Deriving Query Intents from Web Search Engine Queries

Dirk Lewandowski, Jessica Drechsler, Sonja von Mach

Hamburg University of Applied Sciences, Department Information, Finkenau 35, D—22081 Hamburg, Germany

Corresponding author: Dirk Lewandowski; dirk.lewandowski@haw-hamburg.de

This is a preprint of an article accepted for publication in *Journal of the American Society for Information Science and Technology* copyright © 2012 American Society for Information Science and Technology.

Abstract

The purpose of this paper is to test the reliability of query intents derived from queries, either by the user who entered the query or by another juror. We report the findings of three studies: First, we conducted a large-scale classification study (approximately 50,000 queries) using a crowdsourcing approach. Then, we used click-through data from a search engine log and validated the judgments given by the jurors from the crowdsourcing study. Finally, we conducted an online survey on a commercial search engine's portal. Since we used the same queries for all three studies, we were able to compare the results and the effectiveness of the different approaches, as well. We found that neither the crowdsourcing approach using jurors who classified queries originating from other users, nor the questionnaire approach using searchers who were asked about their own query that they just entered into a web search engine, lead to satisfying results. This leads us to conclude that there is little understanding of the classification tasks, even though both groups of jurors were given detailed instructions.

While we used manual classification, our research has important implications for automatic classification, as well. We must question the success of approaches using automatic classification and comparing its performance to a baseline from human jurors.

Keywords: search engines, information needs, query classification, user intent, web queries, web searching

Deriving Query Intents from Web Search Engine Queries

Search engines are by far the major means to finding information on the Web. In just one month, 131 billion queries were posed to the general-purpose search engines (ComScore, 2010). Studies report that the performance of search engines in terms of retrieval effectiveness is only moderate (Bar-Ilan, Keenoy, Yaari, & Levene, 2007; Griesbaum, 2004; Véronis, 2006; Lewandowski, 2008), and search engines' responses to more complex search tasks like exploratory searches (Marchionini, 2006) are considered poor (Singer, Norbisrath, Lewandowski, Vainikko, & Kikkas, 2011). However, when considering simpler tasks such as finding home pages, the performance of search engines is quite good (Lewandowski, 2011). It is clear that first, the quality of search engines' results depends on the difficulty of the task, and second, that search engines perform better on simpler tasks. To identify task types and to provide the best suitable results for the individual tasks, it is important to identify users' intent behind their queries.

In this paper, we focus on identifying the following query intents:

1. Informational, where the user aims at finding some documents on a topic he is interested in.
2. Navigational, where the user aims at navigating to a web page already known or where the user at least assumes that a specific web page exists.
3. Transactional, where the user wants to find a web page where a further transaction (e.g., downloading software, playing a game) can be performed.
4. Commercial, where the query has "commercial potential", i.e. the user might be interested in commercial offerings. This can also be an indicator whether to show advertisements on the search engine results page (SERP).

5. Local, where the user is searching for information near his current geographic position.

Identifying query intent in search engine logs is important not only for the information science research community, but also for search engine vendors and vendors of other information systems. While in the last years, some studies have been published mainly on automatically identifying query intent, it is still unclear how reliable classification tasks concerning user intent are. Studies that propose automatic classification of queries according to their intent are usually based on a set of manually classified queries, which is used for training and evaluation. It is assumed that this baseline set is reliable. However, from our experience, this assumption does not hold true. Therefore, the motivation for this paper is to systematically investigate whether query intents can be reliably derived from the queries.

Our results are important for the research area of query understanding, which recently has received a lot of attention. This great interest can be seen from the number of recently published papers (e.g., Church & Smyth, 2009; Mendoza & Baeza-Yates, 2008; Mendoza & Zamora, 2009; Pitler & Church, 2009; White, Bailey, & Chen, 2009) and from workshops such as those held at SIGIR 2010 (Croft, Bendersky, Li, & Xu, 2010), SIGIR 2011 (Li, Xu, Croft, & Bendersky, 2011), and the session track in TREC (Kanoulas, Clough, Carterette, & Sanderson, 2010).

In this paper, we address the reliability of classifying user intents on different levels. We present the results of three studies, using (1) a crowdsourcing approach to manually classify approximately 50,000 queries, then (2) identifying query types through click-through analysis, and (3) finally conducting an online survey on a major German search portal.

The rest of this paper is organized thus: First, we review the literature on query intent classification, and then we state our research questions. After that, we present our methods

for the three studies conducted. Then follows the presentation of the results, a discussion of these, and our conclusions and suggestions for further research.

Literature Review

This paper deals with query intents in Web search queries. We will discuss the approaches to Web query intent classification, and the major studies using manual classification, and automatic classification, as well.

It must be stated that the scope of this paper is on deriving query intent from isolated queries, i.e. we do not consider queries in the context of a searching session. While query understanding can greatly benefit from session information (He, Göker, and Harper, 2002), (1) there is still room for improvement on the query basis, (2) session data may not always be available, (3) sessions in Web searching are generally short and leave much room for interpretation (Jansen, Spink, Blakely, and Koshman, 2007), and (4), there are many studies conducted using only query information. As one aim of our study is to check the reliability of the approaches used on the query basis, it is justifiable to stay on that level, while keeping in mind that the reliability of query classification can be improved through session information.

The idea of deriving query intents from Web search queries was proposed by Andrei Broder in his seminal paper "A Taxonomy of Web Search" (2002). Since then, much research has been done, especially on automatically deriving query intents.

Jansen, Booth, and Spink (2008) define three sub-areas of query intent research: "(1) empirical studies and surveys of search engine use, (2) manual analysis of search engine transaction logs, and (3) automatic classification of Web searches". A review of studies on browsing behavior and browsing types can also be found in Jansen, Booth, and Spink (2008). In our review, we concentrate on the classes of query intents used, as well as on manual and automatic classification. We review the body of literature, which we divided into three sections. First, we look at different classifications of query intents, and then we review

studies measuring the ratio of these intents in Web search engine logs, divided into manual and automatic classification.

Classifications of Query Intents

The distinction between different query types has long been a topic of interest. Even long before the advent of the web and the query specialities that searching its contents brings with it, certain query types (derived from information needs) have been proposed. Kantor (1976) proposed "known item searches", i.e. a user in a library searches for an item already known to him. Frants, Shapiro, and Voiskunskii (1997, p. 38) separated concrete from problem-oriented information needs from which certain query types can be derived. However, while this research from library and information science forms the basis of query type distinction, the query type research we focus on in this paper considers queries posed to web search engines and the underlying information needs of web search engine users.

Researchers have proposed different classifications of query intents for web search engine queries. While there is some understanding on basic intents that are mainly based on Broder's paper (2002) on web query classification, it seems that there are as many classification approaches as there are investigations on the topic.

Following Broder's definition, informational queries users want to find information on a certain topic. Such queries usually lead to a set of results rather than to just one suitable document. Informational queries are similar to queries sent to traditional text-based information retrieval systems. According to Broder (2002), such queries always target static web pages. But the term "static" here should not refer to the technical delivery of the pages (e.g., dynamically generated pages by server side scripts like php or asp), but rather to the fact that once the page is delivered, no further interaction is needed to get the desired information. An example for an informational query is "life in Kazakhstan").

Navigational queries are used to find a certain web page that the user either already knows or at least assumes exists. Typical queries in this category are searches for a homepage

of a person or organization. A typical example is the search for a company (e.g., "General Motors"). Navigational queries are usually answered by just one result; the information need is satisfied as soon as this one correct result is found.

However, not all people searches are navigational. Rose and Levinson (2004) believed that most in fact are not. They pointed out that "a search for celebrities such as Cameron Diaz or Ben Affleck typically results in a variety of fan sites, media sites, and so on: it's unlikely that a user entering the celebrity name as a query had the goal of visiting a specific site." This may be true, but on the other hand, these queries cannot be seen as informational because one cannot assume that the user wants to read a variety of documents. What we can assume is that there are queries that fall in between "navigational" and "informational" (cf. Ricardo Baeza-Yates, Calderón-Benavides, & González-Caro, 2006), because they do not target a clearly defined website, but a single result is still enough to satisfy the user's information need.

The results of transactional queries are websites on which a further interaction is necessary. A transaction can be the download of a program or file, the purchase of a product, or a further search in a database. E.g., a user searching for "free games" would like to download such games or even play a browser-based game. In both cases, a transaction will be necessary on the website found.

Rose and Levinson (2004) specified this class of queries and used the term "resource" instead of "transaction."

Broder's taxonomy was a starting point for some researchers who refined and amended it. Kang and Kim (2003) used the same classes as Broder, but used different notations ("topic relevance task" refers to an informational query, "homepage finding task" to a navigational query, and "service finding task" to a transactional query).

Probably the most comprehensive classification of query intents can be found in Calderon-Benavides, Gonzalez-Caro, and Ricardo Baeza-Yates (2010). The authors

described nine facets that can be used to characterize a query. They considered genre, topic, task (which refers to Broder's classification and considers "informational," "not informational" and "both"), objective (in which "action" is an indication of a commercial intent, while "resource" refers to a non-commercial intent), specificity, scope, authority sensitivity, spatial sensitivity, and time sensitivity.

To help jurors classify query intents, some studies used only the queries from the search engine transaction logs themselves. Other studies use queries in the context of search sessions (e.g., Rose & Levinson, 2004). While the first approach only needs the queries themselves (which can also be derived from live ticker data if one does not have access to a search engine's log data; see Höchstötter & Koch, 2009), the latter approaches can only be used if one has access to search engine logs. While it would be preferable to use complete search sessions with their richness of data, not all search services are willing or even able to provide such data. Therefore, methods using only data on the query level may be more applicable than the more complex methods involving lots of data surrounding the query.

Studies Measuring the Ratio of Query Intents

In this section, we report the findings from studies measuring the ratio of query intents. First, we review the studies using a manual classification approach, and then we review research automatically classifying queries according to their intents.

Studies Using Manual Classification

The interest of studies that manually classify queries is often more the topics of the queries than their intents. Spink, Wolfram, Jansen, and Saracevic (2001) reported findings on the topic distribution in a web search engine log file. Deeper analysis for certain topics, such as e-commerce, health, and sexuality, is reported in Spink (2004, p. 127-160). Lewandowski (2006) combined topics with query intents to derive the distribution of Broder's query types

within topical classes. A similar approach was used by Ricardo Baeza-Yates, Calderón-Benavides, and González-Caro (2006).

Broder (2002) used a two-fold approach to classify queries. First, he conducted a user survey. In an online survey, users of the AltaVista search engine were asked to specify what they were trying to achieve with their present query. Some 24.5% of the queries were classified as navigational. Because the methods used did not allow a clear distinction between informational and transactional queries, the ratio of these queries were estimated based on the lower bounds derived from the user survey. The estimates were 39% for informational queries and 36% for transactional queries, respectively.

In addition to the survey, Broder analyzed 1,000 queries from AltaVista's transaction logs. The author gave no further information on how and by whom the classification task was done. Results are that navigational queries accounted for 20% of all queries, informational queries for 48%, and transactional queries for 30%. For the missing 2%, no explanation was given.

Rose and Levinson (2004) used session information from the AltaVista search engine. Jurors were able to see additional information such as the results clicked on and further actions of the user. It is unclear who performed the classification tasks. It should also be noted that in this study, only session-initiating queries (i.e., queries that start a new session and may or may not be followed by further queries) were used. The authors conducted three separate studies with approximately 500 queries each. While the classification is much more detailed than the three classes of query intents Broder used, the top-level categories are nearly the same. Findings are that informational queries account for 61-63% of all queries, navigational queries for 11-15%, and transactional (resource) queries for 21-27%.

Ricardo Baeza-Yates, Calderón-Benavides, and González-Caro (2006) used 6,042 queries from the Chilean web search engine TodoCL, which are classified into informational, not informational, and ambiguous. Results were that 61% were informational, 22% not

informational, and 17% ambiguous. The authors noted that the high number of informational queries resulted from the classification of names of artists, searches for products, etc. as informational.

Lewandowski (2006) used live-ticker data from three German web search engines. A total of 1,500 queries (500 from each engine) were classified by two jurors each, who had to reach agreement on each query. Taking the average over all queries from the three search engines, 45% were classified as informational, 40% as navigational, and 15% as transactional.

In Table 1, we compare the results from certain studies that measured the ratio of informational, navigational, and transactional queries (i.e., Broder's original query intents). As the studies (1) are based on different datasets, (2) were conducted at different times, and (3) used different methods, they are not directly comparable. However, the wide-ranging data give an impression that the ratios depend heavily on the dataset, the study design, and/or participants.

It is also questionable whether the amount of queries used for manual classification is adequate. As already stated by Jansen, Booth, and Spink (2008), studies classifying query intents use only small quantities of queries. However, this is understandable because classifying queries manually is a laborious task. Furthermore, one can doubt that a classification simply based on the queries without additional information produces relevant results. As Broder put it, "Inferring the user intent from the query is at best an inexact science, but usually a wild guess" (2002). However, that did not stop him from doing exactly that. Later in this paper, we will try to answer the question of how reliable such inferred query intents are.

Table 1: Comparison of the results from query intent studies using Broder's taxonomy

Study	Data from	Number of queries analyzed	Method	Informational	Navigational	Transactional

(Broder, 2002)*	AltaVista search engine	3,190	User survey	39% (est.)	24.5%	36% (est.)
(Broder, 2002)	AltaVista search engine	1,000	Manual classification; log analysis	48%	20%	30%
(Rose & Levinson, 2004)±	AltaVista search engine	approx. 1,500 (3x 500)	Manual classification; additional data such as results clicked available to the jurors	61-63%	11-15%	21-27%
(Lewandowski, 2006)	Three different German search engines	1,500	Manual classification; live ticker data	45%	40%	15%
(Ricardo Baeza-Yates, Calderón-Benavides, & González-Caro, 2006)	TodoCL	6,042	Manual classification	61%	n.a.	n.a.

*Results in Broder's user survey add up to more than 100% due to respondents placing some queries into multiple categories.

±Results in the Rose and Levinson study are not exact because they aggregate data from their three studies.

Studies Using Automatic Classification

In this section, we briefly review studies that used automatic classification approaches. Studies using automatic classification are usually based on manually classified query sets, and machine learning techniques are trained on that basis. Therefore, manually classified query sets to be reliable if they are used as baseline for the automatic approaches.

While some studies try to classify queries into the agreed-upon intents of informational, navigational, and transactional, others attempt only to derive one certain type of query intent, e.g., commercial.

Jansen, Booth, and Spink (2008) used around one and a half million queries and classified them into informational, navigational, and transactional categories. They used a small set of 400 queries to validate the accuracy of the automatic classification and found that their algorithm had an accuracy of 74%. The authors said that the rest of the queries were "problematic" and may have had more than one intent. Ricardo Baeza-Yates, Calderón-Benavides, and González-Caro (2006) used a combination of supervised and unsupervised learning to detect user intents. Their experiments were based on a sample of 6,043 manually classified queries.

Kathuria, Jansen, Hafernik, and Spink (2010) classified queries based on indicators for a certain query intent (e.g., a navigational intent is identified if the query contains companies', organizations' or people's names and domain suffixes, if the query was less than three words, and the searcher only visited the first results page). Again, the automatic classification was evaluated against a set of 400 manually classified queries.

Some studies tried to identify only one query intent, mostly whether a query had a commercial intent or not (Ashkan & Clarke, 2009; Dai, Zhao, Nie, & Wen, 2006; Kang, 2005). This comes as no surprise, since search engines heavily rely on advertising revenue and the ads (or at least the number of ads) shown could be adjusted to the query intent.

Some studies used click-through data to identify navigational queries. The click-through rate is the number of clicks on a certain result shown on the search engine results pages (SERP) divided by the number of times this result was displayed on the SERP for a certain query.

The assumption is that with navigational queries, most clicks in the search engine results pages (SERP) are allotted to just one results position, foremost the first position. Lee, Liu, and Cho (2005) used this approach to show that combined with anchor-link distribution, it is very reliable in identifying such queries. These findings are supported by another study (Lu, Peng, Li, & Ahmed, 2006).

Also, click-through investigations can be seen as a means of identifying navigational queries. The usefulness of click-through data to determine the relevance of search engine results was discussed by Joachims (2002) and confirmed in several other studies (Macdonald & Ounis, 2009; Dou, Song, Yuan, & Wen, 2008; Chao, Guo, & Wand, 2009). While for informational queries, it can be seen from the click-through data which results are relevant, for the navigational queries, a clear aggregation of the clicks on just one result should be observable. However, the position of a result does influence the number of clicks it will get (Craswell, Zoeter, Taylor, & Ramsey, 2008; Höchstötter & Lewandowski, 2009).

Because the performance of automatic classification of search queries is evaluated using a small set of manually classified queries to measure the conformance to automatic classifications, the manual classification itself must be reliable. However, in the studies reviewed, we did not find any approach to manual query classification that even addressed the shortcomings of manual classification.

Objectives and Research Questions

As we have seen from the literature review, there are no studies classifying a large number of queries manually. However, to validate automatic approaches, such work needs to be done. Our study aims at finding whether manually classifying queries according to their intents are reliable enough to be used as test data for automatic approaches. Only if this is the case can we rely on the results from automatic classification. If not, we must find better ways to derive intents from the queries, e.g., through using additional information instead of relying on just the queries themselves. However, it is hard to define an exact threshold for reliability, as the reliability threshold depends on the purpose of the query set. E.g., when using pre-classified query sets in a user study in a lab setting, classification results can be discussed with the testing persons (and ill-classified queries can be filtered out even at this stage), while when the classified queries are used as a training set for machine learning, the reliability of the classification is utterly important.

To guide our research, we use the following research questions.

1. How reliable is user-based query intent classification?
2. What is the ratio of navigational, informational, commercial intent in the query logs of T-Online, a major German search portal?
3. Is it possible to derive navigational intent from simple click-through data (i.e., the ratio of clicks on the first result presented)?
4. Is it possible to derive commercial intent from simple click-through data (i.e., the ratio of clicks on the ads section of the search engine results page)?
5. How reliable are automatic classifications of user intents?

Methods

For our investigation, we used a multi-method approach. To classify a large number of search queries, we used a crowdsourcing approach. Additionally, we analyzed click-through data to determine commercial and navigational queries. Finally, we conducted an online survey using the same queries as we used in the crowdsourcing study. To see which method is most suitable for the classification of search queries in the latter case, the same queries were used. Based on the data sample classified by human jurors, the click-through analysis was evaluated and the forms of the online survey were selected.

With the three studies conducted, we can make a qualified statement about the query intents as well as determine the reliability of the different approaches.

Study 1: Manual Classification through Crowdsourcing

One method to classify search queries into query types is to instate human jurors. The different investigations in this work are based on such a rating process. The search queries used were a random sample from the year 2007 provided by Deutsche Telekom AG, which operates T-Online, one of the largest German search portals. While the portal uses results from Google's search engine, the queries provided are original to T-Online. The sample queries were classified by workers for the crowdsourcing company Humangrid. In total,

49,919 search queries were categorized by these human jurors. The jurors were asked to assign the queries to one or more classes based on Broder's classification (2002). While in Broder's original work, each query was only assigned to one class, we found in pilot testing that there are queries that could be assigned to two or even all three classed.

We extended the classification with two additional classes, "local" and "commercial." Thus, the queries could be put into navigational, transactional, commercial, or local classifications. As mentioned, it was also possible to assign more than one class to a query. The instructions for the jurors included some examples of the particular classes.

A query is navigational when its purpose is to reach a particular site that the user has in mind (Broder 2002). The information need of the user can therefore be satisfied with one "right" result. A transactional query aims at a transaction in the web, e.g. the purchasing of software or the download of a document (Broder 2002).

In our "commercial" category one can find queries that are motivated by a commercial interest. This means the query aims at an action that can lead to costs for purchasing, such as searches for products.

Finally, a local query must have a local connection, e.g. to a town, a region or a state.

The class "informational" was not included in our classification. Instead, we decided that every query that was not put in any other category has an informational need. An informational query's purpose is to collect information; the user therefore (in most cases) needs more than one result to satisfy his information need. The information should also be available on the web in a static form, which means that no further action, except reading, is necessary (Broder 2002).

In the crowdsourcing study, it was possible to assign more than one query type to each query. The complete instructions as given to the raters can be found in the appendix.

The rating process is shown in Figure 1. Each query was rated by two jurors using binary judgments. If the jurors were at odds, the query was classified as 0.5. In this case, we call the classification of the query "non-explicit." Additionally, the current performance of the human juror was included, i.e., when a juror had experience in rating and performed their tasks (also from other experiments conducted with the same crowdsouring solution) well in the past, his or her classification was given more confidence than the classification of the inexperienced juror. However, the influence of the current performance was very low, so nearly all queries in which the human jurors were at odds were classified as 0.5.

One search query can be rated as "explicit," which means that both jurors classified the query to the same class. Such queries equated the value of one. The search query then was assigned to the given class. It was also possible for jurors to rate the query with 0, which means that the respective class did not apply to the query. The final classification is based on the rating by the human jurors, their agreement, and their past individual performance.

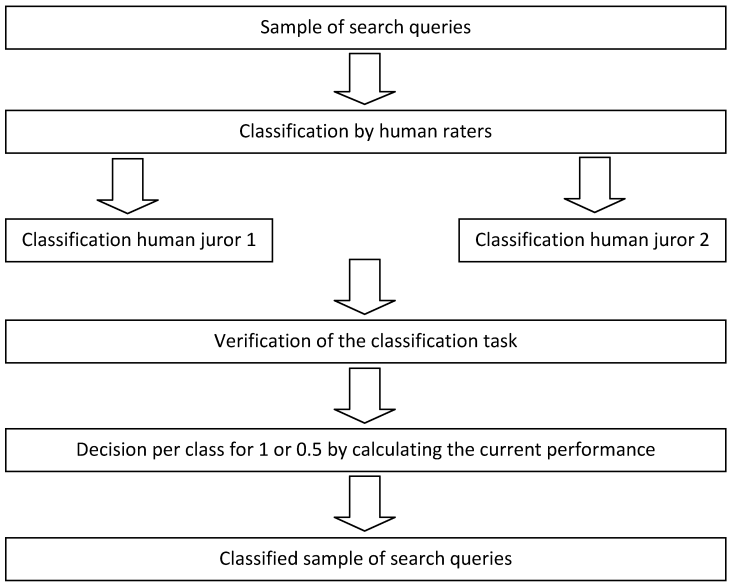

Fig. 1: Rating process for the crowdsourcing study

Study 2: Analysis of Click-through Data to Determine Commercial and Navigational Queries

As mentioned in the literature review, another method used to classify search queries is an analysis of click-through data. The following hypotheses are a basis for classifying a sample of queries into commercial and navigational queries.

- H1: Navigational queries: The majority of the clicks are made on the first search result.
- H2: Commercial queries: Compared to the average distribution of the clicks onto the search results positions, a large number are made on the ads.

The first assumption is based upon different studies that show that search engines can place the most relevant result on the first position of the search engine result page for navigational queries. Lewandowski (2011), for example, explored the success of navigational queries in different search engines. One of his research results was that Google can in 84% of the cases present the adequate result on the first position. Furthermore, in a study by Lu et al., different features for identifying navigational search queries were tested. They found that click features are most reliable for classifying queries as navigational. They concluded that this is "intuitively understandable because if a query is navigational, the navigational URL is the most clicked one" (Lu, Peng, Li, & Ahmed, 2006).

Hypothesis 2, which assumes that when a query receives many ad clicks it is more likely to have a commercial intent than a query with lower ad clicks, is supported by a study by Ashkan and Clarke (2009). Their research was a term-based analysis of commercial query intent. They posited the hypothesis that a query is likely to receive ad clicks based on the contributions from its individual terms. They were able to verify that terms with respect to their ad click-through are effective in exposing the commercial intent of queries that include such terms.

The click-through data was also collected for our sample of 49,919 queries. The data contains information about the clicks on the first search result page (and therefore on the first ten organic search results), as well as an accumulated number of clicks for the rest of the search results (clicks on the second result page up to the last result page). The number of clicks on the first eight ads was also collected. The click-through data was collected for a period of 19 months, from August 2007 to March 2009. We only used queries that got at least 100 clicks in this time-span. This limitation was chosen to avoid a bias in the results, which could emerge when only a marginal number of clicks are recorded for a query. Therefore, our first sample of 49,919 queries was stripped down to 24,807 queries that achieved a minimum of 100 clicks in the described period.

Study 3: User survey

To get to know the intention of real searchers using the queries from our sample, an online survey was arranged. The questions were based on Broder's 2002 survey. It should be noted that, as the survey was conducted on a German-language search portal, the questions also were in German. The translation provided in this article may not exactly be in accordance with all nuances of the German text. As the German versions of the questions were tested in a pilot test and no considerable problems were identified, we conclude that problems identified in the survey data (see below) will not (or at least, only to a small degree) result from the wording used.

- Question one ("Which of the following statements describes best what you tried to find with this search?") has the same meaning and deviates from Broder's question only in the way it is posed.
- Question two ("Which of the following statements describes best why you conducted this search?") is basically the same as in Broder's study, but was extended by more multiple choice answers.

- Question three ("Please write down in your own words what the perfect search result for your search would look like.") is new and was asked to control the plausibility of the answers given in questions 1 and 2.

We decided to use an online survey because this method affords an easy way to catch participators while using a search engine. The survey was shown as a link in the search results pages of the T-Online search portal. When a participant decided to join the survey, a pop-up window opened, and the participant was instructed with a short text before the survey started. A progress bar showed the progress of the survey. After finishing the online survey, the participants could return to the search results.

The collection of data was adjusted to the sample classified by humans. This means that the same queries were used: every searcher who sent one of the search queries used in the sample was asked to take part in the survey. Of course, not everybody who was invited took part, and not every search query from the classified sample was posed to the search engine while the survey was active.

The online survey ran for over a week and was delivered to the user in form of a link after he or she sent his or her search terms to the search engine. In total, 549 users participated in the survey.

As described above, the survey was structured into three questions to identify the several information needs and query types. The first question evaluated the navigational need.

1. Which of the following statements describes best what you tried to find with this search?
    - I want to get to a specific website that I already know.
    - I want to find one or more excellent websites related to my query, I do not have a specific website in mind.

While the first answer suggested a navigational search intention, the second ruled it out.

The second question should identify the other classes. We integrated a filter question that identified a commercial search query.

2. Which of the following statements describes best why you conducted this search? (check all that apply)
   - I am searching for a product or a service which I possibly want to acquire (e.g., borrowing, booking, download with costs).

Filter question

- You have marked that you are searching for a product or service. Please give us more information about your search query (one answer possible):
  - I am searching for websites with online shops and may buy something there.
  - I am searching for a seller or service provider in a **specific city/region** and do not want to buy on the internet.
  - I am searching for a seller or service provider and do not want to buy on the internet.
  - I am searching for a file that I want to download (e.g. music, pictures, software).
  - I am searching for videos or pictures which I want to look at.
  - I am searching for a website where I can communicate with other participants (e.g., chat, e-mail, voip telephony, forums or Myspace, xing, Facebook, or the like).
  - I want to do online banking on a website.
  - I want to play a game online.

- I am searching for a website where I can continue my search or rather begin the search (e.g., databases, price-comparison search engines, or the like).
- None of these statements apply to my search query.

The first answer ("I am searching for a product or a service…") indicates a commercial intent. The filter question narrows this intent either to a transactional and commercial need or to a local and commercial need. The third response option of the filter question verifies a straight commercial intent. A local search intention could only be chosen in connection with the class "commercial." This weakness of the online survey may have caused non-valid results in the class "local."

The second answer to the regular question verifies a transactional intent. The option "download" is highlighted to contrast with the next option, "watching a video or pictures," which is not a transactional but an informational intent. To make the difference clear, the words "to look at" and "download" were highlighted.

The option that asks about communication with other internet users, online banking, an online game, and the website where the search could be continued or rather begun conduces the identification of a transactional intent.

The last answer option ("none of these statements apply") suggests an informational need.

The third and last question of the survey was an open question that acts as a control question for the given answers.

3. Please write down in your own words what the perfect search result for your search would look like.
   - The perfect result for my search would be:

     ……

     ……

……

We used the answers to this question to verify the plausibility of the prior answers and to avoid misunderstandings.

The questions of the online survey were pilot-tested with a group of people who differed in age and gender. The test persons did not have any previous experience in web information retrieval. To check whether the questions were self-explanatory, we asked the test persons to remember their last search queries and then typing these queries using the questionnaire. The pilot test showed that just small improvements were necessary. After that, the questions were adjusted and their final versions as presented in this paper were used.

Synthesis of Data Collection

Table 2 gives an overview of data collected for our three studies. Figure 2 summarizes the three studies conducted in a flow-chart diagram.

Table 2: Data collected for the studies reported

Study	Number of queries	Labels
Classification task (crowdsourcing)	49,919	Navigational, transactional, commercial, local, informational*
Analysis of click-through data	24,807	Navigational, commercial
User survey	549	Navigational, transactional, commercial, local, informational

*Where no other category applicable.

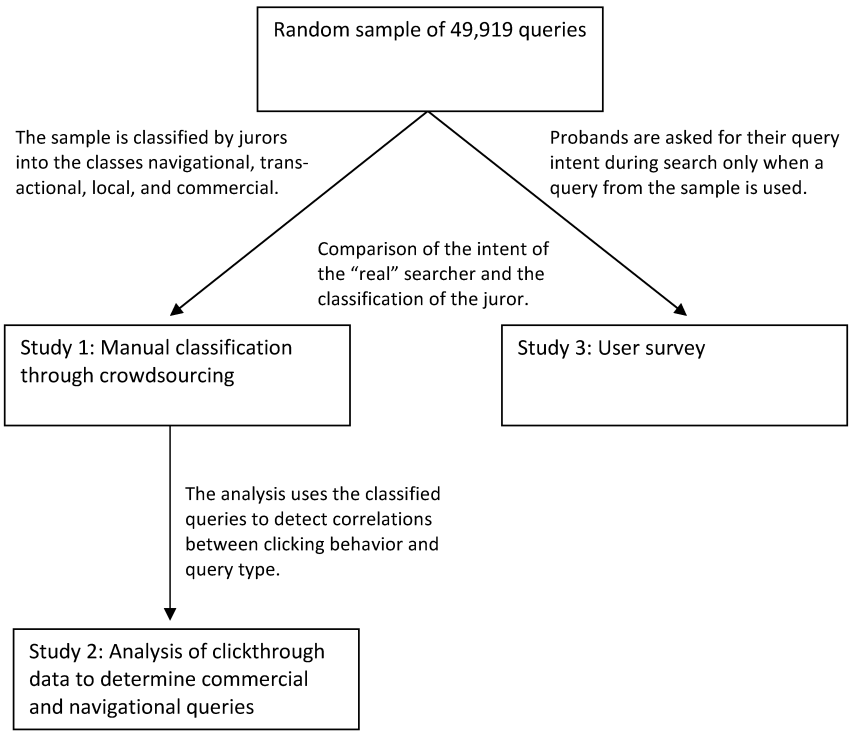

Fig. 2: Flow chart of studies

Results

In this section, we present the results of our three studies separately. However, since the studies were designed to be based on each other, we will also compare the results of the different studies to each other.

Study 1: Manual Classification through Crowdsourcing

The allocation of the several classes shows that the ranges in the categories (explicit classification: both jurors rate the same category; non-explicit classification: the jurors were at odds with each other) are very high (Table 3).

The non-explicit classification of the queries shows that, although the jurors were instructed in the classification task, queries could not always be classified into one class. If the human jurors were at odds with each other, the non-explicit (0.5) classification was assigned.

Table 3: Ratio of queries classified into individual classes

Class	**Span (explicit and non-explicit) in %**

Navigational	27 - 42
Transactional	11 - 39
Informational	22
Commercial	19 - 46
Local	9 - 26

The span of non-explicit versus explicit judgments in the several classes is high. The queries that may have a navigational intent have a span from 27 to 42% (where the queries in this class have been rated by the jurors with 0.5 or one). Transactional queries are represented in a span from 11 to 39%. The informational category (which was defined as a kind of "leftover category") contains 22% of the search queries. Commercial queries with the intent to find a product or service have a span from 19 to 46%. The queries with a local intent are the fewest, with a range of 9 to 26%.

Our findings show that simply asking users to assign queries to intentions does not produce exact results and that users often do not agree on that category should be assigned.

Study 2: Analysis of Click-through Data

To verify the method of using click-through data to identify navigational and commercial queries, we used the classified queries from study 1.

Navigational queries. The verification shows that the assumption regarding the navigational queries can be confirmed, because when we investigate the queries that received a click ratio of 90% or more on the first organic search result (see also Figure 3, second bar from the right), we found that a ratio of 80.55% were queries explicitly rated as navigational by the participants of the crowdsourcing study. In this case, 14.90% of the queries are not explicitly rated as navigational, and a ratio of only 4.55% of the queries is rated as not navigational.

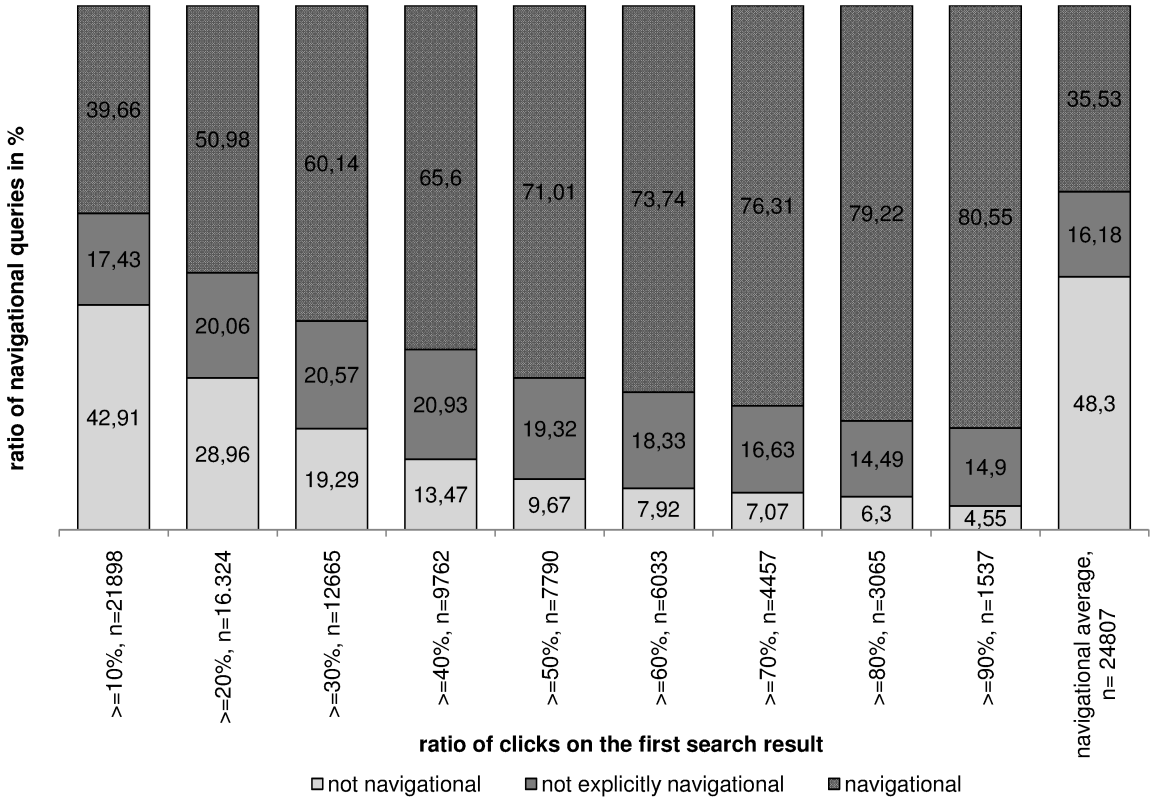

Fig. 3: The ratio of queries for navigational queries on the first search result

The other query types did not reach such a high ratio of clicks on the first result. For these, there is a wider distribution of clicks throughout the results page.

Figure 4 shows the click-through graphs for the different query types (as rated by our human jurors). Queries that are explicitly classified as navigational get the highest ratio of clicks (60.38%) on the first result. A comparison with the other query types shows the following: The queries explicitly rated as transactional achieve a ratio of 38.38% of the clicks

on the first position, the explicitly local queries a ratio of 47.06%, and the commercial search queries a ratio of 31.45%.

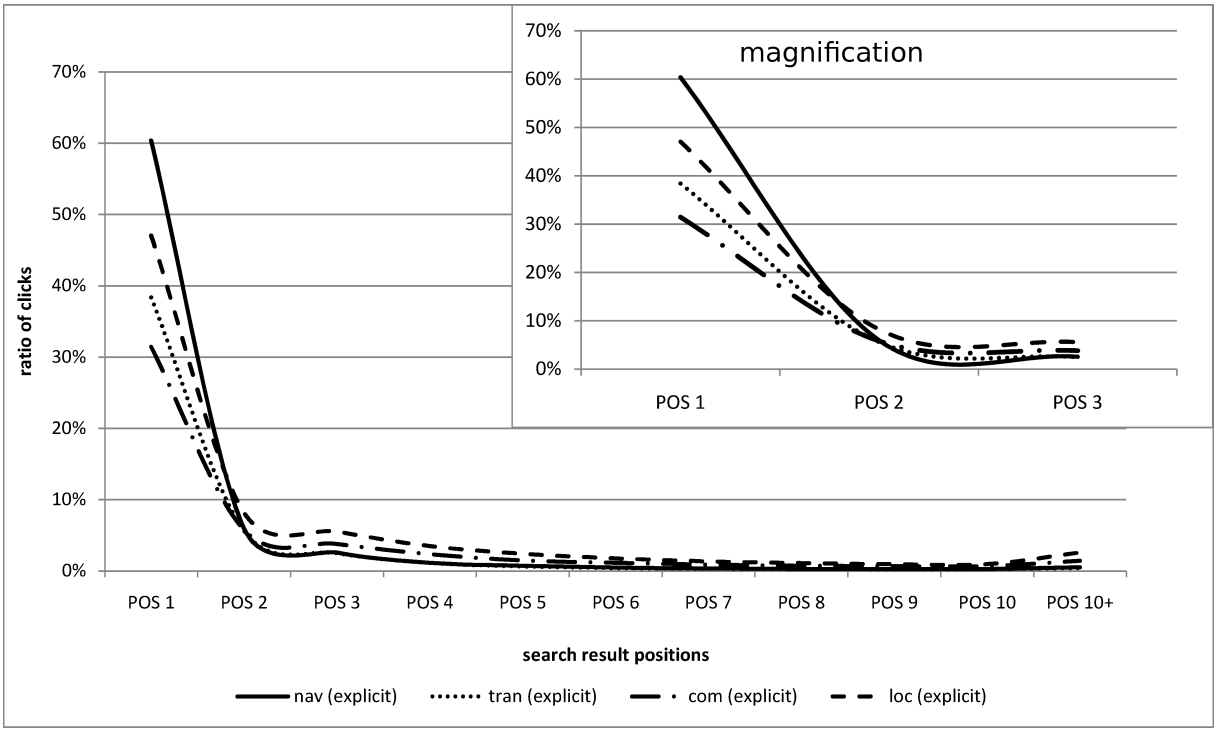

Fig. 4: Distribution of clicks on search engine results pages for navigational (nav), transactional (tran), commercial (com), and local (loc) queries

Our data supports our assumptions about navigational queries and shows that the analysis of click-through data is adequate to identify explicitly navigational queries that are easier to recognize than not explicitly navigational rated queries. When considering the error ratio (i.e., where a query rated as not navigational has more than 90 percent of clicks on the first position), we can see that with only 4.55% of queries, classifying navigational queries on the basis of clicks on the first result is of sufficient reliability.

Commercial queries. To verify the assumption about the click-through behavior for commercial queries, we looked at the ratio of clicks on the different ad results positions. Figure 5 shows that the assignment to the query type "commercial" increases with a rising ratio of clicks on the ads. The highest ratio of commercial queries is found when we have a ratio of 60% or more clicks on the ads, where a ratio of 52.07% of the queries are commercial, and only 28.43% are not explicitly commercial.

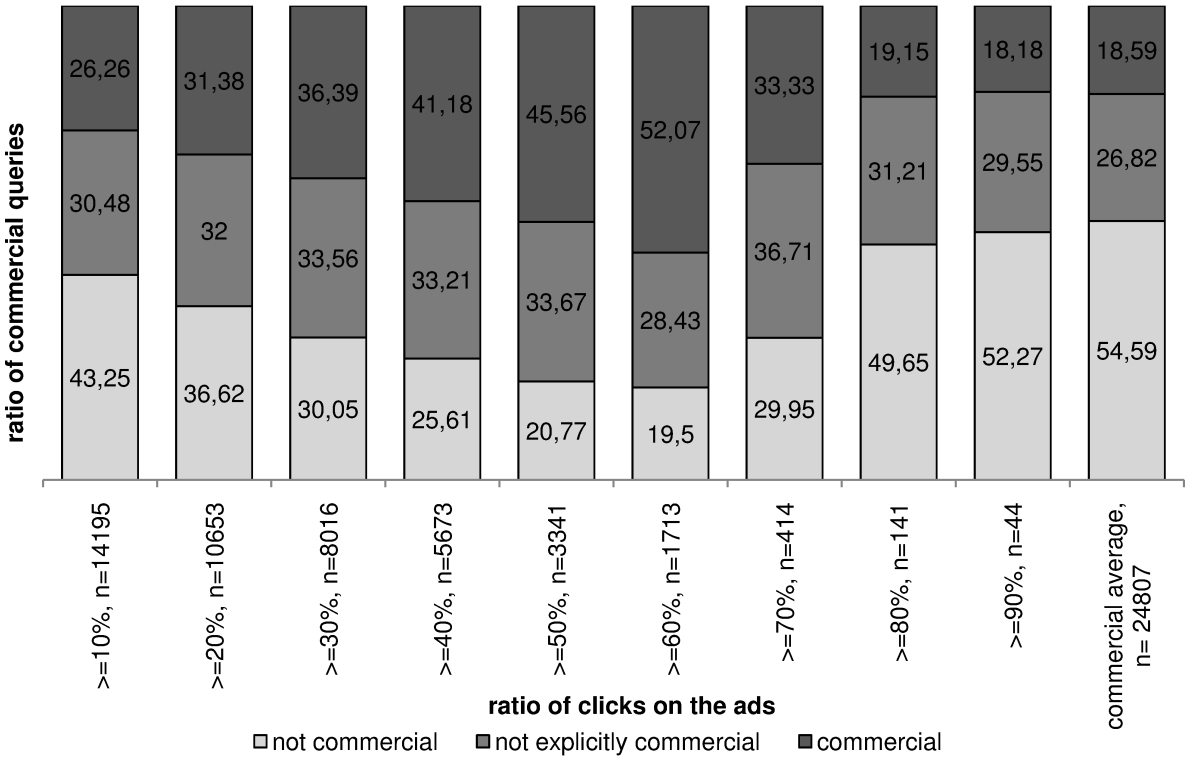

Fig. 5: Ratio of clicks on the ads for commercial queries.

When we look at the different query types and their clicking behavior, we find that compared to the other query types, the commercial queries get the highest ratio of clicks on ads (when accumulated to all positions: 20.29%), followed by the transactional queries (when accumulated to all positions: 12.67%). Therefore, we can infer that our hypothesis concerning the clicking behavior for commercial queries is right. We found a heightened ratio of commercial queries when the click rate on the ads is high, too. However, the analysis

also exposes an irregularity that occurs when clicks on the ads reach a ratio of 70% (see Figure 5). Here, the ratio of commercial queries declines again, even though the ratio of clicks on the ads increase.

Figure 6 shows that commercial queries receive more ad clicks than any other query type, and that this holds true for all ad results positions. A noticeable issue is that navigational queries also receive a large amount of clicks on the first ad result.

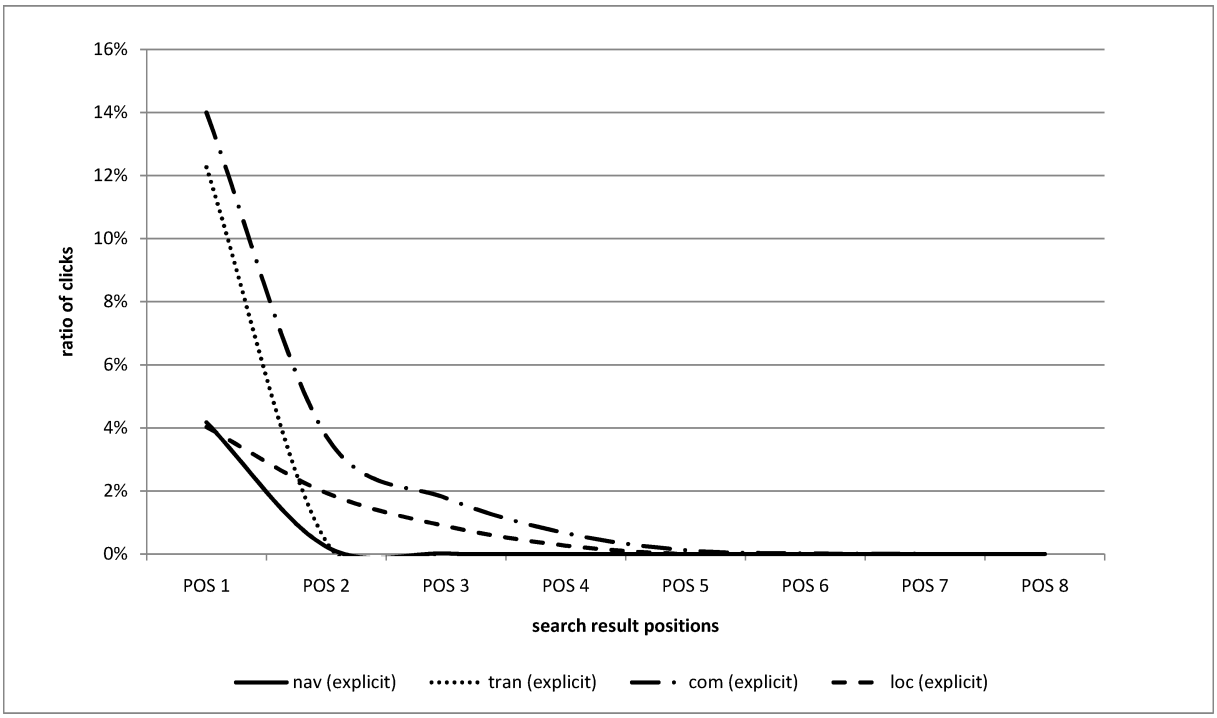

Fig. 6: Ratio of clicks on the ads for navigational (nav), transactional (tran), commercial (com), and local (loc) queries.

Study 3: User Survey

The user survey was initiated to avoid the problem of not knowing the original information need of a searcher. Because we directly asked users what they searching for and therefore gave them the possibility to classify their own queries, we were able to obtain data that can be used to verify the other methods employed to classify search queries.

Table 4 gives an overview of the query types the participants in the online surveys chose. We compared the data to the results from the crowdsourcing study (shown in the right column).

Table 4: Comparison of the query types for the user survey and the crowdsourcing method

Query type	**Ratio of search queries in the online survey (N=549)**	**Ratio of search queries in the sample of the 49,919 queries—only the same queries as in the online survey (N=549)** The range shows the explicit classification (first number) and the not explicit classification (second number)
Navigational	34.61%	68.31-79.60 %
Transactional	78.14%	31.33-64.85 %
Commercial	30.78%	24.77-44.62 %
Local	6.74%	4.01-9.29 %
Informational	38.62%	8.00%

We compared the query types for the same search queries of both samples. The transactional queries stand out because they reach a very high ratio in the online survey. In the classification task done through crowdsourcing, the navigational queries accounted for the majority of the queries.

The great differences in the frequency distribution of the query types in both samples led us to a more detailed analysis of the data. We investigated how often the human jurors and the participant of the online survey matched in choosing the same query types. Here, we found that only 11% are rated exactly the same.

We also examined which query type had the highest matching between human juror and online survey participant. Here we identified local queries, which show the highest conformity with a ratio of 85.06%. The remaining query types did not reach such a high conformity between the participant from the user survey and the crowdsourcing juror. The navigational queries only match with a ratio of 44.99%, the commercial queries with a ratio of 60.47%, and the transactional ones with a ratio of 62.11%.

A further question was if the participants chose the same query types for the same queries. In all, only 75 queries were used more than once in the user survey. With this data, we analyzed the conformity between the participants. The evaluation shows that the participants in most cases did not agree about the query type although they searched with the exact the same query. Only seven participants agreed completely when choosing a query type. However, it should be mentioned that in the user survey, there was only a relatively low number of queries classified by more than one participant (due to the relatively low number of participants).

The analysis reveals different and inconsistent results. Therefore, we decided to test the data of the online survey for plausibility by means of different criteria. As one criterion, we checked the free answer in question three of the user survey ("The perfect result for my search would be…"). If the answer to this question shows that the participant applied the questionnaire not to his or her single query in question, but instead to his or her searches as a whole, we declare this data as not plausible. Furthermore, we describe answers to the user survey as not plausible when the query is not classified as navigational even though the search term contained explicit hints on a navigational query, such as ".de", ".com" or "www." We also analyzed the data set of the survey with regard to queries that are with a high plausibility navigational because they contain domain names, but that were not labelled as navigational. As a last criterion, we checked queries that do or do not suggest a commercial interest, but were not labelled accordingly. All criteria used are summarized in Table 5, which also gives some examples.

Table 5: Criteria applied to the data of the user survey to check the plausibility

Criteria	**Example**
Free answer (question three) All data that show that the participant applied the questionnaire to his or her searches as a whole (instead of to his or her current query).	Example for a free answer: *To select intelligently*
Identification of a domain All queries that were not classified as navigational, although the search term contained	*Facebook.com*

explicit hints on a navigational query, such as ".de," ".com," or "www."	
Domain name Queries that are with a high plausibility navigational as they contain domain names, but were not labelled as navigational.	*Yahoo, Google, or Myspace*
Commercial interest Queries that lead to a/do not lead to a commercial interest, but were not labelled according to that.	Example for a commercial interest: *eBay* Example for a non-commercial interest: *Myspace*

The examination of the plausibility shows that at least one of the criteria can be applied to 75.77% of the queries. When the first criterion remains unaccounted, we have a ratio of 47.36% of plausible answers. In a next step, we checked to see if the remaining 260 plausible answers now corresponded to the queries that were rated in the classification task through crowdsourcing. The result can be seen in Table 6. We see an unchanged high ratio of transactional queries. Also, the ratio of informational queries is still twice as high as in the sample of the queries rated through crowdsourcing.

Table 6: Comparison of the query types exposed through user survey after testing the plausibility and crowdsourcing

Query type	**Ratio of query types after testing for plausibility** (N= 260)	**Ratio of query types for the same queries, rated through crowd sourcing** (N= 260)
Navigational	56.15%	63.46%
Transactional	73.46%	57.31%
Commercial	27.31%	41.15%
Local	4.62%	11.54%
Informational	38.85%	16.54%

Discussion

In this section, we first discuss the overall results, and then the click-through-data, the inter-rater agreement, and the survey results.

Compared to Broder's (2002) results, in our classification study, the range of the navigational queries (27 to 42%) is much higher than the number of queries that Broder allotted to this class (24.4 to 26.4%). The range of transactional queries is between 11 and

39%. Broder allocated 23.8 to 32.3% of the queries to the class “transactional” (2002). We are aware that our study is not directly comparable to Broder’s due to the different data used and the different points in time when the studies were conducted, but the comparison gives a first impression on possible problems in collecting such data.

The fact that the data from the different data collection methods used do differ considerable may result from the basic classification scheme used. While Broder’s distinction between informational, navigational and transactional queries is now a standard for classifying queries according to intent, these basis concepts may not be as evident to search engine users, as researchers assume. To our knowledge, no research thoroughly testing users’ understanding of these basic query types has yet been conducted.

We think that analyzing click-through data is an adequate technique for identifying navigational, as well as commercial queries. In this finding, we are at one with the published literature (e.g., Joachims, 2002). Our advanced hypothesis that for navigational queries, the majority of clicks are made on the first search result, could be confirmed; only the assumption concerning the commercial queries must be limited. We can assume a commercial interest when we have an increased ratio of clicks on the ads, but when this ratio increases up to 70% or more, the probability that the query is commercial decreases. Therefore, we conclude that in this case, other query types cause this heightened ratio of clicks on ads. An examination of the classification for this click ratio shows that the portions of the different query types converge to the average classification for the sample of 24,807 queries. Admittedly, the population is only 414 search queries in which the ratio of clicks on ads is up to 70% or more.

In this context, it should not be disregarded that we cannot be sure if the intention and therefore the type we assign to a query using click-through data is correct, since we did not ask the initiator of the query about his or her information need. We confront this objection with a reference to different studies that show that click-through data can be used as feedback

from the users and therefore also contains a reference to the search intention (cf. Guo, Liu, & Wang, 2009; Dou, Song, Yuan, & Wen, 2008; Joachims, 2002; Macdonald & Ounis, 2009).

Inter-rater Agreement

In our classification task, we found that for many queries, raters were at odds. This, together with the results from the crowdsourcing study, shows that in query intent classification, we face the problem of low inter-rater correlation. The span of non-explicit and explicit classifications in the several classes may be an indicator of the insecurity of the human jurors. This would mean that human jurors need better guidance for a task like this, perhaps training with films and the chance to ask questions.

To see if the jurors were right with their classifications, the search queries were used in the click-through data analyses and the classification by humans was checked. Significant fraction defectives for the navigational and commercial queries were discovered there, which we will discussed in the next section.

We believe the classification by human jurors could be more adequate if some things are borne in mind. The human jurors need to be educated in the classification of search queries. They should have the opportunity to ask questions during their work. This avoids mistakes in classifications, may make them aware of a substantial span in the several classes, and generates a clear result. Also, they should be given additional information (such as click-through and session data).

Another possibility would be to use more jurors. For example, if three jurors were used, a majority decision could be reached. There is also the option to derive multiple query intents from the data when multiple jurors are used and to weight these judgments and use probabilities for the intents of each query. However, this should not distract from the fact that many queries are ambiguous or even non-classifiable. Therefore, just increasing the number of jurors is not sufficient.

Survey Data

The survey has some limitations. The query type "local" could only be chosen when the participant of the online survey also chose the query type commercial. This could be one reason for the small ratio for this query type. Another limitation is that the numbers for the informational queries are also not comparable because the human raters of the crowdsourcing did not knowingly choose this query type.

The user survey on the T-Online search portal lead to quite different results than those of the crowdsourcing study. This could lead to quite different conclusions. On the one hand, we can infer that the use of crowdsourcing is not adequate to classify search queries. That would mean that the human raters could reduce the correct query type from a search query alone. This assumption would lead to the conclusion that humans cannot detect the query type of a "foreign" search query, or that the human raters were not sufficiently trained to classify the queries. On the other hand, we can infer that the chosen method in the form of an online survey is not adequate to collect query types. This assumption is supported by the different, partial contradictory analysis. We designed the online survey to be self-explanatory, which was confirmed in pilot testing. However, it could well be that while our pilot testers did read the instructions carefully, the actual users did not.

Another possibility which backs the conclusion of the online survey not being the adequate method to collect query types is that the results also could reflect that the searchers may not know consistently what they are looking for when they start a search and want the search engine to give them inspiration. If this assumption were correct in at least some cases, it would be an explanation why users were not able to describe their query intent correctly. However, from our analysis, we saw that even in some obvious cases, users did classify their own queries incorrectly. Perhaps the participants did not understand the questionnaire because the questions were too complicated or confusing. Then, our results also question the validity of the results from Broder's study, since he used a similar approach.

As the online questionnaire was displayed directly after a query was entered (i.e., the invitation to enter the survey was displayed on the search engine result page), we assume that issues of incorrectly recalling the query intent will not have influenced the answers given.

This investigation lead us to the assumption that the more people one asks about a query type of a search, the more different opinions and therefore query types one will get. However, this might not be due to mistakes, but to the many possible information needs a query can express. When deciding upon the query type that should be used for further investigation, one could follow the approach of Huffman and Hochster (2007), who used multiple jurors to determine a majority vote. However, when using multiple jurors and considering that queries could have multiple meanings, one could also use probabilities for different query types, as indicated by the votes of the jurors.

Conclusion

In this section, we answer our research questions and make some suggestions about further research. The first research question was, how reliable is user-based query intent classification? As has been shown, neither the crowdsourcing approach using jurors who classified queries originating from other users, nor the questionnaire approach using searchers who were asked about their own queries that they had just entered into a web search engine, lead to satisfying results.

Results from the user survey show that users have difficulty understanding query classification tasks. While we had different users at different points in time (and the results are therefore not directly comparable), we think that it is highly probable that the users in Broder's 2002 study had similar problems, and therefore that the results are questionable. Speaking more generally, we can say that from all our findings, we can see that letting non-expert users classify queries according to their intents is highly error-prone, and we therefore suggest using expert jurors. Additionally, clear instructions should be given, and several

jurors should be used, first to measure the inter-rater agreement, and second to identify the queries for which a classification based on the query alone is not feasible.

We were surprised how different the results from the three studies conducted were, as we had expected a more consistent picture. Considering the differences, we cannot give a clear recommendation on which approach to use when classifying query intents. While the automatic approach performs well on navigational queries (and to some degree, on commercial queries), the crowdsourcing approach and the online survey lead to mixed results. When using one of these approaches, reliability checks should be applied to avoid misclassified queries.

Regarding the ratio of the different query types in the query logs of the T-Online search portal (research question 2), we found that navigational queries account for the largest number of queries (27 to 42%), while informational queries account for 22%, and transactional queries for 11 to 39%. The data, however, shows that from the crowdsourcing data, we cannot give exact numbers, but rather rough estimates. This results from jurors being at odds.

The third research question about whether it is possible to derive navigational intent from simple click-through data can be answered with a clear yes. Our click-through experiment shows that the larger the ratio of clicks on the first result presented, the larger the possibility that this query is navigational. Our crowdsourcing results are robust enough to support this claim.

The results for the commercial intent derived from the ratio of clicks on the ads section of the search engine results page (research question 4) do not give such a clear picture, but also show that the higher the ratio of clicks on the ads, the higher the possibility that a query is commercial. However, while we reach a possibility of 80% for the navigational queries, for the commercial queries, we only reach a prediction reliability of 52%. Also, this value decreases when considering a click ratio of 70% or more on the ads.

The last research question ("How reliable are automatic classifications of user intents?") is hard to answer precisely. The results of our studies leave us in doubt as to whether user intents can be reliably derived by asking users to classify search engine queries without using additional information. We found that even when using the same dataset, results for the classification tasks differ considerably, depending on the methods used. This leads us to conclude that there is little understanding of the classification tasks (which was can be seen from the survey data). We assume that even though jurors in the crowdsourcing task were given detailed instructions, they at least to a certain degree misunderstood the tasks. We showed that human jurors are unreliable in producing exact results.

This leads us to the conclusion that while *automatically* classifying queries according to their intent is certainly useful, current approaches suffer from an unreliable basis in human judgments. When the human judgments are at least to a certain degree not trustworthy, how trustworthy can the automatic classification be if it corresponds to, say, 80% of the human judgments? Therefore, we must question the success of approaches that use automatic classification and compare its performance to a baseline from human jurors. Query classification, although at first seemingly straightforward, seems to be a problem of inter-rater agreement.

Further research should focus on improving the reliability of human classification. We suggest classifying large data sets that can be used as baseline sets for automatic classification and bearing the following recommendations in mind: (1) Use multiple jurors and derive multiple query intents from the data. Weight these judgments and use probabilities for the intents of each query. As jurors might disagree on the query intents, it may be useful to further ask for the reasons for such disagreement. (2) Use expert jurors and give them clear instructions and the possibility to raise a query about the classification task, as well. (3) The questionnaire or the instructions that the jurors use to classify the queries should be very detailed and perhaps also contain "traps" to detect decisions which were not made properly.

References

Ashkan, A., & Clarke, C. L. A. (2009). Term-based commercial intent analysis. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 800–801). New York: ACM.

Baeza-Yates, Ricardo, Calderón-Benavides, L., & González-Caro, C. (2006). The intention behind web queries. In F. Crestani & M. Sanderson (Eds.), *String processing and information retrieval* (Vol. 4209, pp. 98-109). Heidelberg: Springer.

Bar-Ilan, J., Keenoy, K., Yaari, E., & Levene, M. (2007). User rankings of search engine results. *Journal of the American Society for Information Science and Technology*, *58*(9), 1254-1266.

Broder, A. (2002). A taxonomy of web search. *ACM Sigir forum*, *36*(2), 3-10.

Calderon-Benavides, L., Gonzalez-Caro, C., & Baeza-Yates, Ricardo. (2010). Towards a deeper understanding of the user's query intent. In *SIGIR 2010 Workshop on Query Representation and Understanding* (pp. 21-24). New York: ACM.

Chao, L., Guo, F., & Wand, Y.-M. (2009). Efficient multiple-click models in web search. *Proceedings of the Second International Conference on Web Search and Web Data Mining* (pp. 124-131). New York: ACM.

Church, Karen, & Smyth, B. (2009). Understanding the intent behind mobile information needs. In *13th International Conference on Intelligent User Interfaces* (pp. 247-256). New York: ACM.

ComScore. (2010). comScore reports global search market growth of 46 percent in 2009. Retrieved from http://comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009.

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the International Conference on Web Search and Web Data Mining* (pp. 87-94).

Croft, B., Bendersky, M., Li, H., & Xu, G. (Eds.). (2010). Query representation and understanding: Workshop of the 33rd Annual International ACM SIGIR Conference on research and development in information retrieval. Retrieved from http://ciir.cs.umass.edu/sigir2010/qru/QRU_proceedings.pdf

Dai, H. K., Zhao, L., Nie, Z., & Wen, J. R. (2006). Detecting online commercial intention (oci). In *Proceedings of the 15th International Conference on World Wide Web* (p. 829–837). New York: ACM.

Dou, Z., Song, R., Yuan, X., & Wen, J.-R. (2008). Are click-through data adequate for learning web search rankings? In *Proceeding of the 17th ACM conference on Information and knowledge management* (p. 73-82). New York: ACM.

Frants, V. I., Shapiro, J., & Voiskunskii, V. G. (1997). *Automated information retrieval: theory and methods*. *Library and information science*. San Diego: Academic Press.

Griesbaum, J. (2004). Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. *Information Research*, *9*(4).

Guo, F., Liu, C., & Wang, Y. M. (2009). Efficient multiple-click models in web search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (p. 124–131). New York: ACM.

He, D., Göker, A., & Harper, D. J. (2002). Combining evidence for automatic web session identification. *Information Processing & Management*, *38*(5), 727-742.

Huffman, S. B., & Hochster, M. (2007). How well does result relevance predict session satisfaction? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (p. 567–574). New York: ACM.

Höchstötter, N., & Koch, M. (2009). Standard parameters for searching behaviour in search engines and their empirical evaluation. *Journal of Information Science*, *35*(1), 45.

Höchstötter, N., & Lewandowski, D. (2009). What users see—Structures in search engine results pages. *Information Sciences*, *179*(12), 1796-1812.

Jansen, B J, Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing and Management*, *44*(3), 1251-1266.

Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on Web search engines. *Journal of the American Society for Information Science and Technology*, *58*(6), 862–871.

Joachims, T. (2002). Optimizing search engines using click-through data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 133–142). New York: ACM.

Kang, I. H. (2005). Transactional query identification in web search. In: *Information Retrieval Technology*, *LNCS 3689*, (pp. 221–232). Heidelberg: Springer.

Kang, I. H., & Kim, G. C. (2003). Query type classification for web document retrieval. In *Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 64-71). New York: ACM.

Kantor, P. B. (1976). Availability analysis. *Journal of the American Society for Information Science*, *27*(5-6), 311-319.

Kanoulas, E., Clough, P., Carterette, B., & Sanderson, M. (2010). Session track at TREC 2010. *SIGIR Workshop on the Simulation of Interaction* (p. n.a.). Retrieved from http://kanoulas.staff.shef.ac.uk/research/mypapers/sigir10e.pdf

Kathuria, A., Jansen, Bernard J., Hafernik, C., & Spink, Amanda. (2010). Classifying the user intent of web queries using k-means clustering. *Internet Research*, *20*(5), 563-581.

Lee, U., Liu, Z., & Cho, J. (2005). Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web* (p. 391–400). New York: ACM.

Lewandowski, D. (2006). Query types and search topics of German Web search engine users. *Information Services & Use*, *26*, 261-269.

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: Considering results descriptions. *Journal of Documentation*, *64*(6), 915-937.

Lewandowski, D. (2011). The retrieval effectiveness of search engines on navigational queries. *ASLIB Proceedings*, *61*(4), 354-363.

Li, H.; Xu, G.; Croft, B. Bendersky, M. (eds.): Proceedings of the Query Representation and Understanding Workshop held at SIGIR 2011. Retrieved from: http://ciir.cs.umass.edu/sigir2011/qru/proceedings-qru2011.pdf

Lu, Y., Peng, F., Li, X., & Ahmed, N. (2006). Coupling feature selection and machine learning methods for navigational query identification. In *International Conference on Information and Knowledge Management, Proceedings* (pp. 682-689).

Macdonald, C., & Ounis, I. E. T.-F. (2009). Usefulness of Quality-through Data for Training. In *Proceedings of the 2009 Workshop on Web Search Click Data* (pp. 75-79). New York: ACM.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, *49*(4), 41–46.

Mendoza, M., & Baeza-Yates, R. (2008). A web search analysis considering the intention behind queries. In *Proceedings of the Latin America Web Conference* (pp. 66-74).

Mendoza, M., & Zamora, J. (2009). Identifying the intent of a user query using support vector machines. In *String Processing and Information Retrieval* (Vol. 5721, p. 131-142). Heidelberg: Springer.

Pitler, E., & Church, Ken. (2009). Using word-sense disambiguation methods to classify web queries by intent. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3* (p. 1428-1436). Association for Computational Linguistics.

Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web* (p. 13–19). New York: ACM.

Singer, G., Norbisrath, U., Vainikko, E., Kikkas, H., & Lewandowski, D. (2011). Search-Logger -- Tool Support for Exploratory Search Task Studies. *SAC2011* (pp. 751-756). New York: ACM.

Spink, A. (2004). *Web search: public searching on the Web*. Dordrecht: Kluwer Academic Publishers.

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, *52*(3), 226–234.

Véronis, J. (2006). A comparative study of six search engines. Retrieved March 18, 2011, from http://sites.univ-provence.fr/veronis/pdf/2006-comparative-study.pdf.

White, R. W., Bailey, P., & Chen, L. (2009). Predicting user interests from contextual information. In *32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 363-370). New York: ACM.

Appendix: Instructions given to the participants of the crowdsourcing study

Dear ClickWorker,

Thank you for taking the time to read this information leaflet.

We are working for a large Web portal. The client is interested in the types of search queries its users enter into the search boxes on its websites.

We will ask you to answer four yes/no questions.

The questions about the search queries are not self-explanatory. Please read these instructions carefully before starting the task.

Please note: If a certain search query is totally unknown to you or impossible to understand, please skip it.

Explanations of the Four Yes/No Questions

1. Is the user searching for an individual website?

Users usually aim for one of two types of results:

- Type A: The users want a **specific website that they already know about** (in this case, please answer "yes")
- Type B: The users want to see **a number of Web pages** that match their query (in this case, please answer "no")

Type-A search queries often contain **parts of an Internet address**. For example, they begin with **www** or end with **.de, .com,** etc. Some type-A queries do not contain such hints, but the user is nonetheless searching for a certain website (e.g., "postbank," "stiftung warentest"). In these cases, only one website will satisfy the query.

Type-B queries often refer to a general topic, not an individual website. For example, this holds true for queries referring to **cities, countries, people, or sports**. A user expects a variety of results, and these results are **not known before conducting the search**.

Note: Most likely, you will not immediately recognize when a query refers to a specific website. In some cases, you will need to do some research on the query.

Positive examples for queries referring to certain websites:

eBay, amazon.de, sparkasse Dortmund

Negative examples:

berlin, football, angela merkel

2. Is the query transactional?

We refer to a search query as **transactional** if, after using the search engine, the next step is obviously or probably to perform some kind of **Web-mediated activity** that goes beyond collecting information. Web-mediated activities include, for example, a **purchase**, a **sale**, **lending** or **ordering a product**, **booking** a flight, or the **download** of software from the Internet. Other examples include **chatting** over the net, **online gaming**, **online banking**, and voice chatting over IP **telephony**.

Note: **Looking at pictures or videos on websites like YouTube or Clipfish is not considered transactional.**

Queries leading to websites where Web-mediated activities should be conducted are also considered transactional. Examples include Amazon.de and eBay. However, the name of a product does not satisfactorily indicate a transactional query.

Positive examples:

car rental, chatserver, eBay, free games, skype, buy books, amazon.de

Negative examples:

YouTube, Clipfish, Audi A5

3. Does the search query have a commercial intent?

A query has a commercial intent when it is very probable that the user will conduct a fee-based activity after leaving the search engine, such as **buying something, lending something, or downloading software for a fee**. This activity does not need to be Web-based.

When a user searches for a **product or service**, the query's intent is commercial, even if this product or service cannot or will not be obtained on the Internet. Additionally, searches looking for websites where commercial products are sold or where commercial transactions are prepared are considered to have a commercial intent. Examples include mail-order houses like **quelle.de** and real estate portals like **immobilienscout.de**. Queries for certain Web applications, such as online banking, are not relevant here.

Positive examples:

aldi, alice-dsl.de, pharmacy online, hotel, hairdresser

Negative examples:

online banking, route planner

4. Does the query have a local intent?

A query is considered to have a local intent when it obviously refers to **a town, a region, or a country**, or if it asks for a local shop or service. A query does **not** have a local intent **if it refers solely to a continent** (Africa, America, Asia, Australia, or Europe).

Note: You may not recognize each town's name immediately, so you may need to do some research before answering this question.

Positive examples:

aachener newspaper, berlin adressbuch, aok niedersachsen, adelboden, dentist dresden, hairdresser tübingen, hotel berlin

Negative examples:

news, africa, hairdresser