# Reliability and Validity of Query Intent Assessments

## Compressed version of paper accepted for publication in JASIST

Suzan Verberne
s.verberne@cs.ru.nl

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

Max Hinne
mhinne@cs.ru.nl

Maya Sappelli
m.sappelli@cs.ru.nl

Saskia Koldijk
saskia.koldijk@tno.nl

Eduard Hoenkamp
hoenkamp@acm.org

Wessel Kraaij
w.kraaij@cs.ru.nl

## Keywords

Query intent classification, User studies, Data collection, Validation

## 1. INTRODUCTION

The quality of a search engine critically depends on the ability to present results that are an adequate response to the user's query and intent. If the intent (or the most likely intent) behind a query is known, a search engine can improve retrieval results by adapting the presented results to the more specific intent instead of the — underspecified — query [6]. Several studies have proposed classification schemes for query intent. Broder [3] suggested that the intent of a query can be either informational, navigational or transactional. He estimated percentages for each of the categories by presenting Altavista users a brief questionnaire about the purpose of their search after submitting their query. After manual classification of 1,000 queries he warned that "inferring the user intent from the query is at best an inexact science, but usually a wild guess." Later, many expansions and alternative schemes have been proposed, and more dimensions were added.

In many existing intent recognition studies, training and test data for automatic intent recognition have been created in the form of annotations by external assessors who are not the searchers themselves [2, 1, 4]. Post-hoc intent annotation by external assessors is not ideal; nevertheless, intent annotations from external judges are widely used in the community for evaluation or training purposes. Therefore it is important for the field to get a better understanding of the quality of this process as an approximation for first-hand annotation by searchers themselves. Some annotation studies have investigated the *reliability* of query intent annotations by measuring the agreement between two external assessors on the same query set [1, 4]. What these studies do not measure, is the *validity* of the judgments.

In this paper, we aim to measure the validity of query intent assessments, i.e. how well an external assessor can estimate the underlying intent of a searcher's query. We use a classification scheme to describe search intent.

*DIR 2013,* April 26, 2013, Delft, The Netherlands.
.

## 2. OUR INTENT CLASSIFICATION SCHEME

We introduce a multi-dimensional classification scheme of query intent that is inspired by and uses aspects from [3], [2], [4] and [5]. Our classification scheme consists of the following dimensions of search intent.

1. Topic: categorical, fixed set of categories from the well-known Open Directory Project (ODP), giving a general idea of what the query is about.
2. Action type: categorical, consisting of: *informational*, *navigational* and *transactional*. This is the categorisation by Broder.
3. Modus: categorical, consisting of: *image*, *video*, *map*, *text* and *other*. This dimension is based on [5].
4. *source authority sensitivity*: 4-point ordinal scale (high sensitivity: relevance strongly depends on authority of source).
5. *spatial sensitivity*: 4-point ordinal scale (high sensitivity: relevance strongly depends on location).
6. *time sensitivity*: 4-point ordinal scale (high sensitivity: relevance strongly depends on time/date).
7. *specificity*: 4-point ordinal scale (high specificity: very specific results desired; low specificity: explorative goal).

## 3. EXPERIMENTS

In order to obtain labeled queries 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. We asked colleagues (all academic scientists and PhD students) to participate in our experiment. Participants were asked to occasionally (at a self-chosen moment) annotate the queries they submitted in the last 48 hours, using a form that presented our intent classification scheme. To guarantee that no sensitive information was involuntarily submitted, participants were allowed to skip any query they did not want to submit.

In total, 11 participants enrolled in the experiment. Together, they annotated 605 queries with their query intent, of which 135 duplicates. On average, each searcher annotated 55 queries (standard deviation=73). The three topic

Table 1: **Reliability and validity of query intent assessments in terms of Cohen's Kappa, averaged over the assessor pairs. Boldface indicates moderate agreement ($\kappa >= 0.4$) or higher.**

| Dimension | Reliability (stdev) | Validity (stdev) |
|---|---|---|
| Topic | **0.56** (0.19) | **0.42** (0.16) |
| Action type | 0.29 (0.20) | 0.09 (0.08) |
| Modus | **0.41** (0.14) | 0.22 (0.10) |
| Source authority sensitivity | 0.05 (0.05) | 0.10 (0.03) |
| Time sensitivity | **0.48** (0.08) | 0.14 (0.04) |
| Spatial sensitivity | **0.69** (0.07) | **0.41** (0.04) |
| Specificity | 0.26 (0.10) | 0.05 (0.09) |

categories that were used most frequently in the set of annotated queries were *computer*, *science* and *recreation*.

To obtain labels from external assessors we used the same form as was used by the participants. Four of the authors acted as external assessors; all queries were assessed by at least two assessors.

## 4. RESULTS

In order to answer the question "How *reliable* is our intent classification scheme as an instrument for measuring search intent?", we calculated the interobserver reliability as the agreement between the external assessors using Cohen's $\kappa$. The middle column of Table 1 shows the average agreement over the assessor pairs for each dimension. For only one of the seven dimensions from our classification scheme) substantial agreement (0.6 or higher) was reached. For four of the seven, at least moderate agreement (0.4 or higher) was reached: least moderately reliable query intent classification is possible for the dimensions topic, modus, time sensitivity and spatial sensitivity.

In order to answer the question, "How *valid* are the intent classifications by external assessors?", we compared the intent classifications by the external assessors to the intent classifications by the searchers themselves. We calculated $\kappa$-scores per dimension for each assessor–searcher pair. The rightmost column of Table 1 shows the average agreement over the assessor–searcher pairs. The table shows that moderately valid query intent classification is possible on two of the seven dimensions from our classification scheme: topic and spatial sensitivity. The difference between the inter–assessor agreement and the assessor–searcher agreement was significant on all dimensions.

Our experiments suggest that classification of queries into Topic categories can be done reliably, even though we had 17 different topics to choose from. This is good news for a future implementation of automatic query classification because topic plays an important role in query disambiguation and personalisation. The second reliable dimension, Spatial sensitivity, is an important dimension for local search: every web search takes place at a physical location, and there are types of queries for which this location is relevant (e.g. the search for restaurants or events). The finding that external assessors can reach a moderate agreement with the searcher on this dimension shows the feasibility of recognizing that a query is sensitive to location. The search engine can respond by promoting search results that match with the location.

For the implementation of intent classification in a search engine, training data is needed: The features are the query terms (the textual content of the query) and the labels are the values for the dimensions in the classification scheme. Analysis of the queries shows that for many intent dimensions, there is no direct connection between words in the query and the intent of the query. For example, in the 33 queries that were annotated by the searcher with the *image* modus (e.g. "photosynthesis"; "coen swijnenberg") there were no occurrences of words such as 'image' or 'picture', and only 2 of the 90 queries that were annotated with a high temporal sensitivity contained a time-related query word. This means that for automatic classification, it is difficult to generalize over queries. However, the most likely intent can still be learned for individual queries by following the diversification approach in the ranking of the search results: The engine can learn the probability of intents for specific queries by counting clicks on different types of results. This approach requires a huge amount of clicks to be recorded (which is possible for large search engines such as Google) and the long tail of low-frequency queries will not be served.

## 5. CONCLUSIONS

We found that four of the seven dimensions in our classification scheme could be annotated moderately reliably ($\kappa > 0.4$): topic, modus, time sensitivity and spatial sensitivity. An important finding is that queries could not reliably be classified according to the dimension 'action type', which is the original Broder classification. Of the four reliable dimensions, only the annotations on the topic and spatial sensitivity dimensions were valid ($\kappa > 0.4$) when compared to the searcher's annotations. This shows that the agreement between external assessors is not a good estimator of the validity of the intent classifications.

In conclusion, we showed that Broder was correct with his warning that "inferring the user intent from the query is at best an inexact science, but usually a wild guess". Therefore, we encourage the research community to consider - where possible - using query intent classifications by the searchers themselves as test data.

## 6. REFERENCES

[1] A. Ashkan, C. Clarke, E. Agichtein, and Q. Guo. Classifying and characterizing query intent. *Advances in Information Retrieval*, pages 578–586, 2009.

[2] 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.

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

[4] 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.

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

[6] 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.