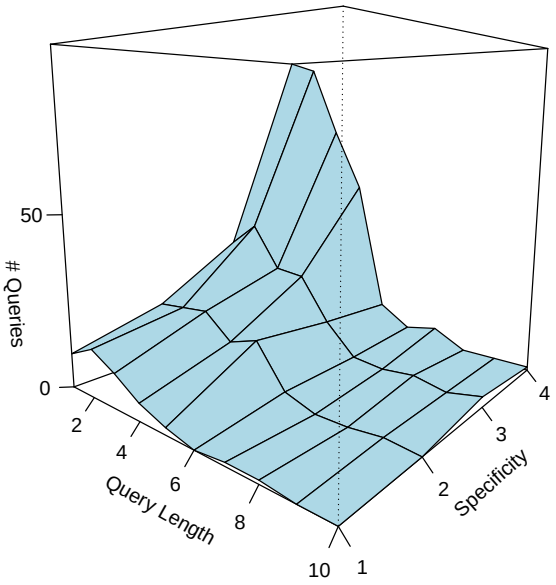

Figure 2: Effect of query length (number of words in query) on specificity as measured on a scale from 1 (low) to 4 (high)

relation between query length and the annotated specificity of the query intent.

6. CONCLUSION AND FUTURE WORK

In the research described in this paper we collected a set of queries that are labeled with their underlying intent. The queries were annotated by the searchers themselves to have a ground truth data set. This set can be used (1) to investigate the reliability of external human assessors and (2) to train automatic classification models.

The data shows that the textual content of the queries does not give many hints as to what annotation in terms of modus, action, source authority sensitivity, location sensitivity, time sensitivity and specificity can be expected. Moreover, query length does not predict the specificity of the query intent. This indicates that it might be difficult for an external human assessor that does not know the searcher or the context of the query to reliably determine what the searcher's query intent was. If it is difficult for a human assessor, it is even more difficult to assess query intent using automatic classification for a system without world knowledge or additional knowledge about the searcher.

We are interested in the differences between query intent annotation by external human assessors and the searcher's own annotations. Currently, the collected ground truth data is being labeled by external human annotators using the same annotation scheme as presented in this work. If it is possible for external annotators to reach consensus about search intent that matches the ground truth search intent, then automated classification may be possible as well. The main contribution of this work is a resource that helps to validate the assumptions that are made by state of the art methodologies for qualitative and quantitative studies of query intent. Indeed, current state of the art studies are based on external assessors, assuming these are sufficiently close to the original intent of the searcher. Our work enables validation of this assumption.

We will use the knowledge about differences and commonalities between external assessors and the searcher for improving automatic intent classification. We expect that knowledge about the user's expertise, search history, and other computer behaviour to be the most important factors to be able to understand the intent of a query. Therefore, in future work we will address the use of the user's search history and current computer activies to assess a searcher's intents. Monitoring computer activities may provide context about the query and disambiguate its intent.

7. REFERENCES

[1] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The Intention Behind Web Queries. In F. Crestani, P. Ferragina, and M. Sanderson, editors, *String Processing and Information Retrieval*, LNCS 4209, pages 98–109, Berlin Heidelberg, 2006. Springer-Verlag.

[2] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[3] L. Calderón-Benavides, C. González-Caro, and R. Baeza-Yates. Towards a Deeper Understanding of the User's Query Intent. In *Workshop on Query Representation and Understanding, SIGIR 2010*, pages 21–24, 2010.

[4] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179:1822–1843, 2009.

[5] C. González-Caro, L. Calderón-Benavides, R. Baeza-Yates, L. Tansini, and D. Dubhashi. Web Queries: the Tip of the Iceberg of the User's Intent. In *Workshop on User Modeling for Web Applications, WSDM 2011*, 2011.

[6] M. Hinne, M. van der Heijden, S. Verberne, and W. Kraaij. A multi-dimensional model for search intent. In *Proceedings of the Dutch-Belgium Information Retrieval workshop (DIR 2011)*, pages 20–24, 2011.

[7] D. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web (WWW 2004)*, pages 13–19. ACM, 2004.

[8] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174, 2010.

[9] S. Sushmita, B. Piwowarski, and M. Lalmas. Dynamics of genre and domain intents. *Information Retrieval Technology*, pages 399–409, 2010.

[10] R. White, P. Bennett, and S. Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1009–1018. ACM, 2010.