Stratified Analysis of AOL Query Log

David J. Brenes, Daniel Gayo-Avello [a]

[a] *Despacho 57, planta baja, Edificio de Ciencias. Calvo Sotelo s/n 33007. Oviedo, Asturias. Spain*

Abstract

Characterizing user's intent and behaviour while using a retrieval information tool (e.g a search engine) is a key question on web research, as it hold the keys to know how the users interact, what they are expecting and how we can provide them information in the most beneficial way. Previous research has focused on identifying the average characteristics of user interactions. This paper proposes a stratified method for analyzing query logs that groups queries and sessions according to their hit frequency and analyzes the characteristics of each group in order to find how representative the average values are. Findings show that behaviours typically associated with the average user do not fit in most of the aforementioned groups.

Key words: query log analysis, user behaviour, user interactions, user intent, user profiling

1 Introduction

An accepted idea in the HCI area is that 'there is no average user' Krug [15]. This is a key concept that expresses the complexity inherent in the diversity of users that could access a tool in the Web environment. This complexity becomes crucial in search engines as they are one of the main access point to information retrieval in the Web [1] and they are used by many different users.

Email addresses: research@davidjbrenes.info (David J. Brenes), dani@uniovi.es (Daniel Gayo-Avello).

[1] Meiss et al. [19] have reported that access from search engines are not as common as assumed (5% of all the web access), but in the same paper is pointed out that it is not the actual value but a lower bound, as only requests coming directly from search engines are taken into account.

Usability issues in search engines do not only concern designing interfaces but also what are the user needs, how he express them and how he reacts to the retrieved information. The diversity of user's profile makes very important to take care about how the user behaves in order to implement better information retrieval web tools. Questions about behaviour (How many searches does a user execute? How many words does he use to express his goal? How many results does he visit?) go beyond the interfaces and deserve answers from research community which has already addressed them.

Beyond analyzing the behavior of a hypothetical average user whose correspondence with actual users in all situation, there are some research initiatives that tries to group queries and user interactions according to some criteria so different access patterns can be identified and intentions can be inferred,

This paper proposes a new criteria for grouping queries and user's interactions with search engines analyzing different measures traditionally extracted in previous research papers in order to check if this criteria allow us to separate different use cases of search engines with different characteristics.

2 Literature review

Since the late 90's researches have studied how the search engines influence users' navigation process. Jansen et al. [12], Silverstein et al. [25] and Lau and Horvitz [16] published first results from analysis on query logs from search engines widely used (*Excite* and *Altavista*) in order to describe the interactions between the user and the tool. Jansen et al. [12], Silverstein et al. [25] had a descriptive aim, extracting analytical results for some metrics in the query log (e.g. length of a query, number of visited results, etc.). According to these metrics, they revealed significant differences between the common assumptions about users in traditional IR systems and the actual behaviour of users in web IR systems.

These studies focus the efforts on obtaining metrics which give us a glimpse on the behaviour of all the users showing an ideal average user's behaviour. This is a powerful approach which have provided valuable results but offers a description which doesn't take into account the existing differences between users or between different search episodes from a single user. This way, these studies doesn't help us to solve questions about how users interact with search engines in different situations with different motivations.

One alternative approach is to group queries and users according to some criteria which allow us to find differences in users' behaviour and analyze them. Following this direction, Spink et al. [27] analyzed behaviour depending

on user's location, comparing the values of statistical measures (e.g. average length of queries, session length, etc.) for European and US users.

Another approach is to group different user's interactions in search sessions. This research, accomplished by many authors (e.g. He and Göker [9]), allow us to group queries in which user's intentions are supposed to be the same or, at least, very highly related.

The underlying goal of the interactions between the user and the search engine is also a good criteria to study the user's behaviour depending on what he is expecting from the search engine.

Lau and Horvitz [16] classified a set of queries and related the information goal corresponding to those queries (e.g. *Current Events*, *Health Information* or *Products and Services*), the actions performed by the user (e.g. generalization or specialization of a query, request for additional results, etc.) and the time elapsed between those actions to infer future users' behaviour.

Following previous research in traditional Information Retrieval, Broder [6] studied users' informational goals from a different perspective, focusing on the abstract intentions (e.g. '*I want to reach a website focused on this subject*', '*I'm searching for varied information about a subject*', etc.) and not only in the subjects of the queries. From this analysis he extracted a taxonomy that 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`).

Broder's taxonomy served as basis for others authors (e.g. Rose and Levinson [23], Jansen et al. [14]) who reviewed and extended it, adding levels of abstraction and dividing the original Broder's categories in more subcategories. In theory, these taxonomies based on behavioural characteristics don't have any language or interpretational dependencies (once the different behaviours are defined), so they seem good candidates to be implemented and run over big sets of queries and some authors have faced this problem.

Although Broder did not believe that a fully automatic classification could be feasible and thus he classified manually a small set of queries some authors (e.g. Lee et al. [17], Jansen et al. [14]or [5]) have defined characteristics which could be recognized by an algorithm in order to automatically detect and classify certain type of queries.

3 Research motivation

3.1 Research questions

The underlying aim of this paper is to show a new criteria to group queries submitted by users and the search sessions they perform so we can have a glimpse about the motivation shown by a single user in a specific search episode and the characteristics of the user's interactions with the search engine.

The criteria will allow us to classify queries in different groups so we can extract usual statistical measures for each group and check if these characteristics depends on the group where queries were classified. The measured characteristics will be:

Navigational Coefficient It's a measure which will report us how much a query fits in the definition of a navigational query as proposed by Broder [6].

Query Length A valuable measure that report us about the number of words of a query.

Number of visited results An interesting characteristics which gives us an idea about how many resources needed the user to fulfill the underlying informational goal.

Failed Submissions The percentage of submissions of a query in which the user didn't visit any result at all.

The criteria will also be applied to search episodes detected by the authors in the query log, which will be grouped in the same way as the queries will be. For search session we will measure:

Session Length The number of queries the user submitted before solving the informational need or leaving it unsolved.

Number of visited results As seen in the case of queries, this characteristic gives us an idea of the complexity of the user's informational need.

Failed Submissions The percentage of search sessions where the user didn't visit any result.

At the end of these experiments we will have learned something about the characteristics of the user's behaviour on each group of queries and sessions. It could be interesting to link those results and check if groups of queries and sessions are related in some way (i.e. if queries from a group are very related with sessions of another group).

3.2 Proposal of a new grouping criteria

As we have seen there are many examples in the literature of criteria used to group queries and extract common characteristics (e.g. the location of the user or belonging to the same search episode).

This paper aims to examine the results obtained for a new grouping criteria, based not on the location of the user or the relation between queries but on the popularity of the query (i.e. the number of times a query has been submitted to a search engine).

In the following subsections, the grouping algorithm will be explained with some examples in order to make it understandable.

3.2.1 Queries

A query log usually consists on thousands (or millions) of record, each of them stores information about the submission of a query and one result visited by the user (more than one record if the user visited more than one result, and *null* values if the user didn't visit any result).

The chosen criteria to classify a query is the number of records corresponding to that query, so a query q will be classified in a group g, which first query is qi, only if the number of records corresponding to q in the AOL query log differs in 15% or less from the number of records corresponding to qi.

An example is shown in Table 1. In this example, query `google` has 332.002 related records in the query log and is the first query of the first group. For the second query (`ebay`) to belong the same group it must appears in 282.202 records (332.002 - 332.002*15%). However, `ebay` only has 139.171 records, so it's classified in another group. The third query, `yahoo`, appears in 130.535 records which differs less than 15% from the number of records for `ebay`, so it's classified in the same group.

We repeat this process for all the queries in the AOL query log and classify all the queries in groups according the number of records they have in the query log.

3.2.2 Sessions

Sessions in the AOL query log have been detected implementing the algorithm described by He and Göker [9] taking a fixed threshold of 20 minutes as the maximum difference between two queries being in the same session. Other

Group ID	Query	# of records
1	google	332.002
2	ebay	139.171
2	yahoo	130.535
3	yahoo.com	97.518
3	mapquest	88.268
4	google.com	79.990
4	myspace.com	77.202
4	myspace	74.362

Table 1
Examples of groups of queries

approaches could have been evaluated, but this is out of the scope of this paper.

Sessions will be grouped according to its first query. The approach is group those queries which begin a new search sessions, as detected by the algorithm previously described, and classify the sessions according to the group corresponding to its first query (e.g. if query `google` is classified in group 2, every search session started by query `google` is classified in group 2).

Whit this purpose, the grouping algorithm explained in previous section (section 3.2.1) is repeated taking into account only those queries which start some search sessions.

4 Research design

4.1 Data description

As this paper aims for the stratified study of query log data no various sources of data were necessary, so one query log would be enough to perform the required analysis. The only requirement for the chosen query log was having a representative number of queries to obtain results with statistic value and being publicly available for research. The log which best fit this requirements was the AOL query log, described by Pass et al. [21].

This log was surrounded by controversial because of its privacy concerns that allowed the press to discover the identity of some of the users recorded on the log (Barbaro and Jr [4]). Some debates were produced about the ethics of

using this kind of data (e.g. Hafner [8] or Anderson [3]) and research has been done on privacy issues (e.g. Adar [1] or Xiong and Agichtein [28]). According to Anderson [3], using AOL query log for research can not be considered unethical as long as the aim is not the identification of actual people. Analysis described in this paper doesn't aim to identify real users and try to process the queries without paying attention to its content.

4.1.1 Query log's characteristics

AOL query log consists of approximately 20 millions of queries submitted by 650.000 users from March to May in 2006.

A record on this query log represents the visit to a result for a query or the submission of a query (if no result is visited). Each record stores:

- An anonymous ID that allows to group queries from the same user without revealing the AOL user's nickname.
- Query submitted by the user.
- Date and hour of the submission of the query.
- Rank position of the result visited by the user on each record.
- Domain portion of the URL of the result visited. For example, for URL *http://www.nasa.gov/multimedia/imagegallery/image_feature_1137.html*, the domain portion is *www.nasa.gov*.

Examples of these records can be found in Table 2.

User ID	**Query**	**Date and Time**	**Visited result rank**	**Visited result URL**
142	`rapny.com`	2006-05-18 09:21:57		
217	`lottery`	2006-03-01 11:58:51	1	http://www.calottery.com

Table 2
Examples of AOL log queries

The AOL query log has been described by Pass et al. [21] and a full description can be found there.

The most submitted query in the AOL log is the *'-' query*, a query which only character is `-`. This query shows a wide range of visited results and it's submitted almost 1.000.000 times.

Visited results for this query (e.g. http://www.theonering.net, http://www.sacramentochoral.com or http://www.market4demand.com) make us to believe that some queries have been masked behind `-` character.

Instead of removing it from the analysis we decided to analyze it like the rest of the queries as it's isolated in one group (the first of them, as we will see in section 5.1.1) and it's easy to be detected.

4.2 Experiments

4.2.1 Queries

4.2.1.1 Navigational Coefficient We reviewed different attempts of queries classification. One of these approaches was presented by Broder [6] who divided queries in three categories: navigational, informational and transactional. We also mentioned research focused on the automation of this classification over large sets of queries. Thus, it's reasonable to think that classification of queries has significant research interest and, because of that, a valuable characteristic to observe.

In this study we will not attempt to classify queries in each group (as it's beyond the intention of the paper), but to apply some of the ideas pointed by Brenes and Gayo-Avello [5] and [17] analyzing the relationship between the group a query belongs to and its navigational coefficient (a value that measures the navigational intent behind a query).

Thus the so called *navigational coefficient* of a query will be measured as the percentage of visits which go to its most visited result (see Figure 1). This coefficient was chosen as it was studied in both of the aforementioned papers and it reflects the association between a query and a website in the mind of the users.

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

Fig. 1. Navigational Coefficient Formula

Thus, data present Table 3 would allow us to estimate the *navigational coefficient* for those queries, being the results 0.33 and 0.68 respectively for queries `baby names` and `jesse mccartney`.

For less frequent queries, this value will not be really meaningful, as they need less visits to a result for obtaining high navigational coefficients. However, grouping the queries will correct this behaviour, although the coefficient won't be as significant as for other queries.

4.2.1.2 Query Length The number of terms for queries is an indicator of the complexity of a query which is usually measured in query log analysis.

Query	Result	Visits
`baby names`	http://www.babynames.com	1
`baby names`	http://www.babynamesworld.com	1
`baby names`	http://www.thinkbabynames.com	1
`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 3
Clicked results and number of visits two queries

In Jansen et al. [13], authors present a Table (Table 6) in which show the distribution of queries according to their number of terms. In Jansen and Pooch [11] this measure is even used to compare different web-user studies.

4.2.1.3 Visited Results Interpretation of this characteristic is very arguable. Does it indicate non-appropriate results for the query (forcing the user to visit more results)? Or Does it indicate a broad informational goal that needs information from various sources in order to be fulfilled? Whatever the answer is, a high number of visited results seems to indicate a more difficult to solve informational need.

The process for calculating this value will be divided in two steps. First, as each record in the query log contains information about a visited result (*null* if no result was visited by the user) we will group every adjacent record (being from the same user), assuming they are visits from the same submission. We can see examples in Table 4. Once the queries have been grouped we calculate the average of visited results. Results for the example can be seen in Table 5.

This value is expected to be not very high in less common queries as they will have the upper bound of the number of records in the query log (e.g. `sherlock holmes` only appears once in this example, so its maximum possible value is 1, while `pink floyd` and `indiana jones leather bags` appears in more records).

4.2.1.4 Failed Submissions A *failed submission* is a submission of a query where the user doesn't visit any result (e.g. Submission 1 in Table 4 is a *failed submission*). This measure (shown as percentage of the total number of failed submissions) complements the previous one (average number of visited documents) as it can give us a hint about what queries doesn't

Submission ID	Query	Visited Result
1	`pink floyd`	*NULL*
2	`indiana jones leather bags`	*http://whatpriceglory.com*
2	`indiana jones leather bags`	*http://www.uswings.com*
2	`indiana jones leather bags`	*http://www.saddler.co.uk*
3	`sherlock holmes`	*http://www.sherlockian.net*
4	`pink floyd`	*http://www.pinkfloyd.com*
4	`pink floyd`	*http://www.pinkfloyd-co.com*
4	`pink floyd`	*http://www.pinkfloyd.net*

Table 4
Example of queries from AOL log query grouped by its submission

Query	Submissions	Average of visited results
`pink floyd`	1, 4	1.5
`indiana jones leather bags`	2	3
`sherlock holmes`	3	1

Table 5
Example of queries from AOL log query grouped by its submission

generate appropriate results [2] (at least judged as appropriate by the user), one of the interpretations of the previous metric.

4.2.2 Sessions

4.2.2.1 Session Length Regarding sessions, one of the most commonly measured metrics is the number of queries (session length).

For those authors who define search sessions according to the user's goal a session is formed by subsequent queries issued in order to solve the same informational need. Thus, higher session length can indicate a more complex goal to be fulfilled.

In order to calculate the session length we will count the number of submissions

[2] Of course, a *failed submission* can have another meanings (e.g. the informational need is solved in the results page itself and doesn't need a visit to any result). Unfortunately the impact of these behaviours have not been analyzed and such an analysis is beyond of the scope of the paper

(as defined in Section 4.2.1.3) for each session and then calculate the average for each group of sessions.

Another aspect of length is the time elapsed between the beginning and the end of a session. We will calculate this temporal length as the time elapsed between the initial and the final queries of a session.

4.2.2.2 Visited Results The number of visited results gives us more hints about the complexity of the goal behind the search session. It would be interesting to find correlation between this measure and those presented in Sections 4.2.1.3 (Number of visited results for each query) and 4.2.2.1 (Number of queries for each session).

4.2.2.3 Failed Sessions Similarly to the concept of Failed Submissions (see Section 4.2.1.4) a Failed Session is a session where no result is visited by the user. Analyzing the percentage of failed sessions on each group could give us more details about the tasks of the users.

4.2.3 Groups' Relationships

A first interesting result will be to discover relationships between the group of a session and the groups of the queries included on it. If users' behaviour is related to the group of the queries they submits then changes of query group in a session would reflect changes on the way users face the informational need along the time.

5 Results

5.1 Groups

5.1.1 Groups of Queries

Graphic 2 shows the distribution of queries over the 60 detected groups in a logarithmic scale (so the trends are better appreciated).

This measure seems to follow a power-law distribution [3]. However, detection

[3] Power-laws distributions is the mathematical name for the 'long tail' phenomenon which is associated to a lot of phenomena, some of them related to different aspects of the web (i.e. Newman [20] reports visits from AOL's Internet Service fllowing a

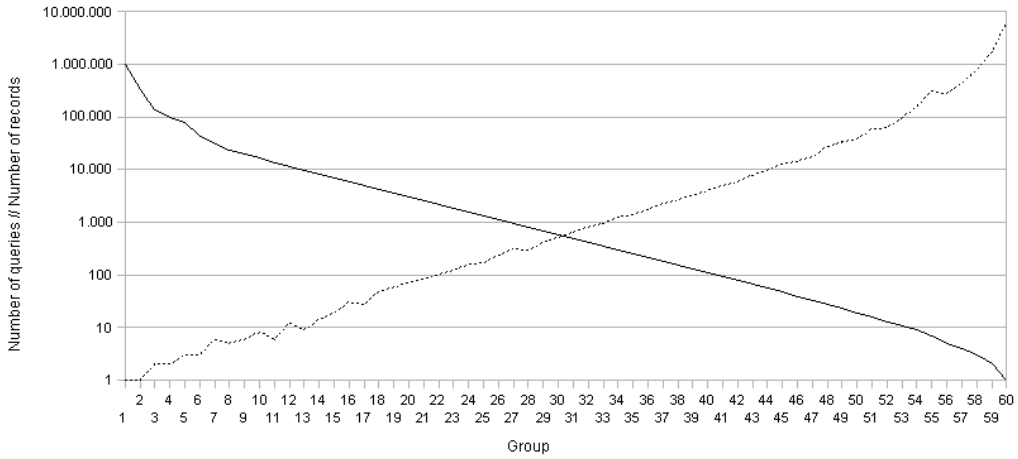

Fig. 2. Distribution of queries along the groups (dotted line) and numbers of records of the initial query of each group (solid line)

of true power-law distributions is not a trivial question [4] and research has been produced on this question (i.e. Newman [20] or Clauset et al. [7]).

The number of apparitions of the initial query of each group also seems to follow this kind of distribution, although it's highly biased by the selected grouping criteria explained in section 3.2.1. The distribution can be seen in figure 2 which shows another power-law distribution on the number of record in the query log related to the initial query of each group.

Despite fitting into a power-law distribution is uncertain, data points out that a small set of queries generates a big proportion of the submissions recorded in the query log. In fact, the second most submitted query appears in the 0.0087% of the log's records, almost 90.000 times more than if submissions were equally distributed.

This could led us to think that paying attention to those popular queries is the most profitable approach. Unfortunately, tail of this distribution is very long and almost half of all the query log's records corresponds to those queries situated in the last 8 groups that have not been submitted more than 11 times.

5.1.1.1 Conflictive groups The *'-' query*, mentioned in section 4.1.1, constitutes the first of the groups of queries. This will allow us to identify it and detect trends without paying attention to the introduced perturbation.

power-law distribution) and has been popularized by Anderson [2].

[4] Although detection of real power-law distribution on query logs could be an interesting research question, due to its complexity authors preferred to postpone it for future research.

5.1.2 Groups of Sessions

After executing the grouping algorithm, 60 groups of sessions were found.

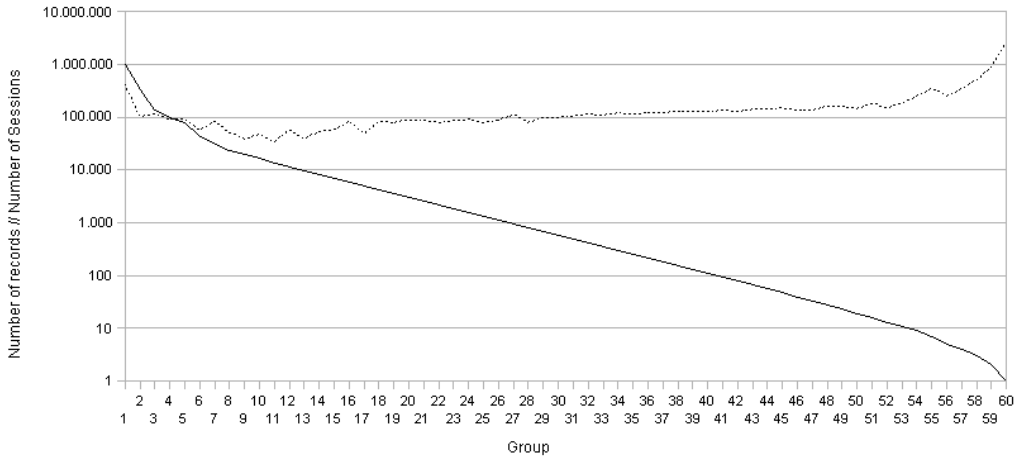

Fig. 3. Distribution of sessions along the groups (dotted line) and number of records of session's first query along the groups (solid line)

Figure 3 shows another power-law graphic regarding the number of records related to the initial query of the first session in the group, very similar to that shown in figure 2.

Instead, sessions' distribution along the query log is different from queries' distribution. This is not surprising too, as the most submitted queries have a larger probability to start a search session [5], so it's not expected that first groups have significantly less sessions than others.

Graphic 3 shows that sessions grouped in first groups usually start less sessions than those situated in latter ones which start most of the sessions (last group accumulates approximately 35% of all the detected sessions and the 6 latter ones groups the half of all the detected sessions).

5.2 Queries

5.2.1 Navigational Coefficient

Figure 4 shows the distribution of the average Navigational Coefficient, calculated on each group of queries.

[5] This probability must be understood in a statistic way in a random session scenario, where no others factors are analyzed than the number of queries and the number of sessions.

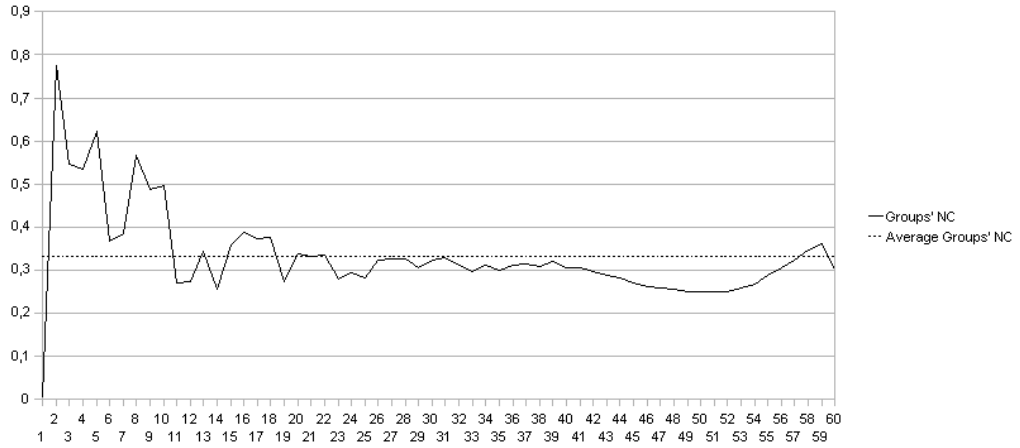

Fig. 4. Distribution of Navigational Coefficient

Data shows that the Navigational Coefficient is higher for the most popular groups of queries [6]. This points out that queries grouped in those groups are more identified with a relevant website by the users. This trend could be stronger for some groups (i.e.groups 6 or 7) but the low number of containing queries makes them very unstable and the introduction of queries with a low navigational coefficient [7] affects the group's navigational coefficient heavily which causes big differences between some correlative groups (e.g. groups 13 and 14 or 18 and 19)

However those differences, a descendant trend can be found which slows down in central groups (from 20 to 40) that show a more constant behaviour.

The latter groups shows a more descendant behaviour but for the latest groups which rises navigational coefficient in 0.1. This behaviour was expected as the less submitted queries were described to have a less meaningful navigational coefficient due to the lack of statistical data.

Figure 4 also shows the average value of the Navigational Coefficient of all the groups. First groups normally get higher values regardless their unstable nature.

5.2.2 Query Length

Figure 5 shows behaviour of the queries' length along the groups. All groups show an upward trend in their number of terms with the exception of some of

[6] With the exception of the aforementioned *'-' query*.

[7] Some of this queries are very common terms which don't have any intuitive website which could turns out into a reference. Some examples of this queries are `internet`, `.com`, `http`, `porn`, `sex`, `m` or `www.`.

the first groups because of their low number of queries.

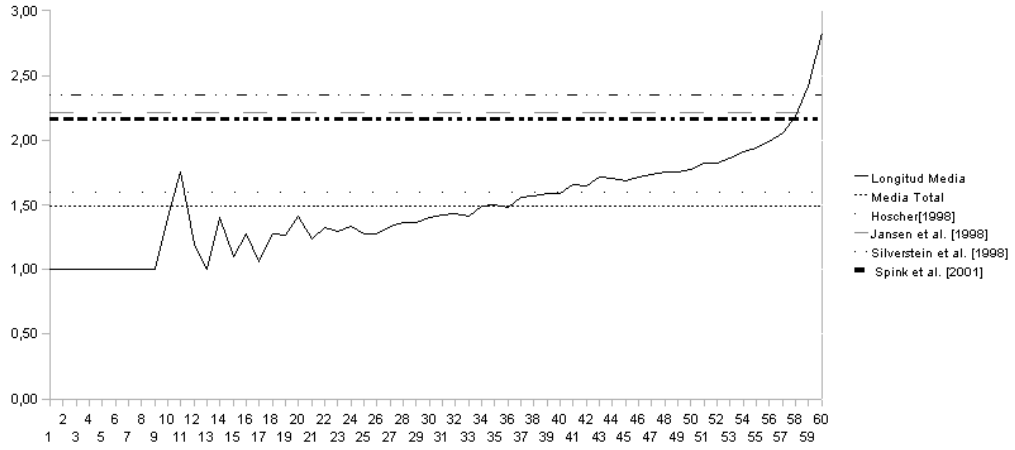

Fig. 5. Average length of queries in the groups compared with previous research

Figure 5 also compares the average length value with values from previous research. It must be noticed that our research doesn't obtain the average length for the query (in contrast to previous research) so comparisons must be cautionary tackled although it points out some valuable information. For example, the average value for the average query length on each group is lower than any of the other measures represented in the graphic. This is because grouping the queries in the latest group we are diminishing their weight.

However, we can observe that not all the groups are well represented by any of the values produced by Jansen et al. [12], Silverstein et al. [25], Hoelscher [10] and Spink et al. [26].

5.2.3 Visited Results

Figure 6 shows how the user behaves visiting retrieved results when submitting queries from different groups. As expected, this behaviour is opposite from that seeing in figure 4 when analyzing navigational coefficient as a high rate of visited results indicates a low navigational behaviour.

In first groups, the average number of visited results is lower, pointing out that queries are solved or abandoned after few visits. This is consistent with data from figure 4 as navigational goals need (if search engine retrieves satisfactory results) less visits to be fulfilled. However, this low value could also be explained by the presence of some non concrete queries which can retrieve few attractive results for the user (e.g. `internet`, `http`, `sex`...).

In the latest groups the average number of visited results rises until it exceeds 2.5 results.

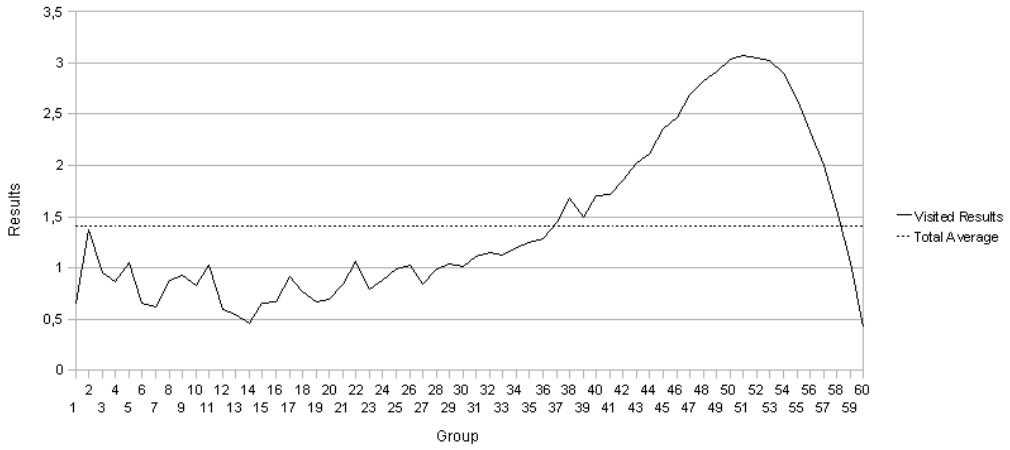

Fig. 6. Average number of visited results

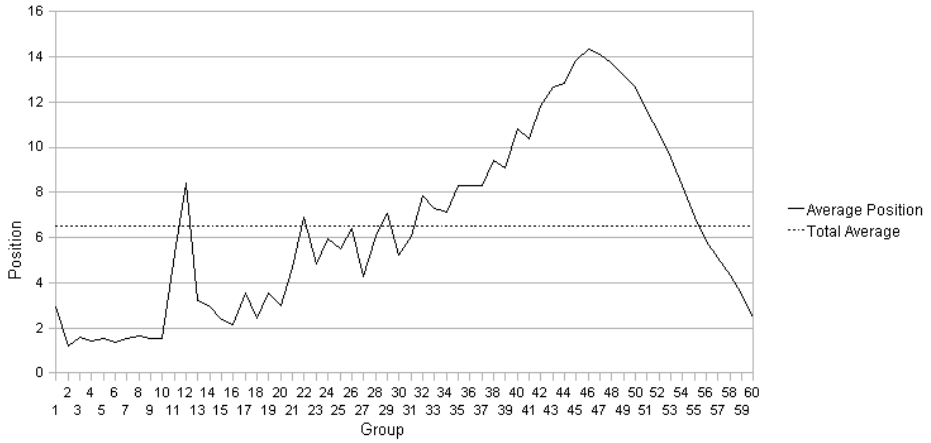

Fig. 7. Average position of visited results

Additionally, figure 7 shows the average position of the visited results which follow a similar behaviour, but shows the highest values in previous groups which points out that although those queries had less visited results, the choice was harder (even making the user to visit the second page of results).

5.2.4 Failed Submissions

Figure 8 shows the proportion of failed submissions (as defined in section 4.2.1.4) in each group.

This data is consistent with those showed in figures 6 and 7 as shows that for most of submissions for most popular queries the users don't visit any result (which is pointed out by the low number average of documents retrieved).

Latter groups present a high rise of failed submissions. Some of these failed

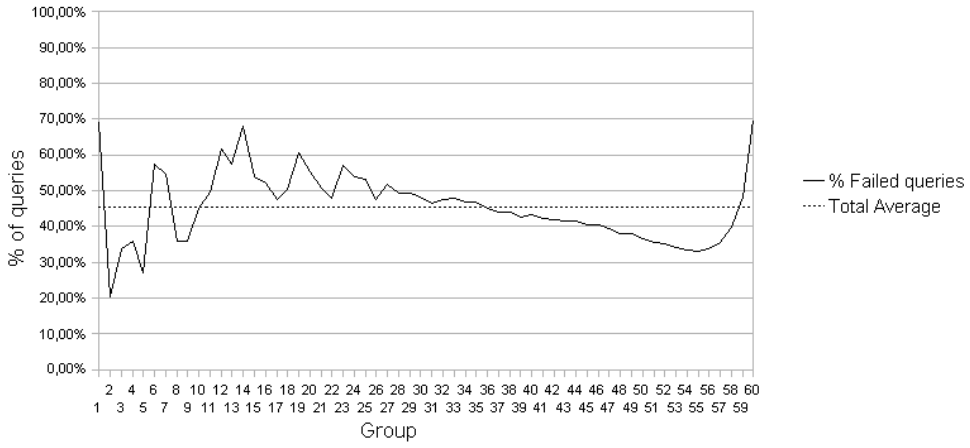

Fig. 8. Proportion of failed queries

submissions can be explained by misspelled queries which are repeated few times and which results are not visited because of the mistake, but the rate of the misspelled queries remains unknown, so we can not assume that this factor has an important effect.

Another explanation for these values is an initial underestimation of the complexity of the query that drives to poor quality queries where no interesting results are retrieved and must be refined without visiting any result.

5.3 Sessions

5.3.1 Average Length

Figure 9 shows that the average session's number of queries along the groups. We can see that variation doesn't seem very significant (almost 1.5 queries between the shortest and the longest sessions) but the trend shows that first groups of sessions consists on sessions which length is somewhat unstable. However, the average length shows a more stable ascending behaviour in following groups which shows that sessions from latter groups are longer.

Results for the first groups are not consistent with the navigational coefficients calculated in section 5.2.1 which pointed out an important navigational behaviour in queries from first groups as doesn't seem reasonable that a navigational query (which goal is associated with the query and easily solved by it) could start a session with more than one query (as the goal was to visit another website).

However, we must not forget that these results are strongly dependent on the session detection algorithm, which is based in a temporal threshold. Although

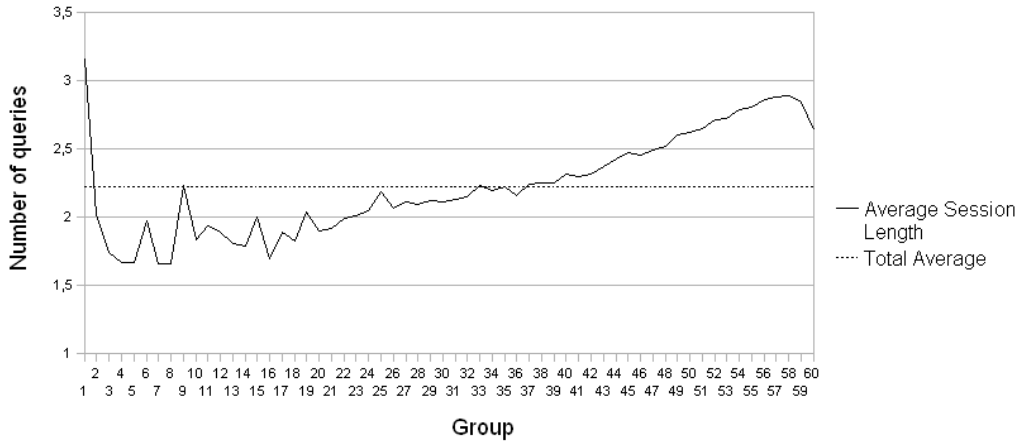

Fig. 9. Distribution of the average number of queries

this method has been widely used, navigational queries could be problematic as they can be easily involved in multitasking (e.g. finding tools or reference websites within another session search) or mixed in other search sessions.

Figure 10 shows the distribution of session's temporal length (time elapsed between the submission of the first query and the last one). This distribution shows more accused differences between first and latest groups.

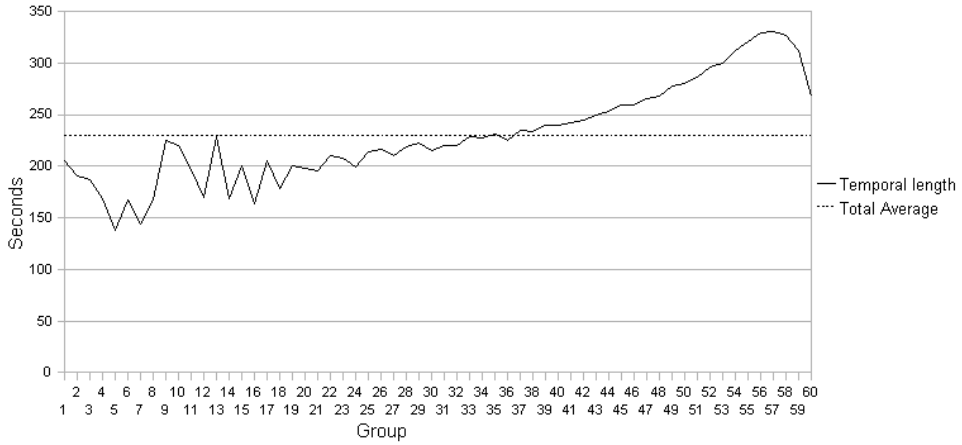

Fig. 10. Distribution of the average temporal length

5.3.2 Visited Results

Figure 11 shows the average number of visited results in sessions of each group.

We can see the same trend shown in figure 6 about the number of visited results per query with the latter groups having the highest value, but the difference between the first groups and the latest ones are less pronounced. This could point out that the evolution of sessions tends to diminish the differences.

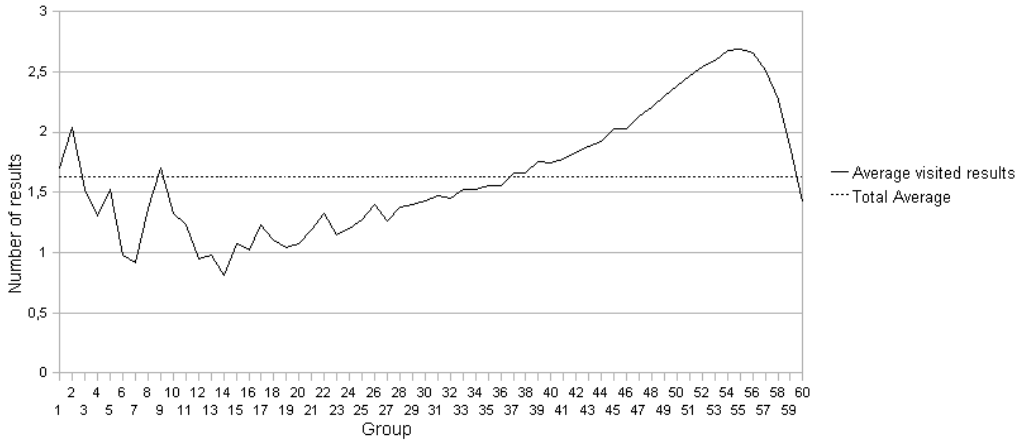

Fig. 11. Distribution of the average number of visited results

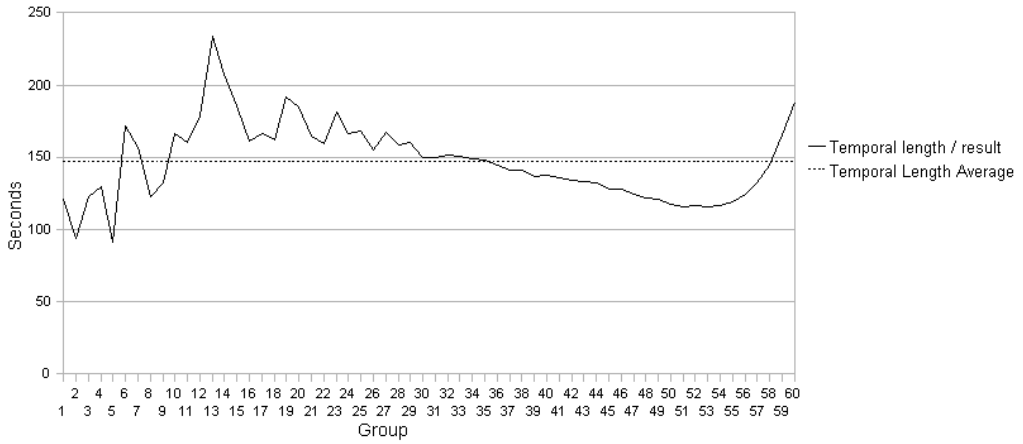

Fig. 12. Time dedicated to each visited result

5.3.3 Failed Sessions

Figure 13 shows the percentage of failed sessions within each group.

We observe that first groups have a small rate of failed sessions (with some exception) but the percentage soon rises and then decreases in a constant rate, being this measure in latter groups lower than in previous groups.

5.3.4 Groups' Relationships

Analyzing how the groups of sessions and the groups of those sessions' queries are related will allow us to find out whether queries in the same session show the same characteristics or whether queries tend to groups with different characteristics.

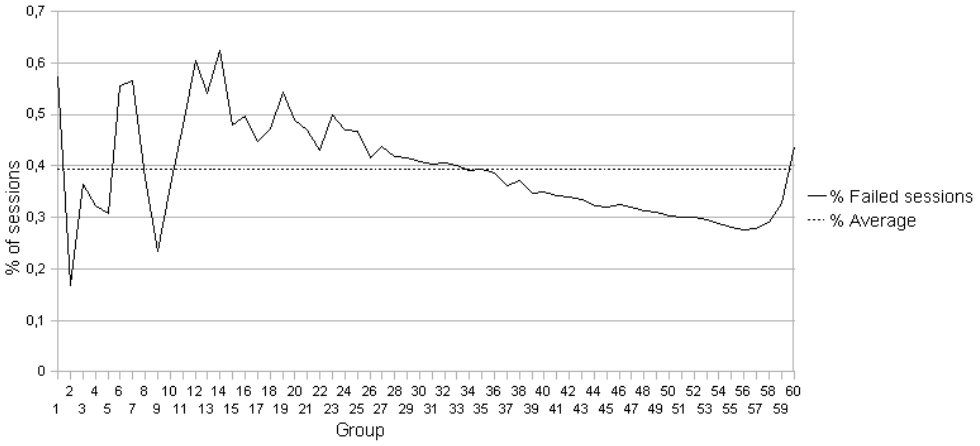

Fig. 13. Proportion of failed sessions

A group of sessions *gs* will be related to a group of queries *gq* if a session *s* grouped in *gs* contains a query *q* grouped in *gq* (and we will refer to this as a *hit*).

To avoid a bias toward certain groups (e.g. initial queries from sessions grouped in *gs* 60 would always have been grouped in *gq* 60, so the relation between those two groups would have been significantly increased) initial queries of each session are ignored. This allow us to observe clearly how the sessions evolve in time.

Figures 14 and 15 shows which groups of sessions are more related with groups of queries and vice versa. In both figures, darker squares means stronger relationships, so the grey square in the intersection of *gs* 15 and *gq* 10 in figure 14 means that there are a remarkable number of queries from *gq* 10 that are submitted in sessions grouped in *gs* 15.

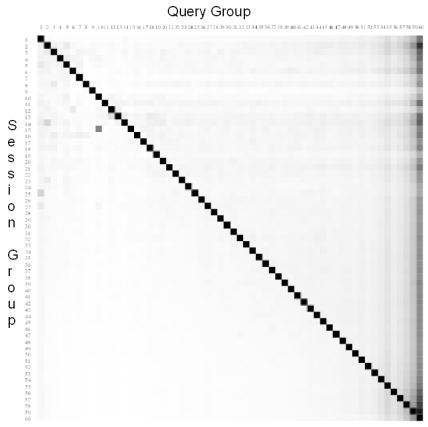

Fig. 14. Percentage of sessions that are related to a group of queries

Figure 14 shows that groups of sessions are mostly related with their equivalent (numerically) group of queries. This points out that characteristics of queries

within a session doesn't tend to change significantly.

Relationships with the latest groups of queries are specially strong. Opposite to this, relationships with previous groups of queries are less usual.

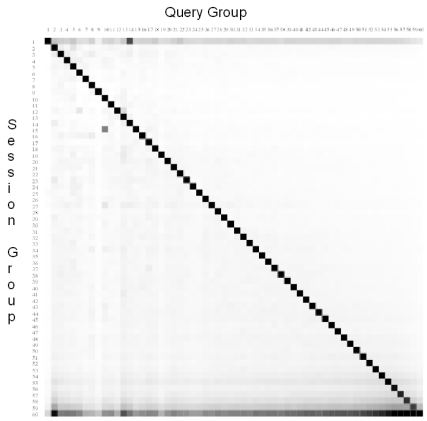

Fig. 15. Percentage of queries that are related to a group of sessions

Figure 15 shows that groups of queries are also prominently relate with their 'equivalent' group of sessions which reinforces the idea that during sessions no radical changes in queries' characteristics are performed.

Conclusions about relationships between queries and sessions are the same as those explained for relationships between sessions and queries (stronger relationships with the groups beyond the 20th and specially with the latest ones). This could seem contradictory (i.e. if latest groups of sessions are not very related with first groups of queries, first groups of queries cannot be highly related with latest groups of sessions) but it has to do with the way data is showed in the graphic.

The relationships between *gq* 13 and *gs* 60 is irrelevant for *gs* 60, as most of its relationships are focused on queries from the latest groups (this is what figure 14 shows), but it's very relevant for *gq* 13 as much of its relationships with groups of sessions are established with *gs* 60. Thus, contradiction doesn't exist.

6 Discussion

6.1 Importance and Complexity of the Long Tail

Uncommon queries are often a problem for navigational aids as they offer few statistical data to use. Mei and Church [18] have pointed out the possibility of ignoring websites without losing much recall precision but this idea is

hardly extensible to queries as almost 96% of the queries have less than 10 submissions. How to extract information about these queries is an important question that deserves deeper research.

Figure 15 shows that queries from this long tail are usually related with sessions started by very different queries and a good number of search sessions are even started by them (as we can see in Figure 3).

Queries from the long tail usually have a higher number of terms which could also give more information to the system compensating for the lack of click-through data in order to extract some information about the user intent.

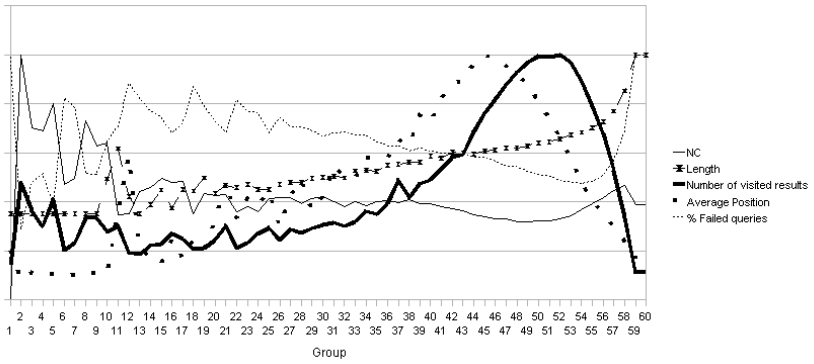

Fig. 16. Summary of the measures extracted about queries after normalizing them

Additionally, queries from this long tail show some behaviours (summarized in table 16)that point out a greater complexity underlying the search. Thus, this queries show a lower navigational coefficient (which shows that visited results are not usually highly related with the submitted query), and higher number of terms and visited results (meaning users have difficulties for expressing their goals in less words and for deciding which results can solve their needs).

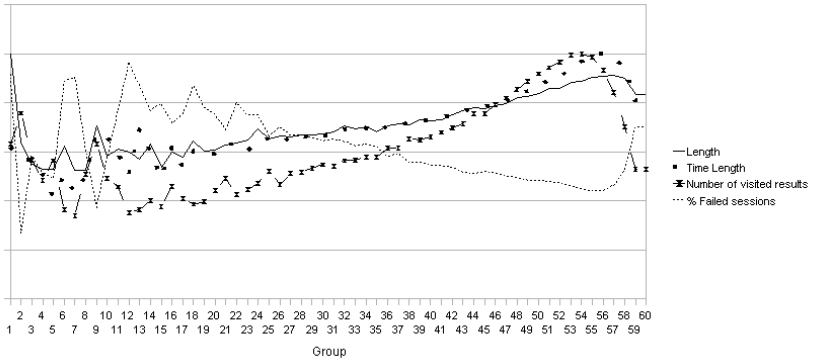

Fig. 17. Summary of the measures extracted about sessions after normalizing them

This trend can be also observed in the results extracted from the groups of sessions (which are shown in 17). In this case the complexity is observed in the

measures of session length (expressed both in number of queries or in seconds) and the number of visited results.

The percentage of failed submissions and queries follows the opposite trend and their percentage decrease in latter groups. This can be explained assuming that the user perceives the complexity or importance of the search and makes an effort to examine some of the obtained results.

6.2 Time consumed examining the results

In previous section we have pointed out that sessions of latest groups are more time consuming that those from previous groups. However, this 'extra-time' seems not to be consumed by an increasing effort on studying search results, as figure 12 shows that differences in time consumed by queries within a session are not really relevant.

This could point out that time dedicated to the analysis of a result is not dependant on how complex the underlying goal is. However, data from query log don't allow us to discard other scenarios such as concurrent visits to multiple results (e.g. tabbed browsing [8] or opening links in a new window) and didn't give information about the time consumed by the latest visited result (it seems pretty reasonable to think that last result could consume more time than other 'false positive' results).

6.3 Evolution of search sessions

Figures 14 and 15 shows how sessions and queries are related. From this relationships we could infer how search sessions evolute.

Sessions have stronger relationships with queries that are placed in higher groups (no matter the group where they have started) , which points out that as sessions evolve users tend to submit longer queries or visit more results. Also, some of these relationships could be explained by misspelling of queries, but measuring the impac of mispelling in a query log is beyond the objectives of this paper.

[8] It must be noticed that in the dates corresponding to this query log multiple tabbed browser had been released (e.g. *Opera*, *Mozilla*, *Konqueror*, *Safari*, etc.) although *Internet Explorer* was mainly distributed in its 6.0 version which didn't include this feature.

7 Conclusions and Further Research

In the introduction and the analysis of previous research we have pointed the weakness of the analysis based on the average user values and introduced some methods used by research community in order to group queries and analyze them. Then we have proposed a new method to group queries based on their popularity and described some experiments which allow us to measure some characteristics along the obtained groups. Results have been discussed and some conclusions have been drawn.

Some research aspects are open so they demand further work which is out of the scope of this paper. As an example, work on confirming the findings of this paper is required to be performed on other available query logs, so the grouping criteria is examined with another data.

Additionally some work on comparing the results extracted for each measure in this paper with results from previous research would be valuable so we could get a deeper understanding on the descriptive value of the previous user's behavior describing attempts based on the existence of an average user. Some findings were shown in Section 5.2.2 when the measured query length is compared with the results from previous researches, but further work is needed.

Work on analyzing some concrete measures is needed too. For example, work on analyzing the real meaning of failed submissions (Pu [22] has analyzed it focused on the image web search) or the impact of misspelled queries would be a very valuable research.

Research efforts could be made in the direction of detecting queries which can bias the actual characteristics of the detected groups of queries and sessions because of its unusual characteristics (e.g. being submitted by bots or reflecting unusual users' behaviours. Some authors have worked on this subject (e.g. Sadagopan and Li [24]) with interesting results.

As a general conclusion, we can assure that our grouping criteria returns very different groups which shows trends describing the complexity in the interactions and goals expressed by the user. However, this criteria needs of more results to be confirmed and analysis of alternative query logs would be a very valuable research.

8 Acknowledgments

We would like to thnak professor Asunción Lubiano-Gómez for many discussions on statistical aspects about data grouping and to the anonymous reviewers for their valuable comments and advice.

References

[1] Eytan Adar. User 4xxxxx9: Anonymizing query logs. In *16th International World Wide Web Conference, Workshop Query Log Analysis: Social and Technological Challenges*, page accessed with no page numbers, Banff, Alberta, Canadá, May 2007. URL `http://www.cond.org/anonlogs.pdf`.

[2] Chris Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, July 2006. ISBN 1401302378.

[3] Nate Anderson. The ethics of using aol search data. online, 08 2006. URL `http://arstechnica.com/news.ars/post/20060823-7578.html`.

[4] Michael Barbaro and Tom Zeller Jr. A face is exposed for aol searcher no. 4417749. *The New York Times*, August 2006. ISSN 0362-4331. URL `http://www.nytimes.com/2006/08/09/technology/09aol.html`. http://www.nytimes.com/2006/08/09/technology/09aol.html?ex=1312776000.

[5] David J. Brenes and Daniel Gayo-Avello. Automatic detection of navigational queries according to behavioural characteristics. In *Woorkshop on Information Retrieval [Accepted for publication]*, 2008.

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

[7] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J Newman. Power-law distributions in empirical data. online, June 2007.

[8] Katie Hafner. Researchers yearn to use aol logs, but they hesitate. *The New York Times*, August 2006. ISSN 0362-4331. URL `http://www.nytimes.com/2006/08/23/technology/23search.html`.

[9] Daqing He and Ayse Göker. Detecting session boundaries from web user logs. In *Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research*, pages 57–66, 2000.

[10] Cristoph Hoelscher. How internet experts search for information on the web. In H. Maurer & R.G. Olson, editor, *Proceedings of WebNet98 - World Conference of the WWW, Internet & Intranet*, pages Published on CD–ROM with no page numbers, Orlando, FL, 1998.

[11] Bernard J. Jansen and Udo Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52:235–46, 2001. ISSN ISSN-3318-3324.

[12] Bernard J. Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic.

Real life information retrieval: a study of user queries on the web. *ACM SIGIR Forum*, 32:5–17, 1998.

[13] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36:207–227, March 2000.

[14] 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.

[15] Steve Krug. *Don't Make Me Think: A Common Sense Approach to Web Usability, 2nd Edition*. New Riders Press, 2nd edition, August 2005. ISBN 0321344758.

[16] Tessa Lau and Eric Horvitz. Patterns of search: analyzing and modeling web query refinement. In *Proceedings of the seventh international conference on User modeling*, pages 119–128, Banff, Canada, 1999. Springer-Verlag New York, Inc. ISBN 3-211-83151-7.

[17] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 391–400, Chiba, Japan, 2005. ACM. ISBN 1-59593-046-9.

[18] Qiaozhu Mei and Kenneth Church. Entropy of search logs: how hard is search? with personalization? with backoff? In *Proceedings of the international conference on Web search and web data mining*, pages 45–54, Palo Alto, California, USA, 2008. ACM. ISBN 978-1-59593-927-9.

[19] Mark R. Meiss, Filippo Menczer, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani. Ranking web sites with real user traffic. In *First ACM International Conference on Web Search and Data Mining*, pages 65–76, 2008.

[20] M.E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, September 2005.

[21] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *The First International Conference on Scalable Information Systems*, pages 1–7, Hong Kong, June 2006. ACM. ISBN 1-59593-428-6.

[22] Hsiao-Tieh Pu. An analysis of failed queries for web image retrieval. *Journal of Information Science*, 34(3):275–289, June 2008. doi: 10.1177/0165551507084140. URL `http://jis.sagepub.com/cgi/content/abstract/34/3/275`.

[23] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19, New York, NY, USA, 2004. ACM. ISBN 1-58113-844-X.

[24] Narayanan Sadagopan and Jie Li. Characterizing typical and atypical user sessions in clickstreams. In *Proceeding of the 17th international conference on World Wide Web*, pages 885–894, Beijing, China, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367617. URL `http://portal.acm.org/citation.cfm?id=1367497.1367617`.

[25] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large altavista query log. *ACM SIGIR Forum*, 33:6–12, Fall 1999.

[26] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52:226–34, 2001. ISSN ISSN-3318-3324.

[27] Amanda Spink, Seda Ozmutlu, Huseyin C. Ozmutlu, and Bernard J Jansen. U.s. versus european web searching trends. *SIGIR Forum*, 36: 32–38, 2002.

[28] Li Xiong and Eugene Agichtein. Towards privacy-preserving query log publishing. In *16th International World Wide Web Conference, Workshop Query Log Analysis: Social and Technological Challenges*, page accessed with no page numbers, Banff, Alberta, Canadá, July 2007.