Automatic detection of navigational queries according to Behavioural Characteristics

David J. Brenes, Daniel Gayo-Avello
University of Oviedo
research@davidjbrenes.info, dani@uniovi.es

Abstract

One of the main interests in the Web Information Retrieval research area is the identification of the user interests and needs so the search engines and tools can help the users to improve their efficiency. Some research has been done on automatically identifying the goals of the queries the user submits, but it is usually based in semantic or syntactic characteristics. This paper proposes the analysis of queries from a behavioural point of view.

1 Introduction

Since the beginning of the Web, one of the first interesting questions for researchers was the behaviour of the user searching in a new environment, different in size and structure from other traditional IR environments, such as catalogues or libraries.

The earliest analysis of search engines' query logs (e.g [Jansen *et al.*, 1998; Silverstein *et al.*, 1998]) pointed out that the common assumptions about users from traditional IR studies are not applicable to the average Web search user. Thus, the main objective was the proper characterization of the needs and goals of Web search users in order to develop better tools which allow the users to improve their efficiency.

One of the first studies analyzing the goals of the users beneath their submitted queries was [Lau and Horvitz, 1999], which classified the queries by the actions the user performed on them (reformulating a query, adding or removing terms in order to specialize or generalize it, etc.) and, in addition to this, attending to the subject (*Current Events, People, Health Information...*).

Such a classification is used to develop a probabilistic user model which could predict the next action for a user besides the topic of the query. This is, to the best of our knowledge, the first attempt to automatically classify queries according to their underlying goals.

Another attempt to classify queries was described by [Broder, 2002], where the proposed taxonomy was based on the behaviour of the Web search users and their searching process.

This classification divided queries into three types: Navigational (the user issues a query in order to reach a unique website, e.g. `yellow pages` or `amazon`), Informational (the user needs factual information which presumes is available in the Web although the website (or websites) is not only unknown but irrelevant, e.g. `auto price` or `history telegram`) and Transactional (the user wants to perform some kind of Web-mediated transaction but the website where fulfill is not important, e.g. `auctions`, `jokes` or `weather forecast`).

The main merit of such a classification is that it pays attention not to the queries but to the users' behaviour and their underlying needs. This level of abstraction circumvents some weaknesses presents in other methods (e.g. [Lau and Horvitz, 1999]) such as language or interpretational dependencies.

It must be noted that Broder's taxonomy has served as background to others such as those by [Rose and Levinson, 2004] or [Jansen *et al.*, 2008] which added more categories and sub-categories to the taxonomy.

Broder applied his classification manually over a small query log (1000 queries). He did not believe that a fully automatic classification could be feasible. Nevertheless, some authors have faced that problem. [Lee *et al.*, 2005] tried to infer the type of query (navigational or informational) with user questionnaires, however, they also tried a first approach to an automatic classification by attending to the way the links from different websites and the user clicks were distributed through the results. [Jansen *et al.*, 2008] took a different approach trying to classify the queries attending to some lexical or semantic features such as the number of words in the query or the meaning of particular terms.

The research approach taken in this paper traces back to the ideas of [Broder, 2002] and [Lee *et al.*, 2005]]; that is, avoiding the analysis of the queries from their superficial appearance (i.e. the terms) and trying to identify behavioural features which can give us hints to infer the search goal for a user within a search session.

Inferring this search goal could mean an interesting step forward in assisting the user in their search session as search engines could adapt their interfaces to this task (i.e. search engine could detect our navigational intent and help us showing a little paragraph about the characteristics of the results in order to allow us to take a decision about which website we are searching for).

In this paper we will only focus on navigational queries and we will propose three types of features (with their corresponding pros and cons) which combined could enrich the knowledge about the user actions. Finally, the results of their application on a query log will be analyzed in order to evaluate their appropriateness.

Structure of the paper

This paper is structured as follows. First, the methodology, tools and data used to perform the research are described. Secondly, the research goals are introduced and the aforementioned features of navigational queries are explained.

Then, some results of the application of those features are described and, finally, some conclusions are drawn and further work discussed.

2 Methodology

The experiments described in this paper were performed on the query log published by AOL in 2006 and described in [Pass *et al.*, 2006].

Although this query log has been surrounded by controversy because of its privacy concerns (as seen in [Hafner, 2006] or [Barbaro and Jr, 2006]), it is the largest and most recent publicly available query log contrasting with those used in [Broder, 2002], [Jansen *et al.*, 1998] or [Silverstein *et al.*, 1998] which are either small or not publicly available.

Besides, according to [Anderson, 2006], research involving such query log could not be considered unethical as long as its aim is not the identification of actual people which is not the case.

The data was loaded into a PostgreSQL database and queried with simple SQL and Python scripts for the execution of more complex algorithms (see Section 3.3).

This algorithms, database queries and other related documents are planned to be released in the author's homepage, in order to be discussed and, maybe, improved by research community.

3 Behavioural Characteristics of Navigational Queries

In this section we will describe the three aforementioned behavioural features. For each of them we will describe the associated Navigational Coefficient (NC) in addition to its benefits and weaknesses.

3.1 Weight of the most visited result

A feature one could expect in navigational queries is as significant high rate of clicks on the same results which would ideally be the website the users had in mind when issuing the query. That is, if users submitting a particular query tend to click on the same result we could safely infer a strong relation between the query and that result (i.e. website).

Table 1 shows an example[1] where users have clicked over different results and none of them shows a clear dominance.

Query	Result	Visits
`baby names`	http://www.babynames.com	1
`baby names`	http://www.babynamesworld.com	1
`baby names`	http://www.thinkbabynames.com	1

Table 1: Clicked results and number of visits for a non-navigational query

[1]These examples has not been executed over the whole query log, but over a small portion of 500 queries.

Table 2 shows just the opposite. A query in which a single result dominates over the rest and, thus, appear to be clearly navigational.

Query	Result	Visits
`jesse mccartney`	http://hollywoodrecords.go.com	13
`jesse mccartney`	http://groups.msn.com	2
`jesse mccartney`	http://jessemccartneyonline.v3.be	2
`jesse mccartney`	http://www.hyfntrak.com	2

Table 2: Clicked results and number of visits for a navigational query

This behaviour was described by [Lee *et al.*, 2005] who used a statistical function in order to define it on his query distribution.

In our research we will assume the NC is equal to the *percentage of visits which go to the most visited result* for the query (see Figure 1).

$$NC = \frac{Number\_of\_visits\_most\_popular\_result}{Number\_of\_visits\_to\_results}$$

Figure 1: Navigational Coefficient Formula for the weight of the most popular result

Following the examples provided in Tables 1 and 2, we could estimate a NC of *0.33* for `baby names` and *0.68* for `jesse mccartney`.

An important issue is whether we must consider those searches where no result was clicked, as they can have a significant effect on the NC value.

As this feature has been defined on the assumption that those queries focusing on unique results have a mainly navigational goal we must consider those searches in which users do not click any result as 'failed' searches and take them into account. Of course, null cannot be the most popular result for a query and, hence, it will just increase the total number of results.

To illustrate this we will count the *null* results for the query `jesse mccartney` (see Table 3). The resulting NC is $\frac{13}{77(13+2+2+2+58)} = 0.1688$, very different from the one calculated with data in Table 2.

Pros and Cons

Although this feature is quite simple and intuitive it seems to only behave properly with those queries with a high number of submissions (i.e. the most frequent queries)

This navigational coefficient depends on the number of submissions for a query and the number of clicks on the

Query	Result	Visits
jesse mccartney	http://hollywoodrecords.go.com	13
jesse mccartney	http://groups.msn.com	2
jesse mccartney	http://jessemccartneyonline.v3.be	2
jesse mccartney	http://www.hyfntrak.com	2
jesse mccartney	*null*	58

Table 3: Clicked results and number of visits for a navigational query (including 'null' results clicked)

most popular results. The more submissions, the more difficult it is to find a small set of focused results except if a few of them are specially relevant for the query.

Problems also arise in less frequent queries as they can reach a high NC without a significant number of clicked results. For instance, a query issued just once and with one clicked result can reach a NC value of 1 which can or not be correct. Because of this, this NC is not meaningful for less frequent queries (which are in fact most of the actual queries a search engine receives).

The other problem is that this coefficient is not affected by the total number of different clicked results for a query. For example, query `New York` and query `Harry Potter` have an approximated NC value of 0.30[2] . These queries seems to have the same navigational behaviour, but `New York` have a little set of results (8 results) with predominance of three governmental websites and `Harry Potter` have a bigger set of results (43 results) with lots of fan pages clicked by the users.

Although both queries to have the same value for the coefficient, it seems that `Harry Potter`'s navigational behaviour is more dispersed in a bigger set of secondary results.

3.2 Number of distinct visited results

Another behaviour we can expect from navigational queries is to have a little set of results common to all the submissions. Some queries involve personal issues as language preference (preferring English versions of the information over Spanish ones, for example), interpretational problems (we could be searching information about a pre-Columbian `american idol` or information about the `american idol` TV show), etc.

To measure the NC for this behaviour we will calculate the rate of different clicked results over the number of clicks to results for the query. This value would give the highest value to a query which clicked results were never repeated (e.g. having 50 different results in 50 submissions would give a value of 1), so we must subtract it from 1 (see formula in Figure 2).

$$NC = 1 - \frac{Number\_of\_distinct\_results}{Number\_of\_visits\_to\_results}$$

Figure 2: Navigational Coefficient Formula for the number of visited results

Pros and Cons

This NC value complements the NC coefficient solely based on the importance of the most popular results (see section 3.1) and, thus, it could support it in those scenarios where the first one shows weakness.

As this NC is based on the size of the set of results, it solves the problem of ignorance of the number of different clicked results (the problem of `New York` and `Harry Potter`).

In fact this NC measure has an analogous problem ignoring the distribution of clicks to the different results. For example, query `Eric Clapton` has near 10 different clicked results, the same as `Pink Floyd`, with the same number of visits. This NC would give the same value to each query, but `Pink Floyd` clicks are more focused in one result while `Eric Clapton` has two prominent clicked results, which points out a more navigational behaviour for `Pink Floyd`.

Another problem is that this coefficient does not provide really meaningful results for less frequent queries because it favours the most frequent ones. Hence, given two queries with the same number of different results, the more frequent one will reach a higher navigational coefficient.

3.3 Percentage of navigational sessions

The last feature studied in this paper is related to the relation between navigational queries and the rest of queries issued within the same search session. Navigational queries are those submitted to reach a particular well-known website where the users will continue their interaction. Thus, these kind of queries are likely to appear isolated within sessions comprised of just one unique query.

Of course, this is oversimplification given that we can imagine several scenarios in which navigational queries are intermingled with transactional or informational queries belonging to the same search session.

Nevertheless, the underlying idea beneath this third NC measure is that navigational queries are more frequent in such mini-sessions, those consisting of a query and a clicked result. Thus, this NC for a query will be computed as the rate of navigational sessions over all the sessions containing that particular query (see Figure 3)

$$NC = \frac{Number\_of\_navigational\_sessions}{Number\_of\_sessions\_of\_query}$$

Figure 3: Navigational Coefficient Formula for the number of navigational sessions

[2]We are discarding 'null' results for this example.

Pros and Cons

Instead, it mainly depends on the searching behaviour of the users. Issues such as multitasking searching can negatively affect to this coefficient. Some studies (e.g. [Spink *et al.*, 2006]) have studied the impact of this phenomenon and have concluded that multitasking is a quite common behaviour; however, other researchers (e.g. [Buzikashvili, 2006]) have put into question that fact and lowering the amount of multitasking to a mere 1% of search sessions. This sharp contrast can likely be explained by a different conception of multitasking, Buzikashvili, for instance, does not consider sessions with sequential searching as multitasking while it seems that [Spink *et al.*, 2006] counted as multitasking sessions all those comprising of two or more different topics.

Another important issue with this approach is the session detection method applied on the query log which can be a line of research on its own. In this paper, the authors have applied the technique by [He and Göker, 2000] ; those researchers applied a temporal cutoff to separate queries into different sessions. With large thresholds the chance of finding 'navigational sessions' decreases but if the threshold is arbitrary low the number of false positives (non navigational queries flagged as navigational) will rise.

4 Results

In this section we will provide some results for each NC in addition to relevant examples.

Although in previous section some examples were explained with few queries, and following examples contains only 50 queries, the extracted statistical data are the results of performing the explained experiments over all the query log and not only over a small set of queries.

4.1 Weight of the most visited result

Table 4 shows the results obtained when applying the NC described in section 3.1. Those queries with a frequency lower than 50 were removed because their results were judged to be no reliable enough.

Most of these queries had a relatively low frequency, specially if we think of 'typical' navigational queries such as `ebay`, `amazon` or `cnn`. This illustrates the fact that this coefficient favours less frequent queries.

All these queries have been issued to the AOL search engine (which actually is a customized version of Google) and many of them seem to be navigational (e.g. the query being part of the URL for some of the results) but in other cases the navigational nature is not so clear (e.g. cosmology book or links for astrology).

Some queries (like `cosmology book` or `links for astrology`) present less evident results and its navigational behaviour seems to be caused by the actions a low number of users submitting the query (1 user in the case of `cosmology book` and 3 users for `links for astrology`). Thus, navigational behaviour of those queries is restricted to a little set of users and its extrapolation to all search users.

4.2 Number of distinct visited results

Table 5 shows the results obtained when applying the NC described in section 3.2. Those queries with a number of clicks to results lower than 50 were removed because their results were judged to be no reliable enough.

Queries in Table 5 could be considered typical examples of navigational queries; in fact they exhibit many of the characteristics described by [Jansen *et al.*, 2008] as important to detect navigational queries: many of them contain URLs or URL fragments (6 out of 20), company names (7 out of 20), or website names (7 out of 20). Thus, it seems that this coefficient behaves in a similar way to the heuristic method proposed by [Jansen *et al.*, 2008]; however it is fully automatic and does not rely on lists of companies or websites.

Query	NC	Submissions
`drudge retort`	1,00	206
`soulfuldetroit`	1,00	138
`cosmology book`	1,00	127
`ttologin.com`	1,00	124
`jjj's thumbnail gallery post`	1,00	108
`beteagle`	1,00	104
`yscu`	1,00	100
`frumsupport`	1,00	89
`cricketnext.com`	1,00	86
`msitf`	1,00	85
`aol people magazine`	1,00	84
`louisiana state university at alexandria`	1,00	84
`links for astrology`	1,00	78
`mod the sims 2`	1,00	73
`richards realm thumbnail pages`	1,00	70
`lottery sc`	1,00	69
`trip advisor half moon jamaica`	1,00	66
`orioles hangout`	1,00	65
`www. secondchance birdrescue .com`	1,00	64
`prosperitybanktx .com`	1,00	63

Table 4: 20 top navigational queries according the 'weight of most popular document'

this NC obtains the highest values with the most frequent queries (`google` is, ironically, one of the most frequent queries in the AOL log) and with those with most clicked results (`fighton` doubles the number of clicked results when compared to the most navigational query according to the first coefficient: `drudge retort`).

4.3 Percentage of navigational sessions

Table 6 shows the results obtained when applying the NC described in section 3.3. Those queries which had participated in less than 50 sessions were removed because their results were judged to be no reliable enough.

This results, as those shown in Table 4, are queries which are far different from the common navigational query. However, most of them fulfill the criteria proposed by [Jansen *et al.*, 2008] because they consist of companies or websites names. However, unlike results in Table 5, like Google or Bank of America, these companies are rather small but obviously known to the user submitting the query (e.g. Mission viejo is a swimming club in California, and El Canario by the Lagoon is a hotel in San Juan, Puerto

Query	NC	Visits to results
google	0,9995	264.491
yahoo.com	0,9990	61.501
mapquest	0,9990	57.459
yahoo	0,9989	107.470
ebay	0,9988	71.521
google.com	0,9987	53.235
bank of america	0,9987	19.287
www.google.com	0,9986	26.867
www.yahoo.com	0,9983	25.536
yahoo mail	0,9983	15.184
myspace.com	0,9983	54.446
fidelity.com	0,9981	9.226
wachovia	0,9981	5.413
hotmail	0,9981	17.233
target	0,9979	6.429
msn	0,9979	16.361
southwest airlines	0,9979	11.013
ups	0,9979	3.818
bildzeitung	0,9978	457
fighton	0,9977	444

Table 5: 20 top navigational queries according the 'Number of distinct visited results'

Rico).

Queries in Table 6 have been issued in a low number of total (not only navigational sessions). This is not surprising, as the higher the number of total sessions, the higher the needed number necessary navigational sessions to get a high NC value.

However, the low values of NC (the lowest from the three behavioural characteristics) indicates that very few navigational sessions have been detected for more 'typical' navigational queries (i.e. google, ebay) because the needed percentage ratio of navigational sessions didn't exceeded 75%.

This could point out either a real lack of navigational session behaviour or a failure in the detecting sessions algorithm, which hides navigational sessions inside another informational or transactional sessions.

5 Discussion and Future Research

5.1 Comparison of results

When studying some of the NCs, negative aspects for each of them were analyzed. Some of these weaknesses were solved by some other NC, which extracts different characteristics from the same queries. Thus, the idea of combining these coefficients appears as a possible way to enhance their descriptive power and effectiveness.

Discovering a metric for combining the values from different NCs is an objective which exceeds the aspirations of this paper but, as a first approach, we can study how all of the described coefficients behave in the queries which have been detected along the research as being clearly navigational, i.e. the 60 discovered queries listed in Tables 4, 5 and 6.

Coefficients have been calculated for each detected query and will be presented in three separated graphics, according to the coefficient in which the query has been detected. Showing the results in tables was considered as an option, but it was judged that using graphics would make more clear their description.

Query	NC	Sessions
natural gas futures	0.867	53
cashbreak.com	0.830	106
allstar puzzles	0.813	59
times enterprise	0.802	71
instapundit	0.796	54
clarksville leaf chronicle	0.790	62
first charter online	0.789	57
mission viejo nadadores	0.787	66
county of san joaquin booking log	0.781	64
thomas myspace editor beta	0.773	53
kgo radio	0.772	79
el canario by the lagoon	0.770	61
yahoo fantasy basketball	0.764	51
kenya newspapers	0.752	85
parkersburg news	0.750	76
slave4master	0.750	56
online athens	0.742	70
ace games.com	0.740	54
family savings credit union	0.739	69
debkafile	0.736	72

Table 6: 20 top navigational queries according the 'percentage of navigational sessions'

Figure 4 shows the evolution of the coefficients along the queries listed in Table 4. The queries has been ordered by the value of the NC based on the importance of their most important result.

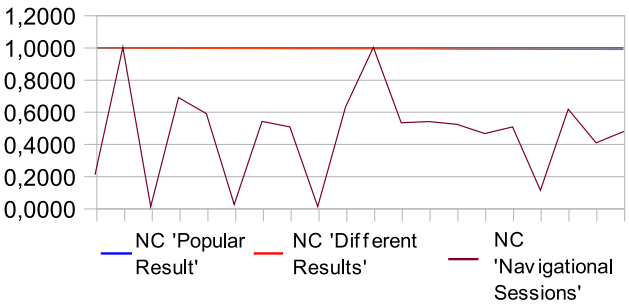

Figure 4: Navigational Coefficients for Top 20 'Most Popular Result' NC

In Figure 4 an important equivalence is shown between the coefficient based on the size of the set of clicked result and the one based in the weight of the most clicked result. In fact, the graphic corresponding to this coefficient hides the other one (because of that, there's no 'blue colour' in graphic).

On the other hand, coefficient based on navigational sessions doesn't have a constant behaviour and fluctuates from the lowest value (0) to the highest (1).

Figure 5 shows the evolution of the coefficients along the

queries listed in Table 5. The queries has been also ordered by the value of the NC based on the importance of their most important result.

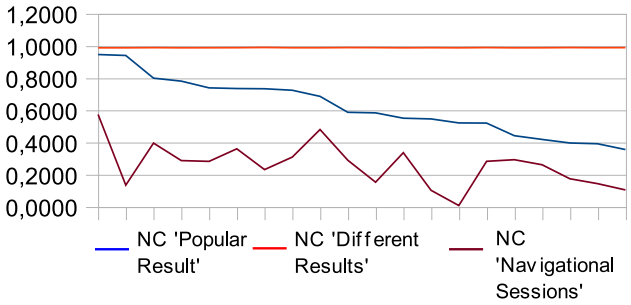

Figure 5: Navigational Coefficients for Top 20 'Number of Distinct Results' NC

The equivalence shown in Figure 4 is not present in Figure 5. Although initial values are very close (but never equal) values from coefficient based on the weight of the most popular document decrease to low values.

In some cases (like `www.yahoo.com`) this difference can be explained by the numerous presence of 'null' results, which reduce the importance of the most visited result (for `www.yahoo.com`, only the 48% of the users clicked in the most important result, and 41% didn't click in any result).

The navigational coefficient shows an erratic behaviour in this Figure too, although its range of values is lower than in the Figure 4, and final queries present more continuous values.

Figure 6 shows the evolution of the coefficients along the queries listed in Table 6. The queries has been also ordered by the value of the NC based on the importance of their most important result.

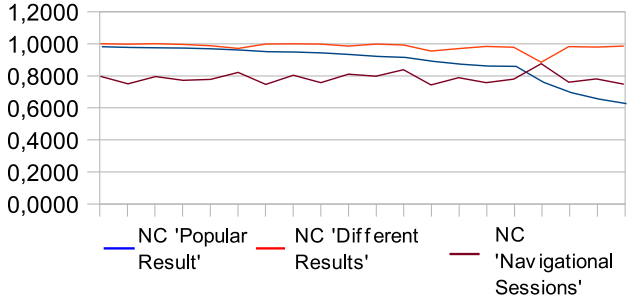

Figure 6: Navigational Coeficients for Top 20 'Navigational Sessions' NC

The coefficients' behaviour shown in Figure 6 is the most stable. In this Figure a relationship between the three coefficients is more insinuated than in the other two.

From these graphics we could extract the following conclusions:

1. Having a result concentrating the traffic usually means having a little set visited of results.
2. Having a little set of visited results doesn't have to mean that a result is taking most of the visits.
3. Having a very popular result, or having a little set of visited results, doesn't imply having a big number of navigational sessions.
4. Having a big rate of navigational sessions, usually means having a very popular result and a little set of visited results.

Of course, it would not be justifiable to extrapolate these conclusions to all the query log, but it shows the behaviour of a little set of automatically extracted queries.

Analyzing the whole query log according to these coefficients would give us a wider view about the behaviour of the coefficients and could be faced in further research.

Also, the confirmation of these behaviours in other query logs would allow us to make sure they are not due to any particularity form the users of AOL's query log.

5.2 Lack of navigational sessions

The lowest values of the three NCs are those based on the existence of navigational sessions which, on the three comparisons, present lower values than the other two NCs.

However, the third comparison shows a convergence between the values of the NCs which is not present in the other two. The question is Why the convergence between the navigational sessions based NC and the other two is not present in the other comparisons?

In section 3.3 the dependence on the algorithm for detecting sessions and the impact of multitasking sessions were pointed out.

Navigational queries are good candidates to be submitted inside other informational tasks in order to check some information in an specific webpage (such as wikipedia, a travel agency, etc.) or use a Web tool (such as a searcher, a dictionary, an agenda, etc.).

Navigational queries are also ideal to be grouped into a unique session because of a big cutoff interval selected for detecting sessions. For example, in the morning a user could look for his mail service to check its mail and, after reading it, search his preferred newspaper in order to read the latest news. These two navigational queries would be grouped into a unique sessions, which doesn't fit into our definition for navigational sessions.

The data shows us a lack of navigational sessions which could be explained by the process of detecting sessions, but Is this loss of navigational sessions real? If we use the other NCs to detect navigational sessions, how could we check that we are not creating artificial sessions?

Further research on this subject is needed in addition to a thorough evaluation through the division of the query log into training and sets.

5.3 Post-hoc characteristics of navigational queries

Some previous works ([Jansen *et al.*, 2008]) on automatic detection of navigational queries were based on lexical and semantic heuristics, such as a short length of query or the presence of domain suffixes.

Such heuristics are based on a post-hoc analysis of some common navigational queries (such as those shown in section 4.2) so they can led to biased methods favouring these specific kind of navigational queries in detriment of other queries where the navigational nature can only be inferred from the behaviour of the users.

In fact, applying such heuristics on our 'top 60' queries we found that: (1) 14 out of 60 queries (shown in Table 7) have three or more terms so they wouldn't qualified as navigational queries, and (2) the suffixes heuristic would only flagged 11 out of 60 queries as navigational, which

Query	Average NC	Terms
clarksville leaf chronicle	0,9013	3
bank of america	0,7702	3
first charter online	0,9215	3
mission viejo nadadores	0,9191	3
aol people magazine	0,9960	3
natural gas futures	0,8365	3
links for astrology	0,8402	3
yahoo fantasy basketball	0,9091	3
order of the stick	0,8186	4
mod the sims 2	0,8341	4
richards realm thumbnail pages	0,8147	4
jjj's thumbnail gallery post	0,8580	4
family savings credit union	0,8940	4
thomas myspace editor beta	0,8022	4
trip advisor half moon jamaica	0,6956	5
louisiana state university at alexandria	0,8381	5
el canario by the lagoon	0,9067	5
county of san joaquin booking log	0,8730	6

Table 7: Navigational queries with more than 3 terms

average NC is smaller than the average NC in the rest of them (0.71 against 0.81).

A curious fact is that those queries which have three variants like `google`, `google.com` and `www.google.com` or `yahoo`, `yahoo.com` and `www.yahoo.com` have less NC as the query is more 'complete' as it's shown in Table 8.

Query	Average NC
google	0,7031
google.com	0,6361
www.google.com	0,6104
yahoo	0,6771
yahoo.com	0,5878
www.yahoo.com	0,5773

Table 8: Decreasing navigational coefficient of '.com' queries

As an example, we could perform an easy experiment to discover the average value for the first NC for queries ending with `.com` (which appears to be the most common gTLD [Zooknic, 2008]).

Figure 7 shows the distribution of the number of '.com' queries according to the value of its NC (calculated following the formula in section 4.1).

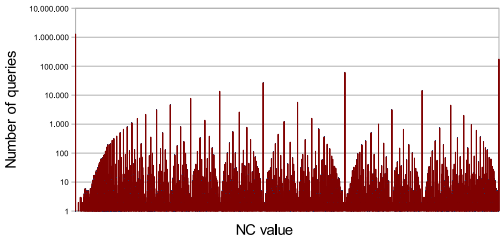

Figure 7: Distribution of '.com' queries according to its NC value

The average value for all '.com' queries is 0.15, which shows that this criteria (containing a URL) could be not as reliable as initially thought. The main reason for this low value is the million of queries which NC value is 0. Ignoring these queries the average NC rises approximately to 0.69, but the amount of ignored queries (almost 75% of all the '.com' queries) turns this decision into a controversial one.

5.4 Extraction of significant terms

Another criteria mentioned in [Jansen *et al.*, 2008] is the presence of companies, people or website names. As the authors of the paper recognize, this forces the dependence on a database of significant names.

We could rely on some distributed databases on the Web, such as FreeBase or even Wikipedia, in order to cover as much names as possible but, even with the greatest database, we could not control the appearance of new significant terms for searchers.

Two first NC criteria forces the discovering of queries which are very related to some Web resources (such as a travel agency, a web-mail interface or some specific software). These queries, for practical purposes, could be considered as 'names' of those Web resources and, therefore, as significant terms.

6 Conclusions

In this paper a method for the automatic detection of navigational queries has been proposed. Some characteristics of the users' navigational behaviour have been discussed and applied to the development of three different coefficients aiming to detect navigational behaviour in an automatic fashion.

We have provided some results obtained with the proposed methods comparing them against each other besides showing benefits and weaknesses of each approach.

The results are promising and, thus, a deeper analysis is required to obtained a combined metric from the three coefficients. A thorough evaluation is needed in addition to experiments on different query logs to check its feasibility on different types of users.

Acknowledgments

This work was partially financed by University of Oviedo through grant UNOV-08-MB-13.

References

[Anderson, 2006] Nate Anderson. The ethics of using aol search data. online, 08 2006.

[Barbaro and Jr, 2006] Michael Barbaro and Tom Zeller Jr. A face is exposed for aol searcher no. 4417749. *The New York Times*, August 2006.

[Broder, 2002] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36:3–10, 2002.

[Buzikashvili, 2006] N.N. Buzikashvili. An exploratory web log study of multitasking. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 623–624, 2006.

[Hafner, 2006] Katie Hafner. Researchers yearn to use aol logs, but they hesitate. *The New York Times*, August 2006.

[He and Göker, 2000] Daqing He and Ayse Göker. Detecting session boundaries from web user logs. pages 57–66, 2000.

[Jansen *et al.*, 1998] Bernard J. Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. Real life information retrieval: a study of user queries on the web. *SIGIR Forum*, 32:5–17, 1998.

[Jansen *et al.*, 2008] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44:1251–1266, 2008.

[Lau and Horvitz, 1999] Tessa Lau and Eric Horvitz. Patterns of search: analyzing and modeling web query refinement. pages 119–128, Banff, Canada, 1999. Springer-Verlag New York, Inc.

[Lee *et al.*, 2005] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. pages 391–400, Chiba, Japan, 2005. ACM.

[Pass *et al.*, 2006] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *The First International Conference on Scalable Information Systems*, page 1, Hong Kong, June 2006. ACM.

[Rose and Levinson, 2004] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. pages 13–19, New York, NY, USA, 2004. ACM.

[Silverstein *et al.*, 1998] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large altavista query log, 1998. http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html.

[Spink *et al.*, 2006] A. Spink, M. Park, B.J. Jansen, and J. Pedersen. Multitasking during web search sessions. *Information Processing and Management: an International Journal*, 42(1):264–275, 2006.

[Zooknic, 2008] Zooknic. Domain name counts, April 2008.