Noname manuscript No.
(will be inserted by the editor)

Reliability and Validity of Query Intent Assessments

Suzan Verberne · Maarten van der Heijden · Max Hinne · Maya Sappelli · Saskia Koldijk · Eduard Hoenkamp · Wessel Kraaij

the date of receipt and acceptance should be inserted later

Abstract 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. Automatic intent recognition is a challenging problem because queries are often short or underspecified. In most intent recognition studies, annotations of query intent are created post-hoc by external assessors who are not the searchers themselves. It is important for the field to get a better understanding of the quality of this process as an approximation for determining the searcher's actual intent.

Query intent annotation quality has different aspects. Some annotation studies have investigated the *reliability* of the query intent annotation process by measuring the inter-assessor agreement. However, these studies did not measure the *validity* of the judgments, i.e. to what extent the annotations match the searcher's actual intent. In this study, we asked both the searchers themselves and external assessors to classify queries using the same intent classification scheme.

We show that of the seven dimensions in our intent classification scheme, four can *reliably* be used for query annotation. Of these four, only the annotations on the topic and spatial sensitivity dimension are *valid* when compared to the searcher's annotations. The difference between the inter-assessor agreement and the assessor-searcher agreement was significant on all dimensions, showing that the agreement between external assessors is not a good estimator of the validity of the intent classifications. Therefore, we encourage the research community to consider using query intent classifications by the searchers themselves as 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.

Institute for Computing and Information Sciences, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
E-mail: s.verberne@cs.ru.nl

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.

The proficiency of search engine users notwithstanding, it seems unlikely that a few keywords can precisely describe what information a user desires, i.e. what a user's *search intent* (also known as *query intent*[1]) is. The exact definition of this concept is a topic of debate (Gayo-Avello, 2009; Silvestri, 2010); roughly, search intent is what the user implicitly hopes to find when he issues a query. This is related to the broader concept of *information need*. Users are often unable to precisely formulate their need (Belkin et al., 1982) and they might need to issue multiple queries to satisfy their information need. For example, when preparing a holiday in the south of France, a user may issue a series of queries with geographical locations. Each query has its own intent. With the query 'flight Nijmegen Avignon', he could have the intent of booking a flight; with the query 'Avignon city centre' he could be interested in a map of the city and with 'events Avignon' he could wish to find recent information about coming events in the city. The information need behind the whole series of queries could be 'planning a holiday in the south of France'.

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 (White et al., 2010). In the case of multiple possible intents, the search engine can apply a diversification strategy to the result ranking, mixing results for the different possible intents (Santos et al. (2011); Sakai (2012)). Several studies have proposed classification schemes for query intent. Broder (2002) suggested that the intent of a query can be either informational, navigational or transactional. Later, many expansions and alternative schemes have been proposed, and more dimensions were added. In Section 2 we summarize the variety of intent classification schemes that have been proposed to date and in Section 3 we present a new, multi-dimensional classification scheme.

A better match between the query intent and the search results increases user satisfaction. Ultimately, a search engine should be able to automatically classify a query according to its most likely intent, so that the search intent of the user can be taken into account in the retrieval result. In existing intent recognition studies, training data for automatic intent recognition have been created in the form of annotations by external assessors who are not the searchers themselves (Baeza-Yates et al., 2006; Ashkan et al., 2009; González-Caro et al., 2011). Post-hoc intent annotation by external assessors is not ideal; for the TREC benchmark tasks it is the preferred practice that relevance assessments are created by the same person who formulated the query. Nevertheless, for practical reasons, intent annotations obtained from external judges are widely used in the community for evaluation or training purposes, for example in the TREC Web track. 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 (Ashkan et al., 2009; González-Caro et al., 2011). What these studies do not measure, is the *validity* of the judgments.

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

The distinction between reliability and validity of judgments is an important one. About a century ago, it was common to assess a person's intelligence by measuring size and form of the skull. These measurements were extremely reliable: someone who measured a second time would get the same answer (the 'test-retest reliability') and when more people made the same measurement each would get the same answer (the 'inter-observer reliability'). Yet, as Binet (1907) showed at the time, these measurements had little to do with the intelligence it was supposed to measure. Then Binet proposed a new test (the Stanford–Binet test) that until today we consider a *valid* measurement, i.e. one that actually measures what it is supposed to measure (intelligence). Note that for a measurement to be valid, it must at least be reliable.

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. To do so, we need an instrument (just as we need a measuring tape for measuring the size of the head). This is problematic because 'intent' is an abstract concept (as is 'intelligence') that is not made explicit during the process of searching. Therefore, we use a classification scheme to describe search intent. The scheme we use is a combination of several classification schemes available in the literature. Our research questions are:

1. How *reliable* is our intent classification scheme as an instrument for measuring search intent?
2. How *valid* are the intent classifications by external assessors?

In order to measure the reliability of the classification scheme (question 1), we collected a set of queries and asked multiple external assessors to classify them according to the scheme. We use the agreement between their assessments as measure of reliability. To measure the validity of the external assessments (question 2), we assume that the searcher himself knows better than an external assessor what he intended to find when issuing his query. Therefore, we have asked the searchers themselves to classify their queries according to the underlying intent, using the same classification scheme.[2] We then approximate the validity of the query intent assessments as the agreement between the external assessors and the searchers themselves.

Our work is an important contribution to the field of query intent classification, because (1) the reliability and validity of intent annotations are indicators for the suitability of these annotations as training data for automatic intent annotation and (2) our multi-dimensional intent classification scheme reveals which aspects of search intent are expected to be more appropriate for automatic classification and which are expected to be less appropriate. We intend to make our data set publicly available.

2 Related Work

The current paper is embedded in a large body of previous research on the intent of search engine users. In all these studies, it is assumed that the user's

[2] Note that we do not aim at identifying all possible intents for a query, but at identifying the intent that was meant by the searcher himself.

query is a textual representation of some underlying search intent or information need. Belkin et al. (1982) consider the information need as an 'anomalous state of knowledge' (ASK). They suppose that the user is unable to precisely formulate his information need. The better a search engine is capable of recognizing the query's underlying intent, the better it can satisfy the user's information need. Several different definitions of search intent and information need are proposed in the literature. Regardless of the specific interpretations, all studies that try to measure either search intent or information need to deal with the issues of reliability and validity as described above.

In the Information Retrieval community, discovering the user's search intent is rephrased as the challenge of tracing the user's information need, i.e. the problem of query understanding (Croft, 2010). The problem is approached from the search engine's point of view: What can a search system learn from the user's queries and how can the search results be improved with the use of this knowledge? Important sources of information for research into search behavior are query log data sets (often referred to as click log data, if they not only contain the queries themselves but also clickthrough information).

In the Information Seeking community, information need is generally modeled as an aspect of user behavior or user communication (Wilson, 2006). Dervin and Nilan (1986) were the first to suggest a shift from the system point-of-view to the user point-of-view: they define information need in terms of the user and suggest to use interviews with searchers for system evaluation. In line with this, Martzoukou (2005) states that a model of information seeking should include characteristics of the user and his/her context in order to fully understand the search behavior. The context of the user can be a good information source for enriching a query with its intent. A better understanding of user abilities and expectations can lead to changes in underlying system mechanics or human–computer interaction (Buchanan et al., 2005).

2.1 Approaches to intent recognition

In the literature, there are three different approaches to intent recognition: classification, verbalization and formalization. The most commonly used approach is classification: to classify the query according to an intent classification scheme — this is the approach that we took. Two alternatives are either to verbalize the query as a longer, more informative question (Law et al., 2009), or to transform the natural language query into a formal query, restricted by a predefined ontology (Zhou et al., 2007; Tran et al., 2007). Law et al. (2009) regard query intent as the (longer) question underlying a short query. In order to find the relations between short queries and the longer underlying questions, they ask users to formulate queries that describe verbalized information needs. Tran et al. (2007) translate keyword queries into descriptive logic queries. They do this by mapping query terms to ontology elements and then expanding the query with neighboring ontology elements, assuming that the underlying intent is broader than the query terms only. Zhou et al. (2007) translate keyword queries into formal logic queries. First they map query terms to ontology entries and then they generate from the ontology a ranked list of formal queries that represent aspects of the keyword

query. In Section 6.3, we come back to these methods and compare them to our approach.

In the following two subsections, we will discuss two types of studies into search intent: (1) studies that aim to develop classification schemes for search intent, and evaluate them through manual or automatic classification and (2) studies that aim to infer search intent from search behavior. In Section 2.4, we summarize the contributions of our work compared to the existing literature.

2.2 Intent classification schemes and the reliability of annotations

Several attempts have been made to make the abstract concept of information need more tangible. These attempts focus at capturing the intent behind a query with the use of intent classification schemes (sometimes referred to as search taxonomies). Below, we give an overview of intent classification schemes from the literature that form the basis for our own intent classification scheme. Most classification schemes have been used for the manual or automatic classification of queries, but most authors are vague with respect to the evaluation of the classification task: only a few measure inter-observer *reliability* (but often without good statistical analysis of the results), and none of them measure the *validity* of the classifications by comparing them to annotations by the searchers themselves.

The earliest work on intent classification is the paper by Broder (2002), presenting the first taxonomy of web search. Broder defines three categories for the intent behind queries: navigational (the user wants to reach a particular 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). The distinction between informational and transactional is important for optimizing advertizement placement and associated clickthrough ratios. Broder (2002) estimates percentages for each of the categories by presenting Altavista users a brief questionnaire about the purpose of their search after submitting their query. He also performs a manual classification of 1,000 queries from an Altavista query log but he warns that "Since inferring the user intent from the query is at best an inexact science, but usually a wild guess, the data obtained from log analysis is very 'soft'."

Broder's intent classification scheme has been refined by Rose and Levinson (2004). 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. The classification scheme by Baeza-Yates et al. (2006) consists of two dimensions: topic (categories taken from the Open Directory Project[3]) and goal (informational, non-informational or ambiguous). They aim at identification of a user's interest based on query logs. They perform a manual classification of 6,000 queries logged by a Chilean web search engine but do not report inter-annotator agreement on the classifications, which makes the reliability of their classifications unclear.

Query intent classification is sometimes considered an important task for web advertisers. For that purpose, the intent classification scheme by Ashkan et al.

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

(2009) has two dimensions: commercial vs. non-commercial and informational vs. navigational. Human judgements were used as reference data in this study. Although absolute agreement scores above 80% are reported, these figures are different to judge since chance agreement has not been taken into account (which is commonly done using Cohen's κ score (Cohen, 1960)).

A number of taxonomies for query *topic* classification are described by Brenes et al. (2009). They state that there is broad consensus on the taxonomy for *intent* classification, referring to the intent classification scheme that was proposed by Broder and refined by Rose and Levinson. For the purpose of evaluating automatic classification methods, Brenes et al. (2009) asked 10 different annotators to classify 6,624 queries according to the Broder-scheme; every query was classified by two different annotators. They do not compute agreement scores but informally report that "the level of agreement between labelers was pretty high" (p. 4).

In more recent work, several aspects of query intent are defined in addition to the classification by Broder. For example, Sushmita et al. (2010) introduce several interesting aspects of query intent in addition to the classification by Broder: they distinguish between 'query domain' (e.g. image, video, or map) and 'query genre' (e.g. news, blog, or Wikipedia). The same study reports experiments on query intent classification, but their evaluation methodology remains unclear. In the work by González-Caro et al. (2011) and Calderón-Benavides et al. (2010), multiple dimensions of search intent are presented. Some of these are very general, such as Genre (with the values news, business, reference and community), Topic and Task (informational or non-informational). Others are defined more narrowly, such as Specificity and Authority sensitivity. 5,000 queries were annotated according to all dimensions. 10% of the queries was annotated by two judges and the authors report inter-annotation agreement in terms of Cohen's κ. They find that the agreement varies largely among the dimensions, from $\kappa = 0.33$ for specificity to $\kappa = 0.98$ for time sensitivity.

The paper by Lewandowski et al. (2012) has the same aim as our work: to measure the reliability of query intent assessments, in order to find out whether manual intent annotations are sufficiently reliable to be used as test data for automatic approaches. They use a Broder-like classification scheme, distinguishing informational, navigational, transactional, commercial (the user might be interested in commercial offerings) and local (the user is searching for information near his current geographic position) intent. In a crowdsourcing experiment, a large sample of 50,000 queries from the German T-mobile search portal were classified by human assessors. The assessors were allowed to assign more than one intent class to a query. The class 'informational' was not included; instead, every query that was not put in any other class automatically obtained an informational intent. The results of the experiment showed that users often do not agree on the intent that should be assigned to a query. After the crowdsourcing experiment, Lewandowski et al. (2012) performed a user study in which they asked users of the T-mobile search portal to fill in a survey (similar to Broder (2002)Broder's (2002) original intent survey) if they issued one of the queries from the crowdsourcing experiment. 549 queries were collected with this study. The results revealed that only 11% of the queries were annotated with the same query type by the searcher and the external assessor. The evaluation also showed that the participants in most cases did not agree about the query type even though they searched with the exact same query. One of the conclusions is that "searchers do not consistently

know what they are looking for when they begin a search and want the search engine to give them inspiration".

2.3 Inferring search intent from search behavior

There is not much previous work that addresses intent classification from the point of view of the searcher himself. In order to collect user data, one could conduct regular interviews with the searchers, use a read-aloud setting in a lab (e.g. Terai et al., 2008), ask users to fill in questionnaires (e.g. Broder, 2002), ask users to label their own queries with respect to some classification scheme, or infer the intent of individual queries by recording clicks on pages in a diversified result list.

A few studies have addressed intent classification by observing the user and his search behavior. Terai et al. (2008) focus on user behavior when performing search tasks with different intents. They use a number of experimental methods (eye gazing, browser logging, read-aloud) and their results show interesting differences in click and view patterns between transactional and informational search intents. White et al. (2010) collect query logs using a browser plugin that saves browsing history, from which they extract queries and click data. They use the activities that the user performed before submitting a query for predicting the intent of the query.

Query-specific intent could be learned from online search behavior. Search engines can present results that answer multiple possible intents (e.g. both images and text, or both location-dependent and location-independent results) (Santos et al., 2011) and record clicks on the result types. This way, the search engine can learn the probability distribution of intents for one specific query.

2.4 Our contributions

As opposed to the literature discussed in Section 2.2, we do not only measure the *reliability* but also the *validity* of query intent annotations, using a measure that takes into account the chance agreement on each dimension (Cohen's κ). To that end, we collect intent classifications from searchers and from external assessors for the same queries. As opposed to the work by Terai et al. (2008), we do not provide search tasks to subjects, but ask them to annotate the Google queries that they formulated in their natural daily work environment. In addition to the work by White et al. (2010), we not only collect queries but also explicit intent annotations according to a multi-dimensional intent classification scheme. We measure reliability as the inter-observer agreement between external assessors, and validity as the agreement between the external assessors and the searchers themselves.

3 Our intent classification scheme

We introduce a multi-dimensional classification scheme of query intent that is inspired by and uses aspects from Broder (2002), Baeza-Yates et al. (2006), González-Caro et al. (2011) and Sushmita et al. (2010). We tried to compile a set of dimensions that together can describe most of the aspects of search intent that are

relevant for improving search results. 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: (i) *informational*, (ii) *navigational* and (iii) *transactional*. This is the categorisation by Broder (2002).
3. Modus: categorical, consisting of: (i) *image*, (ii) *video*, (iii) *map*, (iv) *text* and (v) *other*. This dimension is based on Sushmita et al. (2010).
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).

While many more dimensions can be imagined, we think that these seven capture an important portion of query intent. The rationale behind this set of dimensions is that each of them has a potential value for the adaptation of the search results to the most likely intent behind the query. 'Topic' mainly has a disambiguation function; an ambiguous term such as *java* has a different meaning in the computer domain than in the recreation domain. 'Action Type' and 'Modus' determine the mix of result types that are shown: is the user aiming to buy or download something or is he just looking for information? Is it relevant for the user to view results on a map? Is he expecting video or image results? The ordinal dimensions can influence the ranking of the results. For a query that has a high source authority expectation for the results, user-generated content might be suppressed from the result list. For a query with high spatial sensitivity, links to events/places that are close to the searcher's physical location might be ranked first. For a query with high time sensitivity, pages that match the time slot specified by the user (if there is any), or pages that are about contemporal or near-future events, might be ranked first (e.g. a user searching for an event is likely to be interested in the next edition of that event). A query that expects highly specific results, general pages might be suppressed, such as introductory pages about Java for a Java programmer.

In pilot experiments, we tried out variants of the multi-dimensional scheme and removed the values that were judged as too difficult to interpret by the annotators (e.g. additional values for the 'modus' dimension). The reliability of this classification scheme can be measured per individual dimension, so that it may be further refined by removing unreliable dimensions.

4 Experiments

4.1 Collecting searchers' annotations

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

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	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?
Spatial sensitivity	Are you looking for something at 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)?

locally logs all queries submitted to Google and other Google domains, such as Google Images by selecting URLs that contain the strings *google* and *q=*.[4] 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. Table 1 shows the explanations of the intent dimensions that were given to the participants. Queries were displayed in chronological order. Just like in the work by Lewandowski et al. (2012), participants were allowed to select more than one value in a dimension.

To guarantee that no sensitive information was involuntarily submitted, 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 annotated 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 generated number used to group queries in sessions per participant) and the annotation labels. A screenshot of the query annotation environment is shown in Figure 1.

Before giving statistics on the annotations, we note that we did not intend to collect data that are representative for all queries issued by all search engine users. In fact, representativeness is impossible to reach because we can never create a subject pool that reflects the population of search engine users. Instead, we chose to limit our subject pool to colleagues, all computer scientists. This made it easier to control the experiment: we know beforehand that the searchers and assessors have the same field of expertise. One effect is that the topics covered by the submitted queries are expected to be biased towards computers and science. A second effect is that our expert assessors are probably better in determining the intent behind a query in the computer science domain than assessors without any background knowledge on this topic.

[4] The URL requirement *q=* ensures that only searching — and not browsing — is included in our data set.

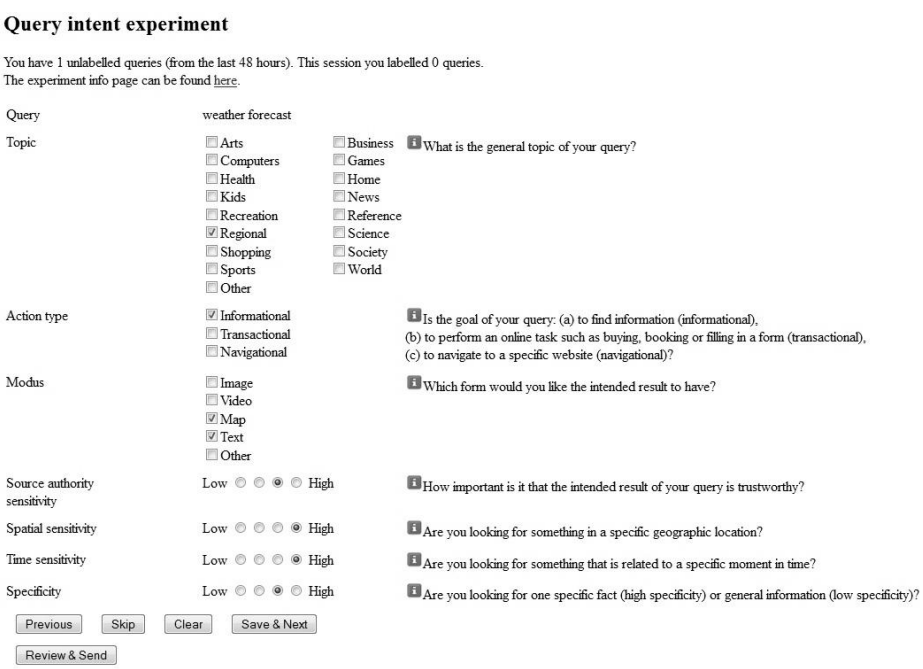

Fig. 1 The query annotation form.

Table 2 Number of queries per topic (most frequent), modus and action type. Sums may be higher than the total number of queries since multiple options could be selected per query.

Topic	# Queries	Modus	# Queries	Action type	# Queries
Computers	250	Text	512	Informational	546
Science	193	Image	33	Navigational	70
Recreation	87	Map	27	Transactional	30
Health	84	Video	10		
Reference	76	Other	6		

4.2 General information on the collected data

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 (see Section 6.2). On average, each searcher annotated 55 queries (standard deviation 73). Table 2 shows the number of queries per topic, modus and action type as annotated by the participants. The three topic categories that were used most frequently in the set of annotated queries were *computer*, *science* and *recreation*. Figure 2 displays the labeling distributions, from low to high, for the ordinal dimensions: source authority sensitivity, spatial sensitivity, temporal sensitivity and specificity of the queries.

4.3 Collecting labels from external assessors

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

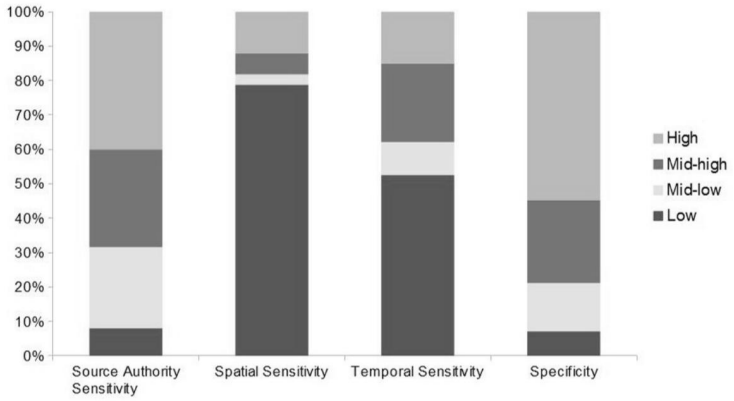

Fig. 2 Distribution of source authority sensitivity, spatial sensitivity, temporal sensitivity and specificity, measured a scale from 1 (low) to 4 (high)

assessed by at least two assessors. Note that although authors acted as assessors they did not see any of the data before making the assessments. The queries were presented in the same order as issued by the participants, which preserves session information between the self assessment and the external assessment. Besides the ordering information no explicit context was provided.

5 Results

In this section we address the research questions that we introduced in Section 1 on the reliability (Section 5.1) and the validity (Section 5.2) of query intent assessments. In Section 5.3, we compare the results for reliability and validity, and in Section 5.4, we investigate for which queries the validity is high and what factors play a role in the differences between validity scores for individual queries.

5.1 The reliability of query intent assessments

In order to answer research question 1, "How *reliable* is our intent classification scheme as an instrument for measuring search intent?", we calculated the agreement between the external assessors using Cohen's Kappa (Cohen, 1960). We will refer to this comparison as the 'κ-inter-assessor agreement' (κ-IAA). The rationale behind this way of measuring reliability is that of the *inter-observer reliability* (Landis and Koch (1977)); results may be seen as reliable if different assessors assign the same classification to given queries.

Cohen's Kappa takes into account the probability that two assessors assign the same labels by chance (in order to prevent over-estimating agreement in the case of very skewed judgments, such as the bias towards the topics computers and

Table 3 Weight matrix for ordinal scales.

	Very low	Low	High	Very high
Very low	0	1	2	3
Low	1	0	1	2
High	2	1	0	1
Very high	3	2	1	0

science in our data set). For calculating the chance agreement, we use the data from the searchers themselves. We made a distinction between the dimensions with categorical values (the dimensions 'topic', 'action' and 'modus') and those with ordinal scales (the dimensions 'source authority sensitivity', 'location sensitivity', 'time sensitivity' and 'specificity'). For the categorical dimensions, each possible value of the dimension (e.g. image, video, map, text and other for the dimension Modus) was considered a binary variable of its own that was either present or absent in the intent classification of a query. Agreement was then calculated for each of these variables separately:

$$\kappa_{\mathrm{N}} = \frac{p_a - p_c}{1 - p_c} \ , \tag{1}$$

in which p_a and p_c represent the absolute agreement and the agreement by chance, respectively, for that value in a categorical dimension. Note that these scores are calculated over all queries to which both assessors assigned a labeling.

For the ordinal dimensions, we applied Weighted Cohen's Kappa (Cohen, 1968). This requires a weight matrix W that indicates how severe a given disagreement is. We chose to let the values of W be given by the distance on the ordinal scale between the two choices, as shown in Table 3, such that the difference between very low and high is more severe than between very low and low.

With this weight matrix, the agreement score per variable is calculated as:

$$\kappa_{\mathrm{W}} = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} a_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} c_{ij}} \ , \tag{2}$$

where k equals the number of possible (ordinal) values and a_{ij} and c_{ij} are the absolute agreement and the expected agreement for choices i by the first annotator and j by the second annotator.

Using Equations (1) and (2) we calculated the agreement between all pairs of external assessors for all dimensions of our intent classification scheme. Table 4 shows the average agreement over the assessor pairs for each dimension. If we want to answer the question: "for which of the dimensions is reliable query intent classification possible?", we have to set a threshold on the κ-scores. According to Landis and Koch (1977), a κ between 0 and 0.20 can be characterized as slight agreement, between 0.21 and 0.40 as fair, between 0.41 and 0.60 as moderate, between 0.61 and 0.80 as substantial, and above 0.80 as almost perfect agreement. For only one of the seven dimensions from our classification scheme (spatial sensitivity) substantial agreement was reached. For four of the seven, at least moderate agreement was reached, so at least moderately reliable query intent classification is possible for the dimensions topic, modus, time sensitivity and spatial sensitivity.

Table 4 Reliability of query intent assessments in terms of Cohen's Kappa for inter-assessor agreement (κ-IAA) on each of the intent classification dimensions, averaged over the assessor pairs. Boldface indicates moderate agreement ($\kappa >= 0.4$) or higher.

Dimension	Mean κ-IAA (Stdev)
Topic	**0.56** (0.19)
Action type	0.29 (0.20)
Modus	**0.41** (0.14)
Source authority sensitivity	0.05 (0.05)
Time sensitivity	**0.48** (0.08)
Spatial sensitivity	**0.69** (0.07)
Specificity	0.26 (0.10)

Table 5 Validity of query intent assessments in terms of Cohen's Kappa for assessor–searcher agreement (κ-ASA) on each of the intent classification dimensions, averaged over the searcher–assessor pairs. Boldface indicates moderate agreement ($\kappa >= 0.4$) or higher.

Dimension	Mean κ-ASA (Stdev)
Topic	**0.42** (0.16)
Action type	0.09 (0.08)
Modus	0.22 (0.10)
Source Authority sensitivity	0.10 (0.03)
Time sensitivity	0.14 (0.04)
Spatial sensitivity	**0.41** (0.04)
Specificity	0.05 (0.09)

5.2 The validity of query intent assessments

In order to answer research question 2, "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 refer to this comparison as the 'κ-assessor–searcher agreement', (κ-ASA). This time we reason that a searcher has a better indication of his own search intent than an external assessor. Although the searcher may not be able to fully express his search intent, intuitively his classification should be closer to his actual intent than if classified by someone else. Thus, if external assessors agree with the searchers themselves, the classification scheme is valid, at least to the extent searchers are able to classify their own intent. Following the same approach as in the previous section, we calculated κ-scores per dimension for each assessor–searcher pair. Table 5 shows the average agreement over the assessor–searcher pairs for each dimension. We again use moderate agreement ($\kappa > 0.4$) as criterion for validity. The table shows that on the basis of this criterion, valid query intent classification is possible on two of the seven dimensions from our classification scheme: topic and spatial sensitivity.

5.3 Comparing inter-assessor agreement and assessor–searcher agreement

In Sections 5.1 and 5.2 we found that it is possible to reach moderate agreement between external assessors on four of the seven dimensions of our intent classification scheme, but that moderate agreement with the searcher was only possible on two of these dimensions.

In order to measure statistical significance of the differences between the IAA and the ASA scores, we take into account the pairwise character of the data: we have classifications for the same queries by both searchers and assessors. Cohen's Kappa is not a suitable measure to measure the pairwise agreement between two

Table 6 Example of a query and its agreement (Jaccard) by two annotators.

Query: *beamer toc color*			
Dimension	**Annotator** 1	**Annotator** 2	**Agreement score**
Topic	Computers, Reference	Computers	0.5
Action type	Informational	Informational	1
Modus	Text	Text	1
Source authority sensitivity	High	Very high	0.67
Spatial sensitivity	Very low	Very low	1
Time sensitivity	Very low	Very low	1
Specificity	Very high	High	0.67

Table 7 Means for PW-ASA and PW-IAA: pairwise agreement values calculated per query and then compared in a pairwise manner for the four reliable dimensions. All reported differences are significant with $p < 0.0001$. The table also shows the effect size in terms of Cohen's d.

	Mean PW-IAA	Mean PW-ASA	Cohen's d
Topic	0.66	0.50	0.43
Modus	0.82	0.74	0.26
Spatial sensitivity	0.95	0.86	0.36
Time sensitivity	0.91	0.70	0.61
Average over dimensions	0.84	0.70	

annotations for one single query because it aggregates annotations over a complete data set. We therefore followed a different approach for comparing the annotations *per query*: For the categorical dimensions, we regarded the assigned values per dimension (e.g. one annotator has selected 'navigational' and 'informational' as values for the dimension Action type) as elements in a set. Two of these sets (the values selected by two annotators for the same query) can then be compared using a set-similarity measure. We choose to use the Jaccard index (Jaccard, 1901), defined as:

$$J = \frac{|A \cap B|}{|A \cup B|} \ , \tag{3}$$

where A is the first set of assigned values and B the second.

For the ordinal dimensions, the weight matrix in Table 3 was used again, but each score was normalized by its maximum attainable value to have a similarity score between 0 and 1. For a given query, its annotation similarity by two assessors now consists of a vector of Jaccard scores (for the categorical dimensions) and normalized distances (for the ordinal dimensions). An example is shown in Table 6. We will refer to the agreements based on these pairwise scores as PW-IAA and PW-ASA to distinguish them from the κ-scores.

By calculating the average agreement between assessor–assessor pairs and searcher–assessor pairs for each query, we created pairwise data for each dimension. This allows us to perform a paired significance test. Table 7 shows the results aggregated over all queries per dimension, as well as the average pairwise inter-assessor agreement (PW-IAA) and the pairwise assessor–searcher agreement (PW-ASA) scores over the four reliable dimensions. We only take into account the reliable dimensions (κ-IAA $>= 0.4$, see Table 4), thereby disregarding the aspects of query intent that cannot reliably be measured.

We performed a MANOVA to ensure that the independent variable (IAA or ASA) influences the difference on at least one dimension, which it did with $p < 0.0001$. A paired t-test showed that in all four dimensions, the results for IAA and ASA are significantly different from each other with $p < 0.0001$. This means that indeed IAA scores are significantly higher than ASA scores and consequently post

hoc assessments are not a valid method for intent classification. The table also shows the effect size in terms of Cohen's d: the standardized difference between PW-IAA and PW-ASA.

Note that the PW-scores in Table 7 are conceptually different from the κ-scores in Tables 4 and 5: the PW-scores have been calculated per query for the purpose of pairwise comparison, and then averaged over all queries. The κ-scores, on the other hand, have been calculated per annotator couple over de complete query set and take agreement by chance into account.

5.4 Are some queries easier to classify than others?

In Table 7 we showed the pairwise assessor-searcher agreement (PW-ASA) for each dimension, averaged over all queries. We can also calculate the PW-ASA per query, averaged over the reliable dimensions. This gives the agreement score between the assessor and the searcher for one particular query. We use this as a measure for the difficulty of the query for intent classification: the lower the PW-ASA for a query, the more difficult it was for the assessor to classify the query according to its intent.

After calculating the assessor–searcher agreement for all queries, we see a large variation in scores: the standard deviation is 0.21, with a mean of 0.70. In this section, we investigate three characteristics of queries that may influence classification difficulty: (1) query length, (2) the position of a query in the session and (3) the type of transition from the previous query (reformulation, or a completely new topic).

First, we hypothesized that the length of the query has an effect on the ease of assessment. One could argue that intent of longer queries is easier to assess than the intent of shorter queries because longer queries contain more information for the assessor. On the other hand, one could argue that longer queries might also lead to more ambiguity. We investigated the correlation between query length and PW-ASA using Kendall τ and we found that these two quantities were very weakly (positively) correlated $\tau = 0.096$ ($p < 0.001$).

Second, we hypothesized that assessor–searcher agreement would increase as a session progresses. According to Silverstein et al. (1999), the session context of the query is a valuable source of information for search intent. It is a characteristic of a user's search behavior that a query is embedded in a series of queries.[5] It seems intuitive that during the session the annotator gains an increasingly better understanding of the searcher's intent. We defined a session as a series of queries issued by a single user, with at most a time span of 30 minutes between two consecutive queries. We performed an analysis of the relation between the PW-ASA for a query and the relative position of the query in its session (the ordinal position of the query divided by number of queries in the session), again using Kendall τ. No correlation could be identified: $\tau = 0.03$ with $p = 0.16$. Thus, query intent assessment does not become easier as a session progresses.

However, observation of the sessions shows that most sessions contain multiple topics. For example, in one session, a user searched for "computing for graphical

[5] We should recall here that for privacy issues, the searchers were allowed to skip some of their queries in their annotations. This may have influenced the continuity of the annotated sessions.

Table 8 Example of session with automatically determined query transition types.

# in session	query	transition type
0	information gain	New topic
1	gain ratio	Reformulation of query 0 (words overlapping: gain)
2	libsvm or timbl	New topic in same session
3	c4.5 representation	New topic in same session
4	c4.5 names file	Reformulation of query 3 (words overlapping: c4.5)

models" and "ticket amsterdam londen city". Each of these queries is the start of a new topic within the session, and each of these queries may be reformulated in order to get better results. We decided to have a look at the query transitions within the session instead of the query's ordinal position. According to (Rieh and Xie, 2006), the transition from one query to the next query in the same session can aid in deriving search intent, because a reformulation of a query can be a refinement of the underlying intent. We hypothesized that queries that are a specification, generalization or reformulation of a previous query are easier to classify than queries that start a new topic in the session, because the user made an additional attempt to get the relevant results.

To analyze the influence of query transition types on the difficulty of intent classification, we use the query transition categorization as proposed by Lau and Horvitz (1999). In earlier work (Hinne et al., 2011), a rule-based classification of query transitions was implemented and applied to the Microsoft query log data. We now applied the same classification rules to queries in our data set. Table 8 shows an example of an automatically labeled session using these rules.

In order to test the hypothesis that queries that are a specification, generalization or reformulation of a previous query have a higher assessor–searcher agreement than queries that start a new topic, we created two groups of queries: (1) Queries resulting from a reformulation, specialization or generalization of the previous query, and (2) queries that start a new topic (at the beginning of a session or within a session). The mean PW-ASA for the first group of queries is 0.76, while the mean PW-ASA for the second group is 0.66. This is a significant difference ($t = -4.2$; $p < 0.0001$, using a Welch t-test). Thus, after a reformulation, specialization or generalization within a session, the intent of the query is easier to annotate than when a query is the start of a new topic.

6 Discussion

In Section 5, we showed that of the seven dimensions in our intent classification scheme, four can reliably be used for query annotation: topic, modus, time sensitivity and spatial sensitivity. Of these four, only the annotations on the topic and spatial sensitivity dimensions are valid when compared to the searcher's annotations. In this section, we comment on our methodology, and discuss the implications of our findings.

6.1 Methodological contributions

For data collection, we used a plugin in Firefox that records all queries issued in a Google domain. The searcher had full control over the submission of his queries

to our experiment, and chose his own time for annotating queries. This made participating in the experiment a task with relatively low effort. We think that this experimental set-up is a good method for collecting 'real life' search data that respects the subject's privacy.

In Section 1, we explicitly assumed "that the searcher himself knows better than an external assessor what he intended to find when issuing his query". We should note that this assumption does not mean that the searcher's own annotations in a intent classification scheme can be considered the 'ground truth' of his intentions. As a measuring instrument, the intent classification scheme is a derivative of the actual underlying intent; the scheme molds the abstract concept of intent into a concrete, measurable form.

In the analysis of our data, we distinguished the seven dimensions of our intent classification scheme, drawing conclusions on each dimension separately. In addition, we looked at differences between queries. This level of granularity is necessary for good interpretation of the obtained results because it reveals effects that would stay hidden if only aggregated results would have been analyzed, the most important effect being the large differences in the reliability and validity of the dimensions.

6.2 The size and nature of our data set

We already stated in Section 4 that we did not intend to collect data that are representative for all queries issued by all search engine users. Instead, we chose to limit our subject pool to colleagues, all computer scientists. As a result, our data set is small compared to previously collected query labeling data sets (Calderón-Benavides et al., 2010; Baeza-Yates et al., 2006). However, the value of our data collection is not in its size but in the fact that all queries have been labeled according to their intent, by the searcher himself as well as at least two external assessors. The limited size of our data set is mainly a problem for the intent dimensions in which the classification is much biased towards one value: 511 of the 605 queries have been classified as 'text' in the dimension 'modus' and 545 have been classified as 'informational' in the dimension 'action type'. We therefore chose a conservative agreement measure that takes into account chance agreement: Cohen's κ.

135 of the 605 queries occurred multiple times in our data set, due to participants issuing the same query multiple times and as an artefact of the data registration (using the browser back button to return to the search result-page generates a duplicate). We saw that sometimes searchers gave different annotations to the same query. This can be explained by two different reasons: a searcher has a different interpretation of the same query on second presentation or a searcher entered the same query twice but with a different intent. We should note here that there was a maximum delay of 48 hours between issuing a query and annotating it, because searchers were presented with the queries they issued during the last 48 hours. This delay may have made the labelling of queries more difficult. In addition, the Firefox plugin that we developed did not save a window or tab ID for each query, only a timestamp. This means that queries that were issued by the same user in different browser tabs are saved in chronological order, as if they belonged to the same session. Some of the topic alternations within sessions that

we encountered in our analyses (see Section 5.4) may have been caused by a user having two or more tabs active with different search sessions.

Another limitation of the data is that both searchers and assessors made mistakes in their annotation. Some of these mistakes were caused by the setup of the annotation task. For the ease of annotation by the searcher, the values from the previous query were kept when going to the next query to classify. We chose to do this because in many cases subsequent queries share aspects of their intent.[6] In some cases, this resulted in obvious errors: 'Health' as topic category for the query "grand theft auto iv wine steam" and 'Computers' as topic category for the query "stomen verkoudheid" (*"inhale steam as cold remedy"*).

6.3 Comparison to previous work

In Section 2.2, we described a number of previous studies on query intent classification. In this section, we investigate how our results compare to the results from those studies, specifically with respect to the reliability of the annotations. We should clarify here that our study is not aimed at identifying many possible intents in order to be able to present a variety of search results to the user (the diversification approach: Santos et al. (2011); Sakai (2012)). Instead, we are interested in the intent that was meant by the searcher himself, and we aimed to discover whether external assessors could recognize this intent. The dimension 'Action type' is of particular interest because it is the original Broder taxonomy of web search (a query is informational, transactional or navigational in nature) and is used very frequently in query intent studies. However, in our experiments, queries could not reliably be classified according to this dimension (κ-IAA was 0.29; κ-ASA only 0.09).

The work that is most similar to our work is the paper by Lewandowski et al. (2012). In an extended variant of the 'Action type' (Broder) dimension, they found that the external assessor agreed with the searcher for only 11% of the queries, which is very much comparable to the low κ-ASA that we found. There are two main differences between the work by Lewandowski et al. (2012) and our work: first, a large portion of their paper focuses on the distribution of queries over intent types in different types of data. Second, they use a larger query set than we do, without any domain restrictions. Although a large data set allows for valuable analyses, the open domain is at the same time a weakness. In fact, the authors recommend to use expert assessors because nonexpert intent judgments is highly error-prone.

In related research, Ashkan et al. (2009) annotated 1700 queries on two dimensions derived from the Broder taxonomy: commercial/noncommercial and navigational/informational. Each query was annotated by 3 annotators and the reported agreement was 81% and 87%. The reported scores are the absolute agreement percentages, not κ-scores—they have not been corrected for chance agreement. κ-scores would turn out to be lower, so a comparison is difficult in this case. Baeza-Yates et al. (2006) annotated around 6000 queries on the dimensions 'topic' (using a set of options derived from the ODP categories that we used) and 'action type' (called 'goal' by the authors: informational/not-informational/ambiguous).

[6] For the external assessors, the form was emptied after each query.

No mention was made of how many annotators performed the task nor of their agreement. A comparison of results is therefore not possible.

González-Caro et al. (2011) and Calderón-Benavides et al. (2010) reported on the annotation of 5249 queries on many dimensions. Most queries were annotated only by a single annotator, but 10% of the queries were annotated by two judges. The authors report both absolute agreement and κ-scores. The reported agreements are considerably higher than the agreements we found. For example on the dimension 'task' (informational/not-informational/both) they achieve considerably higher agreement than we do on the comparable dimension 'action type' ($\kappa =$ 0.63 vs. 0.29). Possibly, the distinction informational/transactional/navigational is more difficult than deciding whether something is informational or not.

Even more striking is the high agreement obtained on 'time sensitivity' ($\kappa =$ 0.98) and 'scope' ($\kappa = 0.93$). Agreement on topic is also higher than in our experiments, but again a direct comparison is difficult due to a slightly different definition (the categories used are derived from ODP, Wikipedia and Yahoo!, while 'News', 'Reference', 'Business' and 'Community' were given their own dimension called 'genre').

Besides these definition differences, a possible reason for the discrepancy could be that the data by González-Caro et al. (2011) and Calderón-Benavides et al. (2010) has been extracted from a search engine query log, whereas our data was gathered from a rather homogeneous and smaller population of searchers. The extremely high κ-scores for some of the dimensions might be explained if the authors provided the assessors with very strict annotation guidelines.[7] For example, if the assessors follow rules such as "The query is time sensitive if it contains a time phrase such as a month or a year", then the annotation task becomes more objective. However, in our opinion, the underlying intent of a query is more than the textual content of the query; part of the intent can be hidden. Thus, the intent of a query can be time specific without a time phrase being literally mentioned.

An alternative approach to query intent discovery (see Section 2.1) is *verbalization*: finding the longer question underlying a short query, as proposed by Law et al. (2009). Advantages of that approach is that it is not necessary to define categories in an intent classification scheme, and that classification of the query into that scheme is not needed. This might be an attractive alternative to query intent classification as we considered it, but it is difficult to directly compare the two approaches because much depends on the actual implementation of intent discovery in a search engine. The *formalization* approaches (Zhou et al., 2007; Tran et al., 2007) are evaluated in specific domains: the scope of the ontology that is the backbone of the system. Zhou et al. (2007) manually formulate (short) keyword queries from (longer) natural language queries in three domains: *geography*, *job* and *restaurant*. The task of their system is to construct a formal query that is semantically equivalent to the original natural language query. If it is, the intent of the short query was recognized correctly. Tran et al. (2007) consider a formal query generated by their system as correct if it retrieves the same answers as the natural language query. These ontology-based retrieval approaches relate to the 'topic'-dimension in our model; they aim to create a detailed representation of the content of the query using a restricted vocabulary. The ontologies do

[7] The annotation guidelines are not included in their paper.

not cover meta-information such as the modus (text/image/video/map/other) or action type (informational/navigational/transactional) of the query.

Ontology-based approaches can be valuable in a restricted domain. In our model, we used the generic ODP categories as values for the 'topic' dimension. This is only relevant for open-domain retrieval. In a more restricted domain (e.g. biomedical literature), the values for the topic dimension should be more specific. Alternatively, in a restricted domain, 'topic' could be replaced by a domain ontology to which the query terms are mapped. This is an interesting direction for future work.

6.4 Implications for future automatic classification and retrieval

In Section 3 we explained the potential of our classification scheme for improving search results. Now we found that human annotators were able to annotate two of the seven dimensions with valid values according to the searcher: topic and spatial sensitivity.

Our experiments suggest that classification of queries into topic categories is a feasible task, even though we had 17 different topics to choose from[8]: on average, external assessors reached a moderate agreement ($\kappa = 0.42$) with the searcher. This is good news for a future implementation of automatic query classification because topic plays an important role in query disambiguation and personalisation (see the example of the Java programmer not interested in traveling to Indonesia or vice versa). 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 ($\kappa = 0.41$) 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 labels are the values for the dimensions in the classification scheme, and the features are the query terms — the textual content of the query. In order to get a feeling for the difficulty of the automatic classification task, we performed a manual analysis to investigate the relation between the textual content of the query and the classified intent. This analysis showed that the dimensions 'modus', 'time sensitivity' and 'spatial sensitivity' were for the majority of queries not reflected by their textual content. 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'; This may explain why it was not possible to reach moderate agreement between assessor and searcher on this dimension (κ-ASA $= 0.22$).

In addition, 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, only 36 were annotated with a high spatial sensitivity. On the other hand, in the queries with low or very low spatial

[8] The bias towards computers and science queries in our data set is accounted for by the chance agreement in the κ-scores.

sensitivity, 11 location references occurred (e.g. “landesvermessungsamt nordrhein westfalen”).

This analysis shows that for many intent dimensions, there is no direct connection between words in the query and the intent of the query. This means that for automatic classification, it is difficult to generalize over queries. For example, the presence of a location reference is not a clear clue that the query is spatial sensitive. However, the most likely intent can still be learned for individual queries by following the diversification approach in the ranking of the search results (see Section 2.3): The engine can learn the probability of intents for specific queries by counting clicks on different types of results. This way, a search engine could learn that someone searching for “photosynthesis” or “coen swijnenberg” is likely to expect an image as search result. 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.

7 Conclusions

The quality of a search engine depends on the ability to present results that match the searcher’s intent. However, recognizing the intent of a user is difficult because queries are often short and/or underspecified.

In the literature on query intent classification, intent annotations are created by external assessors, not the searchers themselves. In some previous studies, the *reliability* of the annotations is measured as the inter-annotator agreement between the assessors, but, to our knowledge, there is no previous literature in which the assessors’ classifications are compared to classifications by the searchers themselves in order to investigate the *validity* of the annotations. In this paper, we measured both reliability and validity for the intent classification of queries. For that purpose, we designed an intent classification scheme with seven dimensions, based on schemes in the literature: topic, action type, modus, source authority sensitivity, spatial sensitivity, time sensitivity and specificity. We collected a set of 605 queries among colleagues using a browser plugin that logged their interactions with the Google search engine during their daily work activities.

We asked the searchers to label their own queries according to the classification scheme. Each query was also classified by external assessors. We used the agreement between the external assessments as measure of *reliability* for our query classification scheme, and we compared the assessor’s labels to the searcher’s labels in order to measure the *validity* of the labels. We analyzed the annotations per dimension, and per query.

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 when compared to the searcher’s annotations ($\kappa > 0.4$). The difference between the inter-assessor agreement and the assessor–searcher agreement was significant on all dimensions. This shows that the agreement between external assessors overestimates the validity of the intent classifications.

A per-query analysis of the results showed that there is almost no correlation between query length and assessor-searcher agreement. Furthermore, we have not been able to find evidence that query intent assessment becomes easier as a session progresses, but queries resulting from a reformulation, specialization or generalization of the previous query are easier to annotate than queries that start a new topic.

A comparison to previous work was difficult because most authors did not report agreement scores, or they measured absolute agreement instead of κ-scores. Studies in which agreement scores are reported showed a higher inter-assessor agreement than we found with our data. This may be due to the nature of the data, differences in annotation schemes and/or differences in annotation guidelines. We emphasized that the underlying intent of a query is more than the textual content of the query; queries are often underspecified with respect to the searcher's intent. A manual analysis of the relation between the textual content and the intent of the queries in our collection confirmed this. For example, none of the queries annotated by the searcher with the 'image' modus contained words such as 'image' or 'picture'.

Web search engines can learn the most likely intent of individual queries by counting clicks on results that represent possible intents. Also, search engines can take into account the context of the query: previous queries from the same session, previous queries from the same searcher in different sessions and, if access is provided, the interest of the searcher obtained from documents or emails. We think that with this information, the search engine can become better than a human external assessor in predicting the underlying intent of a query.

In conclusion, we showed that Broder (2002) 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 ground truth. This is already common practice for relevance assessments in most TREC benchmark tasks. In addition, we recommend that future research explores the broader context (previous queries, other computer interactions) of a searcher for recognizing the hidden intent behind a query .

References

Ashkan A, Clarke C, Agichtein E, Guo Q (2009) Classifying and characterizing query intent. Advances in Information Retrieval pp 578–586

Baeza-Yates R, Calderón-Benavides L, González-Caro C (2006) The Intention Behind Web Queries. In: Crestani F, Ferragina P, Sanderson M (eds) String Processing and Information Retrieval, Springer-Verlag, Berlin Heidelberg, LNCS 4209, pp 98–109

Belkin N, Oddy R, Brooks H (1982) Ask for information retrieval: Part i. background and theory. Journal of documentation 38(2):61–71

Binet A (1907) The mind and the brain. London: Kegan Paul, Trench, Trübner

Brenes D, Gayo-Avello D, Pérez-González K (2009) Survey and evaluation of query intent detection methods. In: Proceedings of the 2009 workshop on Web Search Click Data (WSCD), ACM, pp 1–7

Broder A (2002) A taxonomy of web search. In: ACM SIGIR forum, ACM, vol 36, pp 3–10

Buchanan G, Cunningham S, Blandford A, Rimmer J, Warwick C (2005) Information seeking by humanities scholars. In: Rauber A, Christodoulakis S, Tjoa A (eds) Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science, vol 3652, Springer Berlin / Heidelberg, pp 218–229

Calderón-Benavides L, González-Caro C, Baeza-Yates R (2010) Towards a Deeper Understanding of the User's Query Intent. In: Workshop on Query Representation and Understanding, SIGIR 2010, pp 21–24

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1):37–46

Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70(4):213–220

Croft B (2010) Thoughts (and Research) on Query Intent. Presentation in CSIRO seminar

Dervin B, Nilan M (1986) Information needs and uses. Annual review of information science and technology 21:3–33

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

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

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

Jaccard P (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37:547–579

Landis J, Koch G (1977) The measurement of observer agreement for categorical data. Biometrics pp 159–174

Lau T, Horvitz E (1999) Patterns of search: analyzing and modeling Web query refinement. In: UM '99: Proceedings of the seventh international conference on User modeling, Springer-Verlag New York, Inc., Secaucus, NJ, USA, pp 119–128, URL `http://research.microsoft.com/users/horvitz/ftp/queryrefine.pdf`

Law E, Mityagin A, Chickering M (2009) Intentions: A game for classifying search query intent. In: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, ACM, pp 3805–3810

Lewandowski D, Drechsler J, Mach S (2012) Deriving query intents from web search engine queries. Journal of the American Society for Information Science and Technology 63(9):1773–1788

Martzoukou K (2005) A review of Web information seeking research: considerations of method and foci of interest. Information Research 10(2):10–2

Rieh S, Xie H (2006) Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing & Management 42(3):751–768

Rose D, Levinson D (2004) Understanding user goals in web search. In: Proceedings of the 13th international conference on World Wide Web (WWW 2004), ACM, pp 13–19

Sakai T (2012) Evaluation with informational and navigational intents. In: Proceedings of the 21st international conference on World Wide Web, ACM, pp 499–508

Santos R, Macdonald C, Ounis I (2011) Intent-aware search result diversification. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information, ACM, pp 595–604

Silverstein C, Marais H, Henzinger M, Moricz M (1999) Analysis of a very large web search engine query log. In: ACM SIGIR Forum, ACM, vol 33, pp 6–12

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

Sushmita S, Piwowarski B, Lalmas M (2010) Dynamics of genre and domain intents. Information Retrieval Technology pp 399–409

Terai H, Saito H, Egusa Y, Takaku M, Miwa M, Kando N (2008) Differences between informational and transactional tasks in information seeking on the web. In: Proceedings of the second international symposium on Information interaction in context (IIiX 2008), ACM, pp 152–159

Tran T, Cimiano P, Rudolph S, Studer R (2007) Ontology-based interpretation of keywords for semantic search. The Semantic Web pp 523–536

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

Wilson T (2006) On user studies and information needs. Journal of Documentation 62(6):658–670

Zhou Q, Wang C, Xiong M, Wang H, Yu Y (2007) Spark: adapting keyword query to semantic search. The Semantic Web pp 694–707