Survey and evaluation of query intent detection methods

Article · January 2009

DOI: 10.1145/1507509.1507510

CITATIONS
44

READS
1,000

3 authors, including:

David J. Brenes

8 PUBLICATIONS **139** CITATIONS

SEE PROFILE

Daniel Gayo-Avello

University of Oviedo

74 PUBLICATIONS **1,524** CITATIONS

SEE PROFILE

Survey and evaluation of query intent detection methods

David J. Brenes
Indigo Group
C/Campoamor 28 1º Oficina 5
33001 Oviedo (SPAIN)
+34 985 207 746
david.brenes@indigo.es

Daniel Gayo-Avello
University of Oviedo
Despacho 57, Edificio de Ciencias
C/Calvo Sotelo s/n 33007 Oviedo
(SPAIN)
+34 985 104 340
dani@uniovi.es

Kilian Pérez-González
University of Oviedo
(SPAIN)
i9433245@petra.euitio.uniovi.es

ABSTRACT

User interactions with search engines reveal three main underlying intents, namely *navigational*, *informational*, and *transactional*. By providing more accurate results depending on such query intents the performance of search engines can be greatly improved. Therefore, query classification has been an active research topic for the last years. However, while query topic classification has deserved a specific bakeoff, no evaluation campaign has been devoted to the study of automatic query intent detection. In this paper some of the available query intent detection techniques are reviewed, an evaluation framework is proposed, and it is used to compare those methods in order to shed light on their relative performance and drawbacks. As it will be shown, manually prepared gold-standard files are much needed, and traditional pooling is not the most feasible evaluation method. In addition to this, future lines of work in both query intent detection and its evaluation are proposed.

Categories and Subject Descriptors

H.2.8 **[Database Management]**: Data Mining; H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval; H.3.4 **[Information Storage and Retrieval]**: Performance evaluation (efficiency and effectiveness); H.3.5 **[Information Storage and Retrieval]**: Web-based services

General Terms

Algorithms, Measurement, Performance, Experimentation, Human Factors.

Keywords

Click-through data, Web search behavior, MSN Query Log, Query intent detection, Evaluation.

1. INTRODUCTION

Query classification has been an active research topic for the last years, even deserving an edition of the ACM's KDD Cup [20]. There exist two main "dimensions" in which query classification has been usually performed: *"topic"* and *"intent"*.

WSCD'09, February 9, 2009, Barcelona, Spain.

Query topic classification consists of identifying a query as belonging to one or more categories from a predefined set (e.g. assigning the query `'the gold rush charlie chaplin'` to the category `'Entertainment/Movies'`). To this extent, several query topic taxonomies have been described (e.g. [3, 4, 7, 20, 24, 26, 27, 29]). All of them show some commonalities (e.g. they are highly similar to the structure usually seen in Web directories) but the number of categories and subcategories varies widely. Thus, Spink *et al.* [29] proposed 12 categories; Pu, Chuang and Yang [24] described a taxonomy with 15 categories and 85 subcategories; the taxonomy depicted by Li, Zheng and Dai [20] has got 67 categories; and Broder *et al.* [7] described a hierarchy with 6,000 entries! Nonetheless, as it was said, this particular task was addressed within the KDD Cup 2005 and, hence, there exists an abundant literature describing state of the art methods to perform query topic classification.

On the other hand, query intent classification consists of identifying the underlying goal of the user when submitting one particular query. For instance, a user issuing the query `'apple store'` could be trying to reach `http://store.apple.com` while a user submitting `'telegraph history'` is most likely interested in finding information on that topic but not concerned about the particular website to solve that information need.

With regard to such users intents there exists broad consensus as most of the researchers rely on the taxonomy proposed by Broder [6] and refined by Rose and Levinson [25]. According to Broder, search queries reveal three types of user intents: (1) *"navigational"* (the user wants to reach a particular website), (2) *"informational"* (the user wants to find a piece of information on the Web), and (3) *"transactional"* (the user wants to perform a web-mediated task). Rose and Levinson further improved that classification by introducing subcategories for both informational and transactional queries. Later, Jansen, Booth and Spink [13] provided a comprehensive and integrative view of the different query intent taxonomies proposed in the literature.

Therefore, several methods have been proposed to automatically classify queries according to their intent (e.g. [13, 15, 17, 19, 21]). However, in contrast to query topic classification, no "bakeoff" has been devoted to the comparison of different query intent classification methods and, hence, the research community simply does not know the relative performance of the different techniques, nor their respective drawbacks.

We have recently developed some work on the detection of navigational queries [5] and, thus, our research proposal for this *"Workshop on Web Search Click Data"* consisted of replicating most of the proposed techniques to perform automatic query

intent detection in addition to study the feasibility of a pooling strategy *à-la-TREC* to evaluate the different techniques.

To achieve such objectives, several intent detection methods have been replicated and applied to the MSN Query Log [22]. Then, they were evaluated against a manually labeled sample extracted from the same dataset. As it will be later discussed, such an evaluation led us to conclude that pooling is neither feasible nor adequate to evaluate query intent detection methods.

The rest of the paper is structured as follows. Section 2 discusses most of the techniques to perform query intent detection. The research questions guiding this study are stated in Section 3. Then, in Section 4 we described both the details of the replicated techniques and the evaluation method eventually applied. In Section 5 the results achieved by each technique are shown. Finally, Sections 6 and 7 conclude this paper and introduce future lines of work.

2. LITERATURE REVIEW

As it has been previously mentioned, researchers agree on three different intentions driving user queries, namely, *informational*, *navigational* and *transactional*. Different intents require different answers from the search engine; thus, automatically identifying such query intents has been an open research topic since the publication of Broder's taxonomy. This section reviews the state of the art on such query intent detection methods.

Kang and Kim [15] proposed four different methods to determine whether a web search query was informational or navigational (*topic relevance* and *homepage finding* in their own terminology). Such methods could be used alone but when combined they achieved the best results (91.7% precision and 61.5% recall according to the original authors). Three of the methods require training collections. So, Kang and Kim employed WT10g[1] to build two subsets: DB_{HOME} and DB_{TOPIC}. The first one, DB_{HOME}, comprised those documents acting as entry points for a particular website while the remaining web pages from WT10g were assigned to DB_{TOPIC}. From such subsets it was possible to find out (1) the frequency of appearance of each term in both subsets, (2) the mutual information of term pairs in both subsets, and (3) the frequency with which each term appears in anchor texts and page titles.

In addition to those three information sources they included a fourth method relying on POS tagging: every query containing a verbal form (except for the verb *'be'*) was considered a topic relevance task (i.e. an informational query). All of these sources of evidence were linearly combined and, thus, to obtain the parameters a training subset was necessary.

Lee *et al.* [17] revisited the problem of telling apart navigational queries from informational ones without considering the third class by Broder (i.e. transactional). To automatically determine the query intent they relied on two different data sources: click-through data, and anchor texts. From click-through data they computed the click distribution for each query. When such distribution is highly skewed towards one or just a few domains it can be assumed that the query is navigational. In contrast, when the click distribution is relatively flat, an informational intent can be supposed.

In addition to compute click distributions, click-through data was also employed to find out the average number of clicks for each query. That information is highly relevant because navigational queries are usually associated with fewer clicks than informational ones.

To compute both features (click distribution and average number of clicks per query) each query requires an important amount of prior data. Lee *et al.* proposed an alternative source of information when such click-through data is unavailable or sparse: the so-called anchor-link distribution. Such distribution is very similar to the click-distribution but it is computed from a collection of web pages. The main assumption behind the anchor-link distribution is that navigational queries commonly appear as anchor texts linking to a few domains, while anchors containing informational queries exhibit a much greater variety of URLs.

All of these methods rely on *'ad hoc'* thresholds and parameters and, therefore, some researchers have tried machine learning techniques. For instance Nettleton *et al.* [23] and Baeza-Yates *et al.* [1] applied different clustering techniques to classify users and queries. Thus, Nettleton *et al.* employed Self-Organized Maps to classify user sessions (not queries) into the three aforementioned classes: informational, navigational and transactional. On the other hand, Baeza-Yates *et al.* employed SVMs and PLSA to cluster queries according to their intent. It must be noticed, however, that they employed three categories different from the commonly used: *informational*, *not-informational* and *ambiguous*. In addition to this, classifiers were not actually trained on the queries but on the contents of the clicked documents.

Liu *et al.* [21], as Lee *et al.* [17], exploited click-through data to find out the query intents. According to these researchers click-through data is a good source of information because, when using sufficiently large logs, there is prior information for about 90% of the queries. They also employed anchor texts but, according to them, even when relying on huge collections (about 200 million documents) less than 20% of the queries appear as anchors. These researchers, as many of the previous ones, also focused in the task of separating navigational queries from informational/ transactional ones. To perform such task they applied two sources of evidence: *n Clicks Satisfied (nCS)* and *top n Results Satisfied (nRS)*.

The first value, *nCS*, is just the proportion of sessions containing a given query in which the user clicked, at most, *n* results. The underlying assumption for such value is that users issuing navigational queries click on fewer results than users submitting informational or transactional queries. Hence, when using a small *n* value (e.g. 2 clicks) navigational queries would exhibit larger *nCS* values than informational/transactional queries.

With regard to the second value, *nRS*, it is based on the assumption that users submitting navigational queries tend to click on the top results. Thus, *nRS* is just the proportion of sessions containing a given query in which the user clicked, at most, the top *n* results. As was the case with *nCS*, navigational queries exhibit higher *nRS* values.

[1] http://ir.dcs.gla.ac.uk/test_collections/wt10g.html

In addition to *nCS* and *nRS*, Liu *et al.* also employed click-distributions (as proposed in [17]). To combine these three sources of information they computed a decision tree.

Jansen *et al.* [13] proposed a quite different approach. Firstly, their method only relies on the queries, that is, it does not exploit click-through data. Secondly, the method consists of a number of easily implementable rules to determine the intent of each query.

The approach by Tamine *et al.* [30] combines not only the query features, but also the query context, to find out the probability for each of the three different intents. They applied most of the query features previously used in the literature: query length, use of verbs, use of transactional terms (such as *"download"* or *"buy"*), and the terms usage rates in both anchor texts and page titles (i.e. their method requires an external document collection). With regard to the query context, it consists of those immediately prior queries with the same intent. Then, that context is compared against (1) the query intent, and (2) the expected features for a session exhibiting the context intent in order to compute the probability that the query actually exhibits the same intent of the context.

Finally, Brenes and Gayo-Avello [5] proposed three sources of evidence distilled from click-through data and somewhat related to the work by [17, 21]. The first coefficient, *weight of the most popular result (cPopular)*, exposes the relative size of the most visited result with regard to the whole set of clicked results for a query (see Equation 1). The second value, *number of distinct visited results (cDistinct)*, consists of dividing the number of distinct clicked results by the total amount of clicks and, then, subtracting that value from one (see Equation 2). The third and last value, *cSession*, requires a prior sessionization of the query log. The underlying assumption for this coefficient is that navigational queries tend to appear isolated and, thus, it is the ratio of one-query sessions to all the sessions containing that query (see Equation 3).

$$cPopular(Query\ \boldsymbol{q}) = \frac{\#Clicks\ on\ top\ result\ for\ query\ \boldsymbol{q}}{\#Clicks\ for\ query\ \boldsymbol{q}} \quad (1)$$

$$cDistinct(Query\ \boldsymbol{q}) = 1 - \frac{\#Distinct\ clicked\ results\ for\ query\ \boldsymbol{q}}{\#Clicks\ for\ query\ \boldsymbol{q}} \quad (2)$$

$$cSession(Query\ \boldsymbol{q}) = \frac{\#One\ query\ sessions\ involving\ query\ \boldsymbol{q}}{\#Sessions\ involving\ query\ \boldsymbol{q}} \quad (3)$$

3. PROBLEM DEFINITION

As it has been shown there exist several methods to perform query intent detection. However, none of these methods have been thoroughly evaluated or compared with analogous techniques. Thus, the main research questions addressed in this study are the following: (1) How could the performance of such methods be evaluated? And, (2) Which are the most appropriate methods to perform query intent detection?

With regard to the first research question we were interested in two different aspects. Firstly, we wanted to provide an evaluation method analogous to that employed in the KDD Cup 2005 [20] devoted to query topic classification[2]. Secondly, we wanted to study the feasibility of pooling [14] as an evaluation method for query intent detection techniques.

Once an evaluation method was provided, the second research question could be directly addressed by just running the different query intent detection techniques on the available query log.

Such query log was kindly provided by Microsoft Research and consists of 15 million queries, submitted to the MSN search engine by United States users, and sampled on May 2006. The dataset provides for each query the following attributes among others: time stamp, query string, and clicked URL (if any). Given that information, most of the previously described methods relying on click-through data should be easily replicated.

4. RESEARCH DESIGN

4.1 Query intent detection methods

Not all the aforementioned techniques were reproduced in this study. Those by Kang and Kim [15] and Tamine *et al.* [30] were left for future work because of the unavailability of the WT10g collection to the authors and the tightness of the workshop deadline. The approaches by Nettleton *et al.* [23] and Baeza-Yates *et al.* [1] were also excluded given that, although related, they are not totally analogous to the rest of the techniques.

Hence, this study replicated the techniques proposed by Lee *et al.* [17], Liu *et al.* [21], Jansen *et al.* [13], and Brenes and Gayo-Avello [5]. All of them, except for the technique by Jansen *et al.*, rely on click-through data and the MSN Query Log was thus employed. The techniques by Lee *et al.* and Liu *et al.* also make use of anchor texts and, to that end, about 1.2 million web pages were collected. Such a dataset comprises the 100,000 most frequently clicked URLs in the MSN Query Log and the remaining documents were obtained by means of Yahoo!'s random URL generator[3].

The technique proposed by Liu *et al*. relied on two coefficients: the fraction of sessions with at most *n* clicks (*nCS*) and the fraction of sessions visiting at most the top *n* results (*nRS*). In order to compute such coefficients suitable values had to be assigned to *n*. For this study, 2 and 5 were employed to find out *nCS* and *nRS*, respectively.

Finally, all of the coefficients proposed by the different authors, except for Jansen *et al.*, range between 0 and 1 and, hence, a 0.5 threshold was applied to label a query as navigational.

With regard to the technique by Jansen *et al.*, it consists of a series of rules depending on the query contents. By means of such rules their method classifies a query into any of the three usual intents. However, because the other techniques just deal with navigational queries we only implemented the rules to detect such intent.

Those rules can be summarized as follows: navigational queries contain names of companies, businesses, organizations, or

[2] To perform the evaluation in the KDD Cup 2005, a random sample comprising 800 queries was manually tagged and used as a gold-standard against which compare each solution.

[3] `http://random.yahoo.com/fast/ryl`

people; they contain domains suffixes; or they have less than three terms. As can be seen, some of the rules require external information and, to that end, several lists of pertinent terms were obtained from Freebase[4] by means of MQL [11] queries (see Figures 1 and 2). That way, we obtained lists of companies (120,000 entries), organizations (37,000 entries), websites (4,000 entries), and people names and surnames (300,000 entries). Figure 3 shows some of such terms.

```
{
  "cursor":true,
  "query":[
    {
      "key":[],
      "name":[],
      "type":"/business/company"
    }
  ]
}
```

Figure 1. A MQL query to retrieve the name and keys of all the companies available in Freebase.

```
{
  "key" : [
    "848",
    "Audi",
    "Audi_AG",
    "audi",
    "Audi_Aktien-Gesellschaft",
    "Audi_Sport"
  ],
  "name" : [
    "Audi"
  ],
  "type" : "/business/company"
}
```

Figure 2. One "record" obtained with the previous query.

```
alsa bus company        craigslist
cajastur                digg
microsoft corporation   john
uc los angeles          william
uk labour party         james
unicef                  moore
blogger                 jackson
```

Figure 3. Some of the terms employed to implement the technique by Jansen *et al.* They are companies, organizations, websites, and people names.

4.2 Proposed evaluation method

One of the objectives of this study was providing an evaluation framework analogous to that previously employed to evaluate query topic classification methods [20]. Thus, precision, recall and balanced *F*-score were to be computed for every solution (see Equations 4 to 6).

To obtain such figures, queries from the MSN Query Log needed to be manually labeled. Needless to say, such task was unattainable for the whole dataset (it contains more than 6 million unique queries) and, thus, a random sample was extracted. For the KDD Cup 2005 800 queries were selected from an original log comprising 800,000; hence, 6,624 queries would provide a similar sample for the MSN Query Log[5].

However, the approach followed in this study was different from that of the KDD Cup 2005. In that bakeoff three editors tagged the sample, each participant system was evaluated against the three different answer sets, and, hence, the performance measures were computed as weighted aggregates.

We, in contrast, divided the 6,624 queries among several editors (10 Computer Science students and professionals, in addition to the authors themselves) in such a way that every query was evaluated by two different persons but every subset was unique. This way, each editor just had to label about one thousand queries which was a much lighter work than labeling the whole sample. Once every query was labeled we checked if both labels were equal and, otherwise, a third editor (one of the authors) resolved the inconsistency. Thus, after a couple of workdays the whole subset was completely tagged according to query intent (i.e. *navigational*, *informational* or *transactional*)[6].

It must be said that the level of agreement between labelers was pretty high. However, neither κ, nor χ^2 figures were computed. This was because for any two given raters the amount of common judgments was well below 10%. Instead, after assembling the final tagged query subset, the performance of every labeler was assessed finding that the average precision and recall were 85% and 79%, respectively. With regard to the worst and best editors (according to *F*-measure), they achieved 0.913 and 0.949 precision, and 0.335 and 0.922 recall, respectively.

Certainly, such figures are far from perfect; however, they are much higher than the average precision achieved by labelers in the KDD Cup 2005 when compared against each other, and, in spite of that, they "agreed" in the three best performance teams for that bakeoff. Therefore, the final gold-standard was considered good enough to perform evaluations and to compare different navigational intent detection methods.

Finally, although precision, recall, and *F*-score are well-known in IR, there exist other available performance measures that could be studied for future evaluations (e.g. inferred average precision [33] or *infNDCG* [2]).

$$P = \frac{\#Queries\ correctly\ tagged\ as\ navigational}{\#Queries\ tagged\ as\ navigational} \quad (4)$$

$$R = \frac{\#Queries\ correctly\ tagged\ as\ navigational}{\#Queries\ tagged\ as\ navigational\ by\ human\ labeler} \quad (5)$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (6)$$

[4] http://www.freebase.com/

[5] With that sample size the error rate at 99% confidence would be 1.59% (assuming the largest standard deviation, i.e. 0.5)

[6] It must be noticed that the distinction between so-called *informational* and *navigational* queries is, in many cases, highly subjective and clearly context- and user-dependent. Thus, the assumption of a clear boundary between those two intents should be considered for future research and, perhaps, a more cognitive approach could be followed (such as in [31]).

5. RESULTS

The results obtained with each of the query intent detection methods are shown in Table 1.

In addition to the techniques described in previous sections, a naïve method based on Monte Carlo simulation was also evaluated. The idea is extremely simple: From the gold standard file the probabilities for navigational, informational, and transactional queries are computed[7] (0.270, 0.631, and 0.099, respectively). Then, for each query 10 random numbers are produced in the range [0, 1]. If the number is equal or below 0.099 a vote for *transactional* is issued; if it is between 0.099 and 0.369 the vote is for *navigational*; and, otherwise, for *informational*. Finally, the most voted label is eventually assigned to the query.

Five methods are highlighted in Table 1; they are the top three achievers (according to *F*-measure) and two techniques that are just 3.74% below the third place. The top achiever techniques are those by Liu *et al.* [21] which mainly rely on the number of clicks and visited results. Their approach is highly related to the *click-distribution* method proposed by Lee *et al.* [17] and their technique even employs it, but, interestingly enough, such method does not show a great performance[8].

Table 1. Performance achieved by every evaluated method. Best performance figures are shown in bold and top achievers are highlighted.

Method	Precision	Recall	F-measure
Lee *et al.* anchor distribution	0.426	0.011	0.022
Lee *et al.* click distribution	0.258	0.087	0.130
Liu *et al. nCS*	0.310	0.484	0.378
Liu *et al. nRS*	0.347	0.467	**0.398**
Liu *et al.* decision-tree	0.292	0.522	0.374
Jansen *et al.* company names	0.218	0.087	0.130
Jansen *et al.* organization names	0.253	0.023	0.042
Jansen *et al.* people names	0.189	0.281	0.226
Jansen *et al.* domains	**0.997**	0.178	0.302
Jansen *et al.* websites	0.306	0.011	0.021
Jansen *et al.* combined	0.272	**0.532**	0.360
Brenes & Gayo *cPopular*	0.345	0.409	0.374
Brenes & Gayo *cDistinct*	0.753	0.034	0.117
Brenes & Gayo *cSession*	0.403	0.180	0.248
Monte Carlo simulation	0.304	0.115	0.167

Another technique by Lee *et al.*, namely anchor-distribution, also shows poor recall (although the precision is pretty good). However, this result is not really surprising given that just 50,000 out of 6.6 million queries appear in the anchor texts collected (a mere 0.76%).

It must be said that such performance results contrast sharply with the claims by Lee *et al.* about 90% accuracy. However, as they pointed up in their paper, their experiment was conducted on a log comprising queries issued from a CS department and, thus, they could be widely different from general user queries. Anyway, we think that click- and anchor-distributions deserve deeper study.

A technique that achieves reasonable performance (unnoticeable[9] differences with regard to the third best achiever) is that proposed by Jansen *et al.* [13]. Actually, this method obtains the best recall figure. This technique relies heavily on external term lists and, thus, it is plausible that tuning such lists (i.e. removing noisy and ambiguous terms) would greatly improve its performance.

With regard to the techniques devised by the authors; one of them, *cPopular*, reached the third place while the other two measures, *cDistinct* and *cSession*, obtained better precision than the average but much lower recall. This is due to the fact that both coefficients require a big amount of click data in order to obtain significant results for a given query. On the other hand, *cPopular*, is not greatly affected by the total number of clicks and, thus, achieves much higher recall.

Hence, three different methods relying on very different sources of evidence are the best achievers. As many other researchers (e.g. [17, 21]), we were also interested in the performance that could be achieved by combining different techniques. Therefore, every conceivable combination of the aforementioned techniques was evaluated and the best results are shown in Table 2.

As can be seen, it is possible to greatly improve the *F*-measure (about 18-19%) by just mixing the results from two or more techniques. Besides, such simple approach retrieves about 60% of the navigational queries with near 40% precision. In addition to this, those results also show the most promising features: *nRS*, using domains and websites names, and, rather surprisingly, Monte Carlo simulation.

Table 2. Best combined methods.

Methods	Precision	Recall	F-measure
nRS, anchor distribution, websites, domains, *cDistinct*	0.395	0.593	**0.474**
nRS, websites, domains, *cDistinct*	0.396	0.589	**0.474**
nRS, domains, *cDistinct*	0.397	0.583	0.473
nRS, websites, domains	0.396	0.585	0.472
nRS, anchor distribution, domains	0.396	0.583	0.471
nRS, domains, *cPopular*, *cDistinct*, *cSession*	0.397	0.579	0.471
Monte Carlo, *nRS*, domains	0.375	**0.634**	0.471
nRS, domains	0.397	0.579	0.471
cPopular, domains	**0.404**	0.525	0.457

6. DISCUSSION

We now return to the research questions previously stated. (1) How could the performance of query intent detection methods be evaluated? And, (2) Which are the most appropriate techniques to perform query intent detection?

With regard to the second question, it seems that by combining several sources of evidence such as click-through data, external

[7] Certainly this should be considered "cheating" as the system is to be evaluated on the training data; however, this way it is possible to know the "topline" performance of Monte Carlo.

[8] About 41.5% of the queries in the MSN Query Log do not have any associated click; this could explain such poor performance.

[9] The author is applying the criterion proposed by Spärck-Jones [28] that performance differences lesser than 5% should be disregarded, those in the 5-10% interval are "noticeable", and "material" only those greater than 10%

knowledge, and probabilities inferred from manual labeled data it could be possible to obtain pretty good results.

With regard to the first question, a simple approach based on a reasonable, whilst still relatively small, manually labeled sample can depict the performance of the evaluated techniques.

Nevertheless, we were interested in the feasibility of applying a pool-based evaluation method [14]. Such a method consists of assembling the list of items to be manually labeled as relevant or irrelevant from those detected by the participant systems. In this case, the queries tagged as navigational by each method would comprise the "pool" to be manually edited.

Such strategy has two important problems. First, the pool would not contain every navigational query that a human editor could find. In fact, by taking all the responses of the aforementioned techniques just 81% of the actual navigational queries were found. This way, by evaluating just on the basis of queries flagged as navigational by any of the participant methods the performance measures would be misleadingly high.

The second problem is that most of the queries are flagged as navigational by one method or another. Indeed, even assuming the loss of 19% relevant items, the human editors would have to label nearly 79.7% of the available queries (i.e. 5.3 million queries for the MSN Query Log!)

Arguably, instead of providing just labels for the queries, each system could provide a weight (in fact, most of the described methods produce such output) and, therefore, a pool could be constructed from the top most reliable results. However, we feel that such an approach would only deepen the problem of simply evaluating the systems on the "easiest" items.

Consequently, it seems that the evaluation of the kind of systems depicted in this study would have to rely on manually labeled samples. So, it should be studied the possibility of collecting much larger labeled samples by means of crowdsourcing (e.g. using Amazon Mechanical Turk[10] such as in [16]). Moreover, the work by Buckley *et al.* [8] should shed some light on the issue.

7. IMPLICATIONS AND CONCLUSIONS

The performance of search engines can be greatly improved by providing more accurate results depending on the query intent [15]. Consequently, there have been several works in the field of automatic query intent detection (e.g. [5, 13, 15, 17, 21, 30]). Nevertheless, in contrast to query topic classification (cf. [20]), no bakeoff has been devoted to the evaluation and comparison of such techniques.

Thus, this study contributes to our understanding of this problem in several ways. First, it provides a review of the available query intent detection approaches. Second, it describes a feasible evaluation method to fairly compare such techniques. Third, the study has shown the relative performance of very different sources of evidence and, in addition, it has pointed out several promising lines of work. Fourth, the authors have discussed the unfeasibility of evaluating query intent detection methods by means of pooling.

This study also has limitations. First, only navigational intent has been studied. Second, not all available query intent detection methods were replicated. Third, the combination of different sources of evidence was quite naïve. Fourth, the collection of web pages to extract anchor texts and the manually labeled sample could have been larger.

Hence, further work is needed in the following lines: (1) replicating those query intent detection methods not studied in these experiments; (2) including informational and transactional intent in addition to navigational intent; (3) deeper analysis of the possible ways of combining different source of evidence; and (4) development of larger manually labeled datasets and anchor text collections.

Additionally, such future work should also pay attention to new "dimensions", orthogonal to query topic and intent, such as geographical location (e.g. [12, 32]), commercial [10], product or job-seeking intent [18]. Additionally, the problem of separating queries issued by human beings from those submitted by software agents is still little studied [9, 34].

8. ACKNOWLEDGEMENTS

The authors would like to thank Microsoft Research and, particularly Evelyne Viegas, for providing access to the MSN Query Log. They are also grateful to the anonymous reviewers for their valuable comments and advice. They would also like to thank the following people for labeling the evaluation data: César, Tania, Andrés, Diego, María, Miguel, Alejandro, Guzmán, Ignacio, and Rodrigo. This work was partially financed by grant UNOV-08-MB-13 from the University of Oviedo.

9. REFERENCES

[1] Baeza-Yates, R., Calderon-Benavides, L., and Gonzalez-Caro, C. The Intention Behind Web Queries. *Lecture Notes in Computer Science 4209*, (2006), 98.

[2] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A.P., and Yilmaz, E. Relevance assessment: are judges exchangeable and does it matter. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, (2008), 667-674.

[3] Beitzel, S.M., Jensen, E.C., Chowdhury, A., and Frieder, O. Varying approaches to topical web query classification. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, (2007), 783-784.

[4] Beitzel, S.M., Jensen, E.C., Frieder, O., Lewis, D.D., Chowdhury, A., and Kolcz, A. Improving automatic query classification via semi-supervised learning. *Proceedings of the Fifth IEEE International Conference on Data Mining*, (2005), 42-49.

[5] Brenes, D.J., and Gayo-Avello, D. Automatic detection of navigational queries according to Behavioural Characteristics. *LWA 2008 Workshop Proceedings*, (2008), 41-48.

[6] Broder, A. A taxonomy of web search. *ACM SIGIR Forum 36*, 2 (2002), 3-10.

[10] http://www.mturk.com/

[7] Broder, A.Z., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., and Zhang, T. Robust classification of rare queries using web knowledge. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, (2007), 231-238.

[8] Buckley, C., Dimmick, D., Soboroff, I., and Voorhees, E. Bias and the limits of pooling for large collections. *Information Retrieval 10*, 6 (2007), 491-508.

[9] Buzikashvili, N. Sliding window technique for the web log analysis. *Proceedings of the 16th international conference on World Wide Web,* (2007), 1213-1214.

[10] Dai, H.K., Zhao, L., Nie, Z., Wen, J.R., Wang, L., and Li, Y. Detecting online commercial intention (OCI). *Proceedings of the 15th international conference on World Wide Web*, (2006), 829-837.

[11] Flanagan, D. *MQL Reference Guide*, (2008). Available at: `http://mql.freebaseapps.com/` (Accessed 24 November 2008)

[12] Gravano, L., Hatzivassiloglou, V., and Lichtenstein, R. Categorizing web queries according to geographical locality. *Proceedings of the twelfth international conference on Information and knowledge management*, (2003), 325-333.

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

[14] Jones, K.S. and van Rijsbergen, C. Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development, 1975. Cited by Voorhees, E.M. and Harman, D. The text retrieval conferences (TRECS). *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15*, 1998, Association for Computational Linguistics Morristown, NJ, USA (1998), 241-273.

[15] Kang, I.H. and Kim, G.C. Query type classification for web document retrieval. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, (2003), 64-71.

[16] Kittur, A., Chi, E., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proceedings of the 26th annual SIGCHI conference on Human Factors in Computing Systems,* (2008), 453-456.

[17] Lee, U., Liu, Z., and Cho, J. Automatic identification of user goals in Web search. *Proceedings of the 14th international conference on World Wide Web*, (2005), 391-400.

[18] Li, X., Wang, Y.Y., and Acero, A. Learning query intent from regularized click graphs. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, (2008), 339-346.

[19] Li, Y., Krishnamurthy, R., Vaithyanathan, S., and Jagadish, H.V. Getting work done on the web: supporting transactional queries. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (2006), 557-564.

[20] Li, Y., Zheng, Z., and Dai, H.K. KDD CUP-2005 report: facing a great challenge. *ACM SIGKDD Explorations Newsletter 7*, 2 (2005), 91-99.

[21] Liu, Y., Zhang, M., Ru, L., and Ma, S. Automatic Query Type Identification Based on Click Through Information. *Lecture Notes in Computer Science 4182*, (2006), 593-600.

[22] Microsoft. *Microsoft Research Microsoft Live Labs: Accelerating Search in Academic Research 2006, Request for Proposals*, (2006). Available at: `http://research.microsoft.com/ur/us/fundingopps/RFPs/Search_2006_RFP.aspx` (accessed 24 November 2008).

[23] Nettleton, D., Calderon, L., and Baeza-Yates, R. Analysis of Web Search Engine Query Sessions. *Proc. of WebKDD*, (2006), 20-23.

[24] Pu, H.T., Chuang, S.L., and Yang, C. Subject categorization of query terms for exploring Web users' search interests. *Journal of the American Society for Information Science and Technology 53*, 8 (2002), 617-630.

[25] Rose, D.E. and Levinson, D. Understanding user goals in web search. *Proceedings of the 13th international conference on World Wide Web*, (2004), 13-19.

[26] Shen, D., Pan, R., Sun, J.T., *et al.* Query enrichment for web-query classification. *ACM Transactions on Information Systems (TOIS) 24*, 3 (2006), 320-352.

[27] Shen, D., Sun, J.T., Yang, Q., and Chen, Z. Building bridges for web query classification. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (2006), 131-138.

[28] Spärck-Jones, K. Automatic indexing, *Journal of Documentation 30*, (1974), 393-432.

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

[30] Tamine, L., Daoud, M., Dinh, B.D., and Boughanem, M. Contextual query classification in web search. *LWA 2008 Workshop Proceedings*, (2008), 65-68.

[31] Taylor, A.R., Cool, C., Belkin, N.J., and Amadio, W.J. Relationships between categories of relevance criteria and stage in task completion. *Information Processing & Management 43*, 4 (2007), 1071-1084.

[32] Wang, L., Wang, C., Xie, X., et al. Detecting dominant locations from search queries. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, (2005), 424-431.

[33] Yilmaz, E. and Aslam, J.A. Estimating average precision with incomplete and imperfect judgments. *Proceedings of the 15th ACM international conference on Information and knowledge management*, ACM Press New York, NY, USA (2006), 102-111.

[34] Zhang, Y. and Moffat, A. Separating Human and Non-Human Web Queries. *Web Information Seeking and Interaction*, (2007), 13-16.