Analysing Geographic Query Reformulation: An Exploratory Study

A study submitted in partial fulfilment of the requirements for the degree of Master of Science in Information Systems at THE UNIVERSITY OF SHEFFIELD

By Saad Aloteibi

September 2009
Abstract

Background: Search engines are now the prime tool for locating information on the web. Their simplicity and the extensive scope of knowledge they can cover have attracted many searchers to them. A significant group of those users are looking for their information needs within a specific geographical scope, as have been revealed in literature. Yet, the set of words that those users express their needs with representing the linkage between them and the search engine is claimed to be weak. This limitation is evidenced in the literature either from the user’s side, since they often submit short queries and rarely make use of the advance search operators in addition to the inherent problems of natural language, or from the search engine’s side as a result of their ignorance of the spatial context of such queries. This may lead users to repeatedly state their need in different ways.

Aim: This study aims to comprehensively investigate the ways in which users reformulate their geographic needs in an attempt to fill the gap in the literature, which lacks specific query reformulation studies.

Method: A query log of nearly 15 million records was made available for this study. It was filtered from non-human queries to be analysed as a whole in an investigation of the overall sessions’ length. Then reformulated sessions of three or more queries were sampled to extract 2400 entries that were inspected to filter geo-sessions and then analysed quantitatively and qualitatively.

Results: The results revealed that one in five of the users who reformulated their queries were looking for geographically related ones and mostly reformulated their queries by changing the meaning of them rather than their structure. Simple but diverse seem to be the key words to describe geo-modifications behaviour given that users were not following a unified sequence of modifications and mostly performing a single action. However, in some cases it was possible to anticipate their next move. A number of trends in geo-modifications were identified including: normal, multi-needs, multi-places and hybrid approaches. In addition, the statistical test of whether users submitted
more queries when looking for geo-information showed that there was no statistical significant difference between geo and non-geo sessions.

**Conclusion:** The research concludes that users may not always reformulate their queries because of ineffective retrieval methods and more qualitative research in this field is needed to accurately infer such behaviour. It is also important to specialise query reformulation studies to focus on particular topics rather than generically analysing them as it is apparent that geographic queries have their special reformulation characteristics.
# Table of Contents

Abstract ................................................................................................................................. 2  
Acknowledgements ............................................................................................................... 6  

*Chapter 1: Introduction 7–11*

1.1 Research background and context ................................................................................. 7  
1.2 Research aim and objectives ......................................................................................... 9  
1.3 Dissertation structure .................................................................................................. 10  

*Chapter 2: Literature review 12–27*

2.1 Introduction ................................................................................................................... 12  
2.2 The Informative theme .................................................................................................. 13  
   2.2.1 How Users are searching the web? ......................................................................... 13  
   2.2.2 What users are searching for? .................................................................................. 15  
   2.2.3 Why do users use search engines? .......................................................................... 17  
   2.2.4 The Informative theme from a geographical perspective ....................................... 18  
2.3 Query Reformulation .................................................................................................... 20  
   2.3.1 Query Reformulation as Part of General Studies ..................................................... 20  
   2.3.2 Query Reformulation as a Main Research Problem ............................................... 22  
   2.3.4 Query Reformulation from a geographical perspective ......................................... 25  
2.4 The identification of geographic queries ..................................................................... 26  

*Chapter 3: Research Methodology 28–47*

3.1 Background .................................................................................................................... 28  
   3.1.1 Research Theme .................................................................................................... 28  
   3.1.2 Definition of the research terminology .................................................................. 29  
3.2 Research Design ........................................................................................................... 30  
   3.2.1 Data collection ....................................................................................................... 31  
      3.2.1.1 Query logs as an unobtrusive mean of collecting data .................................... 31  
      3.2.1.2 The research data collection phase ................................................................. 32  
   3.2.2 Data preparation: ................................................................................................... 34  
      3.2.2.1 Data Filtering: ................................................................................................. 34  
         The role of the literature in this stage: ................................................................. 34  
      3.2.2.2 Preliminary work with the data ..................................................................... 35
3.2.2.3 Data Sampling .................................................................................................................. 37
3.2.3 Data Analysis ........................................................................................................................ 39

Chapter 4: Results and Discussion 48-67

4.1 Introduction ............................................................................................................................. 48
4.2 Analysis of sessions length ...................................................................................................... 48
4.3.1 The size of geographic reformulation on the web ............................................................... 50
4.3.2 Analysis of Geo-sessions length .......................................................................................... 50
4.3.3 Geo-modification: A view on changes in successive queries’ length ............................... 51
4.3.4 Geo-Modification: A depiction of users actions ................................................................. 52
   4.3.4.1 Overview analysis ........................................................................................................... 52
   4.3.4.2 Detailed Analysis ........................................................................................................... 54
4.3.5 Transition Analysis .............................................................................................................. 58
4.3.6 Analysis of correlated modifications actions ................................................................. 62
4.3.7 Time based analysis ........................................................................................................... 62
4.3.8 Geo sessions reformulation pattern ................................................................................... 63
4.3.9 Trends in geographic query reformulation: ..................................................................... 64
4.4 Results limitations ................................................................................................................... 67

Chapter 5: Conclusion and future work 69-70

5.1 Conclusion ............................................................................................................................. 68
5.2 Recommendations for Future work: ....................................................................................... 69

Bibliography 70
Acknowledgements

I would like to thank my dissertation supervisor Dr. Mark Sanderson for his encouragement and insightful comments throughout this study. I wish also to thank Dr. Barry Eaglestone, the course director, for his help and support. Thanks are also due to Microsoft research lab for making their query logs available for this study.
Chapter 1: Introduction

1.1 Research background and context:

Historically, technological advances play an important role in shaping our lives. To put this into context, the incremental storage and processing capabilities that computers have provided have facilitated the vast production, saving and sharing of this era most valuable asset which is information. Moreover, it is thought that computers have provided a valid solution for the 1940s information challenge that Van Rijsbergen (1979) describes as the ability to access to specific information in a speedy and accurate manner, to which the Information Retrieval (IR) concept has been introduced. Jones & Willett (1997: 1) define "the task of an IR system is to retrieve documents or texts with information content that is relevant to a user's information need". However, IR tools faced a new information-rich environment with the advent of the Internet and its main service the World Wide Web. As a consequence, online IR systems have been developed as a mean of dealing with the web as an information resource which differs these systems from traditional IR systems. One of the most popular online IR tools is search engines. In fact, it might be the dominant way of fulfilling users’ information needs on the web (Kobayashi & Takeda, 2000). Undoubtedly, this reflects the importance of studying how this tool is used in order to improve its performance given that current search engines are claimed to be retrieving irrelevant results for their users in many cases (Shi & Yang, 2007). Chli & Wilde (2005: 372) agree with the former fact by saying that "many of the users find the current process of searching the web unsatisfactory”.

While this dissatisfaction might not necessarily be caused by the search engine underpinning algorithms and, in fact, could be much more a problem of these systems’ users rather than the engine itself, clearly developing these tools to cope with users' problems is of critical importance in enhancing them. One of these users' problems which seem to be the dominant problem is the inaccurate formulation of their information needs in natural language words (i.e. query). This is a problem of IR tools in general, either traditional IR, as Gauch & Smith (1991) emphasise, or in search engines, as evidenced in Chang & Ma (2008) and Cui et al. (2003). These queries are simply the keys for retrieving information, and each key is related to a set of
information that should be retrieved by the search engine when this specific key is entered. Despite how these keys are assigned to information resources, assisting novice users by providing automatic reformulation of their queries to relevant ones is a valid solution to this problem (Gauch & Smith, 1991).

This concept of query reformulation is defined as "the process of altering a given query in order to improve search or retrieval performance" (Jansen et al., 2009a). Most of these automatic tools rely on the users to judge a document as “relevant” so that automatic rewritings can be generated depending on it (i.e. relevance feedback) or on analysing the top retrieved documents without the user’s intervention, which is known as “local analysis”. However, it has been found that users rarely utilize the relevance feedback options (Spink et al., 2000) and usually reformulate their needs manually (Anick, 2003). This might be attributed to many reasons, such as the one stated in Jansen et al. (2009a: 1358) that “the retrieval or interjection of assistance into the search process may be too much of cognitive load, requiring a task switch from focusing on the search process to mentally processing the intervention. Therefore, the searcher may simply ignore any assistance” or it could be related to ineffectively reformulated terms (Anick, 2003). Nevertheless, this certainly shows that it is crucial to study users’ reformulation characteristics in order to provide the base knowledge that any effective reformulation tool can act upon. Liu & Belkin (2008:21) support the former claim by saying that “a deeper understanding of the structure and process of query reformulation, in particular, could provide further information for system adaptations”

The former claims have motivated a number of researchers to focus on analysing the ways in which users reformulate their queries (Bruza & Dennis, 1997; Lau & Horvitz, 1999; Spink et al., 2000; Rieh & Xie, 2006; Liu & Belkin, 2008; Jansen et al., 2009a). However, none of these studies have paid attention to the query types. That is, web queries are classified as local and global queries, where the local ones are best answered within a specific geographic area, whereas the global requests are those without geographical boundaries. It is believed that each one of them has different implications for the reformulation mechanism. Therefore, it appears that the former investigations report results on average users, which is claimed to be a less effective
Chapter 1: Introduction

approach in web searching studies (Brenes & Gayo-Avello, 2009). Thus, the study of query reformulation from a specific angle is needed either to identify the similarities and differences in how each type should be handled or to introduce newer modification characteristics that suit specific queries.

The focus in this study is on local queries or, in other words, geographic queries. The reason behind this is that geo-queries are thought to be more challenging in automatic reformulations and need more attention than global ones for the following reasons. Firstly, search engines are criticised because of their ignorance to the geographical constraints on users’ queries and, therefore, retrieve less relevant results (Gravano, 2003), which might lead users to modify them. This could be attributed to the way search engines handle queries in general as they adopt a keywords matching approach without spatially inferring the scope of the geographic terms. However, it shall be noted that a number of services to deal with this issue have recently been proposed in major search engines (for example Bing local and Google local) but not in the general purpose tools. Secondly, it is evidenced in a number of studies that such queries account for a significant proportion of web searching (Sanderson & Kohler, 2004; Jones et al., 2008). Thirdly, such queries pose significant commercial opportunities for businesses (Gan et al., 2008), who are interested in specific geographic area customers and understanding how and why geo-queries are reformulated in order to develop better advertising techniques.

1.2 Research aim and objectives

The aforementioned information has all motivated this study to thoroughly investigate query reformulation from a geographical perspective in order to understand users’ reformulation preferences so that search engines can effectively incorporate such knowledge into handling geo-queries. For this purpose, we utilized a query log released by a major search engine company containing real life information requests in an unobtrusive manner. We also extended on the work of Jones et al. (2008) who considered modifications on the geo-part of such queries by comprehensively covering the two parts (i.e. geo and non-geo) and qualitatively inferring the sequence of geo-
Chapter 1: Introduction

modification. It is our belief that when studying geo-rewritings, the two elements of such queries need to be considered since they both represent a single information need and, thus an understanding of when, why and how each part is modified is vital.

Therefore, we aim to achieve the following objectives:

- To identify the size of query reformulation in general.
- To examine the significance of geographic reformulation in web searching.
- To examine the difference between geographic and non-geographic searching sessions in terms of their length.
- To determine whether users tend to add, delete or substitute terms when reformulating their geo-needs.
- To investigate geo-reformulation in terms of the actions that users perform on their queries.
- To discover relationships between successive geo-reformulation actions.
- To inspect of geo-reformulation complexity.
- To determine when users are likely to reformulate their queries.
- To find out whether users follow common reformulation patterns or modify their needs in diverse ways.
- To explore geo-reformulation trends.

1.3 Dissertation structure

The remainder of this report is structured as follows:

Chapter 2 presents a comprehensive literature review covering: query logs analysis in general and from a geographical perspective in specific; query reformulation from both general and geographical angles; and studies in geographical terminology on web queries. Its purpose is to provide the base knowledge published in these fields and to identify any gaps and limitations.
Chapter 1: Introduction

Chapter 3 details the research methodological aspects. It presents the research framework and methods along with a discussion of the role of the literature in designing them and their suitableness for query log analysis studies. It provides this information in great detail given the importance of such in this particular field of research.

Chapter 4 describes the study findings and their implications and usefulness. It also relates these results to the published ones where possible. The limitations of these findings are also discussed.

Chapter 5 concludes this study by summarising the main findings and recommendations. It also provides a direction for future work in the field of geo-modifications or search engines enhancement in general.
2.1 Introduction

Exploiting query logs in order to understand searchers’ behaviour and improve search mechanisms has long been acknowledged in the Information Retrieval (IR) literature either for the traditional IR tools, such as in the early work of Penniman (1975), or in a web search engine concept. However, given that the web is claimed to be a new searching environment (Jones & Willet, 1997; Silverstein et al., 1998; Jansen & Pooch, 2001; Jansen et al., 1998; Park et al., 2005) and that common facts about searching in traditional IR tools are not entirely applicable to search engines (Brenes, Gayo-Avello, 2009), we have focused on covering web-based studies since this is the general theme of this research. A comprehensive review of the other theme is provided in Peters (1993). In the following, we always refer to web-based query logs that are extracted from search engines.

Query logs Analysis (QLA) studies can be classified into two main research streams: Informative and supportive. The informative studies tend to describe how users are searching the web, what they are looking for, and why they search. This informative classification is supported in Rose & Levinson (2004). On the other hand, research under the supportive umbrella uses query logs as a rich-resource that can be used directly to enhance the performance of services, such as query suggestion (Cucerzan & White, 2007), query expansion (Cui et al., 2003) and query rewrites (Jones et al., 2006). Of course, there is an overlap between the two concepts since a method for effective query rewrites, for example, would be based on searchers reformulation characteristics that are found in the informative studies.

In the following, we shall present notable studies from the literature in order to base the grounds of our study. In the first section, we highlight the informative stream from the aforementioned three angles. Then, a placement of geographical query logs studies in the scope of the informative questions is presented. In the second section, we discuss query reformulation findings, followed by geo-modification studies. Finally, we shed light on the identification of geographical queries. We limit the review to the informative stream since it is the major theme
that our research is classified under and many of its findings deserve the reader’s attention.

2.2 The Informative theme

2.2.1 How Users are searching the web?

Brenes & Gayo-Avello (2009) claim that the majority of investigations in this classification “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”. In other words, they are statistical summaries of users' interactions with search engines. The first major study in this category was conducted by Silverstein et al. (1998), who sampled about 1 billion search requests to the AltaVista search engine, from which they reached a conclusion that Internet searchers usually use short queries (average of 2.35 terms per query) and only review the top 10 results. Similar findings are reported in (Jansen et al., 1998; Spink et al., 2001; Park et al., 2005). Such results might not be surprising to IR researchers since novice users are known for generally and imprecisely formulating their information needs; however, Silverstein et al. (1998) found that users occasionally modify their queries. Such a fact could indicate that the majority of information seeking processes are successful and that the precision of search engines is high. However, the absence of analysing the click-through data in Silverstein et al. attempt, which could made us assume that users needs have been satisfied by reviewing some of the retrieved results, made the assumption that a considerable number of searching sessions were terminated without solving users information problems dominant. However, Silverstein et al.’s observation in regard to sessions length might be skewed by the session identification algorithm, which assumed that users’ inactivity for 5 minutes determined the end of a session. Clearly, such a tight delimiter would decrease the number of queries per session. This argument is supported by He & Göker (2000), who claim that a session threshold of less than 10 minutes is likely to increase the number of sessions with one entry significantly and decrease the percentage of sessions with more than 1 query. They state that the optimal session delimiter is within the range of 10-15 minutes, meaning that Silverstein et al.’s limit is too short to be accurate. Jansen & Spink (2006) support this claim. Their study also suggests
that search engines’ performance could be enhanced by considering the queries as part of phrases since their terms correlation analysis shows that the mostly correlated words are constituents of phrases.

In another series of studies conducted on the Excite search engine, a wider range of observations about Internet users searching behaviour was reported (Jansen et al, 1998; Jansen et al., 2000; Wolfram et al., 2001; Spink et al., 2001; Spink et al., 2002a). For instance, their dataset allowed them to investigate the use of a relevance feedback feature provided by the search engine \(^1\). They reported that less than 10% of the queries came from this service, which indicates that users might not understand the potential benefit of using such a tool. Nevertheless, the actual figure might be fewer than this since Excite stores a query resulting from using the relevance feedback as an empty query. Clearly, this would have biased the results given that users might actually have submitted empty queries. This argument is supported in Jansen & Pooch (2001). Another interesting result is that they found query structure to be very simple, rarely containing Boolean operators or other search modifiers, such as + and “ “, and if it contained any search modifiers would be the ones. However, there are two major limitations of these studies. First, they did not report sufficient information about their session’s identification algorithm. This resulted in contradictory results in regard to the size of reformulation on the web. Specifically, they reported in 1998 that the mean session length was 2.8, whereas their figure for 2001 was 4.86. In 2002, their number decreased to 2.3 queries per session. Having such discrepant findings makes the actual session length questionable. Secondly, all their data were collected from a single day log. This is thought to be a very short time frame to reflect a wider picture of Internet searching.

In 2006, Zhang & Moffat (2006) took a different approach than the former studies. They analysed a query log of MSN search engine containing searchers click-through data, which allowed them to discover facts about users’ interaction with search engines. One of their results of relevance to our study is that the majority of users make their reformulation decisions within seconds, indicating that users are often not certain of the query that best represents their information needs and that they have more than one formula in mind, which they try sequentially.

\(^1\) Excite provided an automatic reformulation tool called “More Like This” which reformulates the user query depending on the chosen document and.
Chapter 2: Literature Review

In addition, they noted that some of the users might wait for 20 minutes or less to modify their queries. However, this result shall be viewed with caution since it is thought that they over-filtered their raw data by deleting all the queries that did not have a click-through record and hence those who waited for 20 minutes might have actually submitted queries in between without reviewing any of the results.

Another body of research has tried to focus its attention on a group of users depending on either their nationality, such as in the work of Spink et al. (2002a) for U.S and European searchers, Efthimiadis (2008) for Greeks, Park et al. (2005) for Koreans and Chau et al. (2009) for Chinese, or on their query characteristics, such as in the work of Bendersky & Croft (2009), who focused on long queries, or Spink & Ozmultu (2002) for question format queries. Another grouping criteria is dependent on the perceived users’ needs, such as geographic needs in Sanderson & Kholer (2004).

2.2.2 What users are searching for?

An early study which was directed towards answering this question was conducted by Ross & Wolfram (2000) using an Excite search log. They revealed that adult-related material was the most requested topic on the web, whereas documents related to business and computing were ranked in the middle of their 30 category list. However, they built their classification scheme depending on the mostly co-occurrence pairs of terms on which they argue that “certain topical domains have very narrow vocabularies, at least in usage, and so tend to generate words with a disproportionately high rate of repetition” (Ross & Wolfram, 2000: 955). Thus, although their results might give a rough indication of users’ interests on the web, it is thought that they might have underestimated topics with broad terms space, such as “places” and “education” even if they are ranked among the top 5 areas.

In 2002, Spink et al. (2002b) extended the former study by comparing samples extracted from the same search engine in 1997, 1999, and 2001. They found that users’ interests on the web had shifted from sex and pornography to business related topics. Specifically, they reported that “Commerce, travel, employment or
Chapter 2: Literature Review

economy” and “People, places or things” categories encompassed the most frequent queries. Nevertheless, no proper definitions of these groups were mentioned and nor did they report their study methodology so it is difficult to interpret what they mean by “things” in their “people, places or things” group. A later study by Jansen & Spink (2006) confirms these findings and concluded that these topics are becoming increasingly popular for U.S. based search engines and European ones. Such a trend is also found in Spink et al. (2006).

Pass et al. (2006) used different categorisation criteria to the previously mentioned studies. They showed that queries related to entertainment and shopping were among the mostly asked queries. Computing, news, health, business and places were similarly popular in the middle of their figures, whereas holidays and finance related requests were the least demanded. Moreover, they stated that entertainment and adult related queries were mostly common in the early hours of the day and personal finance ones increased during working hours. Nevertheless, no information was reported about their classification methodology or about their research sample data.

In summary, the previous studies have highlighted the necessity for having a well-defined classification of web queries since it seems obvious that the past studies varied in their categorisation approaches. Thus it is difficult to compare their findings and to generalise which domain dominates users’ interests on the web. However, there are two major lessons to be learned from the above. First, they revealed that users’ needs on the web are quite diverse (Pass et al., 2006), which Silverstein et al. (1998: 9) also support by saying that “the fact that almost two-thirds of all queries are asked only once in a 6 week period indicates that information needs on the web are quite diverse, or at least are specified in diverse ways”. Secondly, they show that the web has become a major resource of information on a variety of topics and more than an entertainment tool (Jansen & Spink, 2006).
2.2.3 Why do users use search engines?

A remarkable study in answering this question was conducted by Broder (2002), who argues that the classical IR belief that informational needs are mostly the driver of searching activities using IR tools is not applicable in the search engines context. He claims that informational queries make up less than half of web queries. The other half are placed either as navigational queries, in which “the immediate intent is to reach a particular site”, or as a transactional one, where “the intent is to perform some web-mediated activity” (Broder, 2002: 5). One of Broder’s claims is that 20% of navigational queries “have usually only one ‘right’ result” (Broder, 2002 :5). What is interesting about this claim is that recently Microsoft search engine, known as “Bing”, has adopted Broder’s view by only retrieving one result called “Best match” for what they interpret as navigational queries. For example “Yahoo mail” query would only show one result leading to “Yahoo mail” website. It could be argues that Google service “I’m feeling lucky” has also adopted this approach (Rose & Levinson, 2004), but it does not differ between informational or navigational as it directs the user to the first ranked result, regardless of the query type.

Rose & Levinson (2004) highlight the potential benefits of understanding the ‘why’ factor. They claim that the user interface and relevance-ranking algorithm could be improved depending on the underlying user goal in order to provide a customized search experience. They state that such an objective could be reached through three steps: firstly, by constructing a conceptual framework of users’ goals; secondly, by automatically detecting the user goal from the queries; and, finally, “modify the engines in order to exploit the goal information” (Rose & Levinson, 2004:13). Broder’s and Rose’s & Levinson’s attempts merely focus on the first aspect and provide quite similar classification from a high level with more detailed view in Rose & Levinson’s work. Jansen et al. (2008) extended the later framework and also proposed a method for inferring users’ intent behind the queries. They analysed random samples from three different search engines in order to identify special characteristics that could be used to differentiate navigational, informational and transactional queries. Their results showed that each type has a set of terms that are usually associated with their queries and that navigational queries and sessions
Chapter 2: Literature Review

tend to be shorter than informational ones. From this they built a database of these terms and used it in their automatic classifier. Their classification algorithm achieved a success rate of 74% and, as they emphasised, it underestimated navigational and transactional requests in favour of informational needs. They argue that it might be the optimal classifier so far to be implemented in a real time environment. Nevertheless, a limitation of this work is that it assigns a query to one group only (Jansen et al., 2008), but other research indicates that informational queries could also be regarded as partially navigational queries since users might have a specific website in mind that they want to fulfil their informational needs from (Tann & Sanderson, 2009).

To sum up, these studies show that search engines are used in a different manner than is generally assumed in traditional IR literature by having two goals, navigational and transactional, along with the informational one. Thus, the view that search engines could be regarded as the gateway to the internet (Amitay & Broder, 2008) could be supported. This argument might also be supported by the claim that Internet searches often repeat their queries (Sanderson & Dumais, 2007; Teevan et al., 2007), which indicates that this tool is used as “a starting point” in a web-based activity, regardless of whether it is an informational, navigational or transactional task.

2.2.4 The Informative theme from a geographical perspective

The first study to focus on geographically related queries was done by Sanderson & Kohler (2004). They answered the questions of what users are looking for when submitting a geographic query and partially how they search for it. Based on a human analysis method of an Excite query log, they revealed that geo queries represent nearly one fifth of the overall queries, with 80% of those geo-queries containing a place name (a percentage of 14.8%). Similar findings in regard to queries with place names are reported in Gan et al. (2008) with a percentage of 13% and 12.7% in Zhaung et al. (2006). Sanderson & Kohler (2004) argue that geo-queries are 25% longer than the average query with a mean of 3.3 terms. They relate it to the structure of geo-queries as having geo and non-geo parts, with the geo-part occasionally containing spatial or directional terms. Their results also show that
users’ interests when searching for geographically related information do not significantly differ from the general trend, as reported in Spink et al.’s (2002b) study in previous section. In addition, it is claimed that there is no difference between geo and non-geo queries in terms of the use of advance search features or for the occurrence of spelling errors (Kohler, 2003). However, although Kohler (2003) argues that geo-queries reformulation is not a common practise, a limitation of this work is that it does not draw a conclusion on the average length of geo-sessions if this had been done, it would have provided a deeper insight into the ‘how’ element and, also, would have statistically assisted the comparison with the general studies.

Jones et al. (2008) studied a 4 million queries extracted from Yahoo! search engine in 2006. They confirmed Sanderson and Kohler (2004) results in regard to geo-queries characteristics and also added that users largely prefer to search on a city level more than a country or a state level. However, Sanderson & Han (2007) argue that searching on a city level is not the dominant narrowing factor in geography searching as countries or states. These facts are not presented for comparison purposes as it was already emphasised that they are derived using two completely different approaches (Sanderson & Han, 2007) but rather for showing that users are willing to limit their search using a geographic entity regardless of its class, and that these terms are not used only for the sake of fulfilling informational needs about them. In other words, Kohler (2003) found that 15.9% of geo-queries solely represent a spatial area, whereas in the rest it could be argued that geo-terms are used to focus the intention on searching for an object on a specific place on the earth. They also found a correlation between the topic of the geo-query and the user location since queries about “High school” or “restaurants” tend to be searched in a local sense despite others such as “Map” or “Hotel” (Zhaung et al., 2008).

In the same context, Gan et al. (2008) analysed a query log of AOL containing about 36 million queries. They found that the likelihood of tagging a query as geo-query increases as the number of terms in the query enlarges. They also attempted to answer the question of why users search for geo-information and reached an unsurprising conclusion that “geo queries are more frequently aimed at locating goods and services” (Gan et al., 2008: 51), which is under the resource category in Rose & Levinson (2004). In addition, they state that some of the geo-queries were navigational queries, in which the user is either trying to distinguish
a site from another by using a geo-flag, such as in “Google uk”, or filtering a particular site content according to the location of interest (e.g. “craigslist Los Angeles”). They also propose a new 23 geo-queries category list. However, it is thought that their classification is rather wide since it mixes two, arguably, completely different aspects of general queries. Specifically, the why aspect, which is best exemplified by Broder’s taxonomy and queries topics. For example, one of their categories is “Medical”, which is defined as “hospitals, doctors, and general health and medical information” (Gan et al., 2008: 52). Certainly, a query from this topic could be placed in another class of their headings, such as “advice”, since the intent behind the query might be to look for advice on a particular health problem, and, meanwhile, it is a “Medical” query. They do not provide a justification for this overlap between the user intent and the query topic even though they acknowledge that their classification is a hybrid between these two elements.

2.3 Query Reformulation

Query reformulation has received a great deal of attention in the QLA literature either as a part of more general studies, such as in Silverstein et al. (1998), or as a main focus for others, such as in Spink et al. (2000). In the following, we shall present query reformulation research according to the said headings.

2.3.1 Query Reformulation as Part of General Studies

Silverstein et al.’s (1998) study gives us an insight into users’ reformulation actions as they reveal that 12% of reformulated queries are a result of adding or deleting a term or a search operator to the previous query. Specifically, they state that the addition of words is often coupled with the use of “-” operator in an attempt to restrict the retrieved documents. Meanwhile, it is shown that over half of users reformulate their queries by using a mixture of adding and deleting tactics, by which they assume that “the user is modifying the query not to change the scope of the query, but to restate the information need” (Silverstein et al., 1998: 11). However,
the limitations of this study discussed in section 2.2.1 clearly have a negative effect on their query reformulation results.

Jansen et al. (2000) explored users’ reformulation behaviour by examining successive queries in order to detect the changes in the number of terms. They assumed that if the follow-up query has (n) number of terms more than the previous query, it represent as an increase by (n) terms and vice versa. In cases where the number of terms is identical between the two queries it is regarded as 0 change since users might add and delete terms at the same time, as shown in Silverstein et al. (1998). They found that 34.76% of users prefer to substitute terms rather than increase or decrease them. A similar figure was found in Spink et al. (2001). The remaining percentage is divided nearly equally between adding and deleting practises in a decrement trend towards a higher number of terms, which is contrary to Spink et al.’s (2001) findings that 41.6% of users modify their queries by adding terms and 25.9% by deletion. Spink et al.’s (2001) findings shows that Internet searchers tend to go from general queries to more specific ones. However, it should be noted that such figures are highly affected by the session identification procedure and that neither Jansen et al. (2000) nor Spink et al. (2001) provide information about the way they identified them. Also, interestingly, these two studies were conducted on Excite query log; as Jansen et al. (2000) studied a sample of 51,473 queries collected on 9th of March 1997, whereas Spink et al.’s (2001) sample contained 1,025,910 queries collected on 16th of September 1997. This shows a considerable change in users’ reformulation behaviour in a period of about six months.

Park et al. (2005:215) draw a different conclusion from Excite studies. They found that “more users changed their queries entirely rather than adding terms to or deleting terms from the previous queries”, which is related to Silverstein et al.’s (1998) observation that 35.2% of users change their subsequent queries completely. However, it is worth mentioning that Park et al. (2005) studied a Korean search engine, which might show that reformulation approaches differ depending on the users’ ethnic group.
2.3.2 Query Reformulation as a Main Research Problem

In an attempt to conduct more comprehensive query reformulation study, Lau & Horvitz (1999) took advantage of the 1997 released Excite query log. Their study focused on building probabilistic models that capture users’ refinement actions. In order to do that, firstly they proposed 7 types of reformulation: new, generalisation, specialisation, reformulation\(^1\), interruption\(^2\), request for additional results and blank queries\(^3\). These classes were then assigned to the relevant queries manually, and they found that users refine their queries by either asking for additional results or starting a new search as opposed to expanding or contracting them. However, it could be argued that “request for additional results” by reviewing more pages is not a reformulation action since the request has not been changed and an obvious result of submitting a query is reviewing a result. Secondly, they looked at the relationship between the said classes and the probability of their occurrence in terms of the time interval between the query and its subsequent from which they constructed a Bayesian network model to predict users’ next reformulation action.

Spink et al. (2000) extended Jansen et al. (2000) by categorising reformulated queries into four groups. These are: unique query, modified query, next page and relevance feedback. Their results are consistent with Lau & Horvitz (1999) since over 70% of the reformulated queries were either a request for the next page or unique, with the former representing 46% of choices. However, it is acknowledged that users might actually submit repeated queries during the session, which might mean that Spink et al.’s (2000) figures are on the high side. In addition, they examined query patterns during search sessions, from which they argue that the most common pattern of query reformulation is to type a unique query followed by viewing more results.

---

1- Reformulation as defined by the researchers refers to “a query on the same topic that can be viewed as neither a generalization nor a specialization, but a reformulation of the prior query”.
2- Interruption refers to “A query on a topic searched on earlier by a user than has been interrupted by a search on another topic”
3- Blank queries were not used in their analysis which makes its inclusion in the category above questionable.
However, Rich & Xie (2006: 755) argue that the above studies deal with reformulated queries in relation to its former only in which they claim that “the interactive IR process is cyclical in the sense that both query terms and search results are constantly selected, evaluated and modified. ... This leads to the belief that an understanding of query reformulation sequences poses a significant research problem”. Thus, as a contribution to answering this research question, they extended previous work (Rich & Xie, 2001) by studying more sessions and merging their conclusions, from which they discovered that modifications lie in three main areas. These are: content in which the user changes the query meaning; format when the structure of the query is modified but not the meaning; and resource when choosing between different document types. These three classes were then divided into 9 sub-classes in order to capture more specific reformulation patterns. As a result, they proposed eight different forms of reformulated sessions. These are presented in table 1. However, although their study provides an in-depth analysis of query modification, it might be limited with the fact that they define a reformulated session as having 6 or more queries which seems to be relatively high and thus they might have ignored some patterns which are worthy of notice for shorter sessions.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Definition as in Rich &amp; Xie (2006: 759)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specified reformulation</td>
<td>“refers to those sessions in which a user persists in specifying previous queries.”</td>
</tr>
<tr>
<td>Generalized reformulation</td>
<td>“refers to those sessions in which a user begins with a narrow concept and continues on to generalize through successive trials”.</td>
</tr>
<tr>
<td>Parallel reformulation</td>
<td>“refers to those sessions in which a user modifies the queries from one aspect of an entity to another or from one thing to another, both of which share common characteristics.”</td>
</tr>
<tr>
<td>Building-block reformulation</td>
<td>“refers to sessions in which a user identifies and combines the concepts from the previous queries and uses them in subsequent queries.”</td>
</tr>
</tbody>
</table>
Chapter 2: Literature Review

<table>
<thead>
<tr>
<th>Dynamic reformulation</th>
<th>“refers to those sessions in which a user employs inconsistent patterns to increase specificity or generality, moving around from one type of query reformulation to another.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-tasking reformulation</td>
<td>“refers to those sessions in which a user looks for two or more topics simultaneously in the same search session.”</td>
</tr>
<tr>
<td>Recurrent reformulation</td>
<td>“represents cases in which a user enters exactly the same query that has been already used two or more steps previously.”</td>
</tr>
<tr>
<td>Format reformulation</td>
<td>“refers to those sessions in which a user attempts to figure out how to formulate his/her queries correctly in terms of format.”</td>
</tr>
</tbody>
</table>

Table 1 (Query reformulation patterns in Rich’s & Xie’s (2006) study)

Among the former patterns, it is worth mentioning that Bruza & Dennis (1997) state that the most frequent modification process is to repeat a query. However, they did not differentiate between repeated queries in a sequential and non-sequential order, and nor Rich & Xie (2006) provide statistics about their patterns. It is interesting that such modification behaviour still represented even after such a long time period, which indicates that users’ modification behaviour might be static.

Jansen et al. (2009a) studied query modification from a different perspective be the aforementioned investigations. They utilized the concept of a state transition map to infer when users are most likely to seek a help from the system and to which reformulation state they will move. Their research was motivated by the fact that searchers rarely make use of reformulation tools, as evidenced in Spink et al. (2000). They provide algorithms to categorise query modifications states and also to segment query sessions automatically. Although their algorithms are not robust in the sense that identifying a session or a reformulation state based on analysing the query terms only is not usually the case with search queries since a user might expresses a single information need using different terms in various levels of abstraction, they made a notable contribution in this field.
2.3.4 Query Reformulation from a geographical perspective

In her general geo-queries research, Kohler (2003) provides a snapshot of geo-reformulation behaviour. She extracted the reformulated sessions that contained the place name “London” and analysed them depending on the occurrence of the addition/deletion of terms in successive queries. Her conclusion does not significantly differ from that of Spink et al. (2001). Moreover, it was noticed that the addition of more geo terms in the query is commonly used to differentiate between places that share the same name, and this is acknowledged in Jones et al. (2008). Nevertheless, such analysis is very limited to one particular place, making it is difficult to get a generalised view about geo-modification.

Another brief mention of geo-modification is made in Gan et al. (2008). They argue that such action is not a common practise in web searching and in cases where it occurs the changes are likely to appear in the non-geo part of the query. Nonetheless, it seems that they paid little attention to this issue and do not provide the analysis method from which they derived these results, which makes them more like assumptions than a research findings.

Jones et al. (2008) appears to be the only research initiative in examining geo-modification as a main research problem. They sampled queries from Yahoo! search engine in order to discover whether users have different geo-modification characteristics depending on the location that they are looking for. Their results showed a difference in geo-rewritings in terms of the distance between the modified geo-places depending on the location of interest. However, this was limited to only two states in the US and thus cannot be generalised as geo-reformulation behaviour. They also evaluated an automated reformulation tool that uses query logs as a source of candidate words (Jones et al., 2006) in terms of the suggestion of geo-rewrites and found that its result is deemed to be relevant when the geo element of the query is unchanged compared with its former.
2.4 The identification of geographic queries

Since most of the transaction logs available are for generic search engines, an initial stage of research in this field is to define and extract those queries that have a geo-intent. Different approaches to tackle this, perhaps challenging (Zhuang et al., 2008), issue have been proposed in the literature. In the following, we shall present them in incremental order based on their perceived complexity, starting with a straightforward method that was adopted in Kohler (2003). Her approach was to inspect the queries manually, looking for terms that were geographically inductive, such as place names, adjective of places and place locators. This approach might be highly accurate but nevertheless infeasible to be used for larger query logs samples, which often contain millions of lines. This infeasibility is either because manual checking is not an option or applying this in an automated manner by using Named Entity Recognition algorithms (NER) would affect the accuracy negatively since simply looking up from a database of geographic terms does not necessarily ensure that the query is about geography or not even if it contains a geo-term (i.e. false positive problem). An obvious example of this problem is the overlap between people and places names. In principle this is not a new concept and has been discussed in geographic information retrieval literature but query logs are more difficult since queries are very short and users enquire search engines in a variety of ways, as shown previously (Martins et al., 2006).

Jones et al. (2008) assigned a score to each term in the query and, similarly, a “location-related probability” value to places in a geographic database. The final judgement as to whether a query was geo or not was affected by the former values along with an inspection of another database of context words and non-place names (e.g. “George Washington”). When the query was tagged as geo, they checked the user IP address and the frequency of occurrence for that particular place in order to identify the intended location in cases when many geographic entities exist with the same name. However, although this approach might decrease the probability of having a high rate of false positive queries, clearly it is used only for “place names” detection and not geographic queries identification. This is because it is evidenced that less than 50% of geo-queries contain a place name (Welch & Cho, 2008). Implicit geo-characteristics in search queries are also discussed in Yi et al. (2009).
Chapter 2: Literature Review

Martins et al. (2006) used two techniques to label geo-queries. For explicit ones, they argue that it could be viewed as a triple of “<what, relation, where>”. The “what” refers to the non-geo component of the request, “where” is the location of interest and “relation” is the spatial term that links the other two parts. For implicit geo-references, they mine the geographic scope of the retrieved documents that matches the query. However, such an approach in terms of the triple is very limited since it appears that they dealt with the geo-queries as having an identical structure, which is usually not the case. In other words, a user looking for a hotel in London might express his need with a query that matches their triple “a hotel in London” but others may type it differently by saying “London’s hotels”, where the relation aspect is eliminated but implicitly pronounced. This is common in geo-queries, but they do not provide information about handling this issue.

Similar to Martin et al. (2006), Henrich & Ludecke (2007) grouped geo-queries into three classes. The first is named as “Topostate”, where related queries are constructed from four constituents. These are: the non-geo part of the information need; a preposition; a toponym; and the name of the state that the toponym is contained within. Clearly such a structure is limited to US places and also suffers from the same limitation discussed in relation to Martin et al. (2006). The second and third classes are for queries that contain optionally a preposition with a state name only or a toponym only. Again, such an approach is for place names identification but not for all geographic queries, as shown previously.

A novel solution to this problem was developed by Wang et al. (2005) which deals with implicit and explicit geo-queries effectively. They coined the term “Query Dominant Location”, which they define as “A QDL is geographical location(s) associated with a query in collective human knowledge”. This term could be related to the fact that geo-queries are ambiguous in the sense that a geo-term could be shared by more than one place, that is “referent ambiguity” (Tjong et al., 2003), or it could be used in a non-geo situation known as “referent class ambiguity” (Martins et al., 2006). Thus, solving these issues could be of importance in identifying the exact intended location. Wang et al. (2005) also attempt to identify the implicit geo-queries by mining the top retrieved documents.
3.1 Background:

3.1.1 Research Theme

It has been claimed that most IR investigations are qualitative and quantitative in nature even though they appear in the first instance to be quantitative (Robertson & Hancock-Beaulieu, 1992). In the QLA realm, this claim might be more applicable giving its definition “as the use of data collected in a search log to investigate particular research questions concerning interactions among web users, the web search engine, or the web content during searching episodes” (Jansen, 2009: 102). Academically, these studies are viewed under the behaviourism umbrella (Jansen et al., 2009b) since the intention is pointed towards modelling users’ behaviour, that is answering questions regarding what they search for on the web and how they react in a given situation, for example when not presented with relevant documents. In this context, using quantitative tools would be useful for quantifying trends in user searching but insufficient in drawing insightful conclusions regarding their behaviour, for which qualitative methods are best. Borgman et al. (1996) suggest that a combination of the aforementioned approaches would yield more productive findings. Strauss & Corbin (1998: 29) argue that research is a “flow of work” in which evaluating the appropriateness of different research methods to particular aspects of the study is crucial. Certainly, the former argument reveals that the decision of choosing between these two research paradigms or combining them is driven by the research questions.

In the meantime, others argue that QLA studies follow a ground theory approach (Jansen, 2009). This indicates the qualitative aspect of QLA given that ground theory is the “pre-eminent qualitative research method” (Bryant & Charmaz, 2008:1). This argument could be supported with the fact that the existing QLA studies derive their findings and conclusions during the course of analysing the data, that is an inductive approach, rather than formalising hypothesis to be tested, that is deductive approach. In support of this view, Ozmutlu et al. (2009: 211) argue that “Most of the studies available in search behaviour analysis and transaction log analysis are based on EDA (Exploratory Data Analysis)” which others regard it as a qualitative method even if quantified figures are presented (Naoum, 2007).
Certainly, the decision as to which research theme was to be implemented and which approach was to be followed in our study was vital since the existing literature on this field does not provide sufficient information about their methodologies and neither of the above scholars agree on the theme of QLA. Nevertheless, given that we are generally concerned with quantifying the size of geographic reformulation along with an inspection of geo-modification, we adopted a centric solution by combining quantitative and qualitative methods in an inductive framework for the sake of drawing a comprehensive picture of geo-modification on the web.

3.1.2 Definition of the research terminology

Jansen & Pooch (2001) highlight the importance of defining the terminology used in QLA studies in order to pave the way for valid comparison between different studies. Hence, we derive our terminology from previous research in the field wherever possible.

1- **Term**: “A term is any unbroken string of characters (i.e. a serious of characters with no space between any of the characters)” (Jansen et al., 2000: 212). Thus in this research, search operators were dealt with as terms.

2- **Query**: “A query consists of one or more search terms” (Jansen et al., 2000: 212).

3- **Initial Query**: “The first query in a session” (Jansen & Pooch, 2001: 244).

4- **Reformulated Query**: “A subsequent query in a session that is different than any of the previous queries” (Jansen & Pooch, 2001: 244). Such a difference between queries occurs when the user restates his/her request in several ways but the underlying information need is unchanged.

5- **Repeated Query**: A subsequent query that is identical to the former one. These queries are ignored since there are two interpretations for this type: either the user is rewriting the exact query again or reviewing the next pages on the result page (Jansen et al., 2000). The dataset does not allow us to judge between these two activities and, therefore, it is excluded from our analysis.

6- **Session**: A session is defined by (Jansen et al., 2000: 211) as “the entire serious of queries by a user over a number of minutes or hours. A session
could be as short as one query or contain many queries”. However, it is rather difficult to be certain that a session is related to only one user since the assignment of session identifiers depend on the user IP address or the use of cookies and this does not always ensure the identification of one user given that different users may share one IP address in large ISPs or may share one machine on which the cookie is stored. There is a body of research which tries to segment users’ sessions either depending on a time cut, such as, in He & Göker (2000), or depending on terms analysis for a query and its subsequent, such as in Jansen et al. (2009a). The method used by Microsoft in their query log provided to this research is a time cut of 20 minutes. That is if the user status is idle for 20 minutes, then the session is deemed to be finished. However, it shall be noted that this procedure is not a foolproof and we have identified some sessions of time-cut more than the 20 minutes which might be due errors in the program used.

Thus, we could define a session for the purpose of this study as “a series of queries submitted by a user, or maybe users, within a timeframe of 20 minutes, mostly, between a query and its subsequent query”.

7- **Geographic query**: “A search query with a geographic intent”. This intent could be explicitly pronounced as part of the query or implicitly pronounced (Wang et al., 2005).

### 3.2 Research Design

QLA studies usually consist of three main phases, which are: data collection, data preparation and data analysis (Jansen, 2009). The first stage might be standardised in that specific data entities are gathered depending on the software used at the search engine side in contrast to the usual procedure followed in research where data are gathered depending on the study questions (Jansen, 2009). The other two stages might be divided to sub-activities according into the investigation direction. In the following, we provide a detailed view of the procedures followed in this research along with the role of the literature in designing them.
3.2.1 Data collection

3.2.1.1 Query logs as an unobtrusive mean of collecting data

QLA is deemed to be an unobtrusive technique for collecting research data (Borgman et al., 1996; Park et al., 2005; Jansen et al., 2009). Such an approach does not require the direct involvement of the users in the research context in contrast to the other methods. Jansen et al. (2009: 8) argue that the alternative approaches can “lead the actors to alter their behavior in order to look good in the eyes of the researcher or for other reasons”. In the same context, Borgman et al. (1996) highlight the significance of analysing query logs when observing searchers’ behaviour, either using the traditional or the new generation of IR tools. Given this, we could argue that data contained within query logs are, in most cases, representative of the “real world”. An example of such is described in (Jansen et al., 2009) when comparing the rarity of adult-related queries in a laboratory study with the richness in QLA studies. Additionally, query logs have the advantage of gathering an enormous amount of data with relatively low cost from a wide range of users.

Notwithstanding the former strengths, the unobtrusive nature of this method certainly raises privacy concerns, either in respect of the user identity (i.e. their IP address) or their queries, which might contain personal information (Cooper, 2008). From the researcher’s point of view, bringing the “life” of users into the hands of unknown parties to them certainly is ethically debatable. However, query logs’ providers have usually dealt with this matter by sharing anonymous data across the field of academia in which this concern is mitigated, but not eliminated (Jones et al., 2007). Nevertheless, providing that an understanding of the ways by which search engines are used is crucial in enhancing them and query logs provide data that would not be available otherwise (Amitey & Broder, 2008), the release of such data might be viewed as a two-edged initiative. From another angle, analysing such anonymous data has its limitations due to the fact that they “do not record the underlying situational, cognitive, or affective elements of the searching process” (Jansen, 2009: 104).
Chapter 3: Research Method

3.2.1.2 The research data collection phase

A query log of Microsoft Live search engine was made available to this study under the “Microsoft live lab: Accelerating Search in Academic Research” plan in 2006. The data was collected over one month, starting from 1/5/2006 to 31/5/2006. It contains about 15 million queries from US users. This log was provided with the click history of each query, if any, in a separate file (i.e. click-stream data). Table 2 provides a general description of these two files, and tables 3 and 4 give a detailed view of each one, with a snippet of the data provided.

<table>
<thead>
<tr>
<th>File</th>
<th>Number of queries</th>
<th>Number of sessions</th>
<th>File Format</th>
<th>File Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query log</td>
<td>14,921,285</td>
<td>7,470,915</td>
<td>Text file (.txt)</td>
<td>1,098,422 KB</td>
</tr>
<tr>
<td>Click stream</td>
<td>12,251,067</td>
<td></td>
<td>Text file (.txt)</td>
<td>1,239,599 KB</td>
</tr>
</tbody>
</table>

Table 2 (General description of the data provided)

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-stamp</td>
<td>The time a query was submitted in this format yyyy-mm-dd hh:mm:ss</td>
</tr>
<tr>
<td>Query</td>
<td>Text field containing the query submitted by the user.</td>
</tr>
<tr>
<td>QueryID</td>
<td>A unique identifier for each query. This field is used to link the queries with the click-stream associated with each query.</td>
</tr>
<tr>
<td>SessionID</td>
<td>A unique identifier for each session.</td>
</tr>
<tr>
<td>Result Count</td>
<td>The number of results on result page. This definition is the one provided by Microsoft; however, it should be noted that this field is not used in this study.</td>
</tr>
</tbody>
</table>

Table 3 (Fields in the file)
In this section, we provide 4 records from this file to aid clarification.

<table>
<thead>
<tr>
<th>Time-stamp</th>
<th>Query</th>
<th>QueryID</th>
<th>SessionID</th>
<th>Result Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-05-08 15:14:23</td>
<td>windows media streaming</td>
<td>fd97569e9810430c</td>
<td>042efcae005f40ba</td>
<td>12</td>
</tr>
<tr>
<td>2006-05-05 02:44:32</td>
<td>morning call</td>
<td>6d7de11cf5db4148</td>
<td>0394866960f141e1</td>
<td>12</td>
</tr>
<tr>
<td>2006-05-05 02:03:47</td>
<td>short leaf trees</td>
<td>61bfe647bc9443eb</td>
<td>0670fe06d83748f2</td>
<td>10</td>
</tr>
<tr>
<td>2006-05-05 02:04:11</td>
<td>assorted maple</td>
<td>3e9fb1bbdc064aeb</td>
<td>0670fe06d83748f2</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 3 (General description of the query log file)

In this section, we provide 3 records from this file to aid clarification.

<table>
<thead>
<tr>
<th>QueryID</th>
<th>Query</th>
<th>Time</th>
<th>URL</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>000009c2848e4a68</td>
<td>north hills school district</td>
<td>2006-05-04 12:29:12</td>
<td><a href="http://www.nhsd.net/">http://www.nhsd.net/</a></td>
<td>1</td>
</tr>
<tr>
<td>000020de950d43c7</td>
<td>yahoo</td>
<td>2006-05-31 10:51:01</td>
<td><a href="http://www.yahoo.com/">http://www.yahoo.com/</a></td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4 (General description of the clicks file)
3.2.2 Data preparation:

3.2.2.1 Data Filtering:

The role of the literature in this stage:

“The web query logs usually contain queries from both human users and agents” (Jansen, 2009:108). This fact highlights the need for thorough consideration of the effect of having two completely different user groups in our study. Since we were merely concerned with human users, we ought to have filtered the data from the non-human queries. In the following, a brief review of the procedures followed in the literature is presented.

- Ignorance was an approach in Silverstein et al. (1998), who did not clean the data from agent queries. As a result, they got quite a large standard deviation in regard to the number of queries per session (123.40). They acknowledged this problem and its effect on their results.
- Filtering queries that did not have a click through record was adopted in Zhang & Moffat (2006). Clearly not all human users would have clicked on a result as well.
- Jansen (2009) suggests a method of cleaning all the sessions that have a length of over 100 queries.
- Park et al. (2005) state that they statistically analysed the time gap between the queries in a session to identify the non-human queries. No information was provided on the time gap used in the separation process.

The filtering method applied:

As shown above, the choice of filtering non-human queries is rather a “question of over-filtering vs. Under-filtering” (Grimes et al., 2007). Our approach was to apply Jansen’s (2009) suggestion by deleting all the sessions that had more than 100 queries. Although this is not a foolproof option, it does reduce the effect of noisy data and, hence, introduce outliers to our findings given that a human submitting over 100 queries in one session is unlikely. On the other hand, agent might submit under this limit.
Chapter 3: Research Method

Table 5 shows a statistical summary of the data before applying this filter and Table 6 shows the results after applying it. Table 7 shows the effect of this function on the final number of queries and sessions.

<table>
<thead>
<tr>
<th>Number of queries</th>
<th>Number of sessions</th>
<th>Slandered Deviation</th>
<th>Max. number of queries per session</th>
<th>Min. Number of queries per session</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,921,285</td>
<td>7,470,915</td>
<td>12.5151543774557</td>
<td>32824</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5 (The data before filtering agent queries)

<table>
<thead>
<tr>
<th>Number of queries</th>
<th>Number of sessions</th>
<th>Slandered Deviation</th>
<th>Max. number of queries per session</th>
<th>Min. Number of queries per session</th>
</tr>
</thead>
<tbody>
<tr>
<td>14,810,989</td>
<td>7,470,610</td>
<td>2.23979827892491</td>
<td>100</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6 (The data after filtering agent queries)

- In the following, we always refer to the data in Table 6 as the “raw dataset”.

<table>
<thead>
<tr>
<th>Queries deleted</th>
<th>Sessions deleted</th>
<th>Percentage of sessions</th>
<th>Percentage of queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>110,296</td>
<td>305</td>
<td>0.004%</td>
<td>0.745%</td>
</tr>
</tbody>
</table>

Table 7 (The effect of filtering agent queries)

Filtering at the Microsoft Lab

Privacy concerns: As emphasised by Microsoft, e-mail addresses and sequences of numbers longer than 8 digits were replaced with special characters. No information was provided about the users’ IP addresses. Although this preserved user privacy, it limited this in that we could not identify the distance between the user location and his/her geographic query.

3.2.2.2 Preliminary work with the data

An attempt to automatically extract geographic queries was made adopting Martin et al.’s (2006) triple “<what, relation, where>”. Since the “What” cannot be determined, the “relation” and “where” could be, to some extent. We used a
Chapter 3: Research Method

database containing 1,884,611 geographic features in the US, taken from Geonames server, and a defined set of 73 spatial terms, for example in, near and north. Firstly, we filtered the data to extract queries that contain a spatial term, and then we filtered the result against the geographic references database as an application of NER\(^1\). However, this approach proved to introduce a high false positive to our potential geo-dataset although a similar approach was followed in Henrich & Ludecke (2007). This led to the following:

- The identification of geo-names certainly requires a disambiguating algorithm and this is an area of research in its own right.
- Our aim is beyond the former objective.
- Such approaches only capture an explicit geo-query, but they have to be coupled with another approach to extract implicit geo-references.

We chosen not to proceed with this dataset and to apply a sampling option where human identification is feasible. Table 8 shows a snippet of three records of the result obtained.

<table>
<thead>
<tr>
<th>Id</th>
<th>Time</th>
<th>Query</th>
<th>QueryID</th>
<th>SessionID</th>
<th>Result_Count</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2006-05-19 07:21:00</td>
<td>anthem insurance blue access in ky</td>
<td>91bc68c3996d4019</td>
<td>0a72317b07944481</td>
<td>16</td>
<td>Geo</td>
</tr>
<tr>
<td>2</td>
<td>2006-05-17 16:18:37</td>
<td>percentage expression in access</td>
<td>370d56dab9f348ef</td>
<td>1e97a8036e7247cf</td>
<td>11</td>
<td>Not Geo</td>
</tr>
<tr>
<td>3</td>
<td>2006-05-19 06:12:57</td>
<td>switch between tabs in excel</td>
<td>ccc9ced503d14b1f</td>
<td>0f6c4084550f42d5</td>
<td>13</td>
<td>Not Geo</td>
</tr>
</tbody>
</table>

Table 8 (Results obtained from the preliminary work stage)

The above table illustrates the significance of applying algorithms to disambiguate geo-references by determining the Query Dominant Location (QDL), as suggested by Wang et al. (2005). In the previous example, query 1 was about “Access”, as in the city in Kentucky State, whereas it was about the computer program in query 2. Query 3 also follows Martin’s et al. (2006) triple but nonetheless have the same problem as 2 since Excel is a city in Alabama State. Other examples, such as “Kentucky” queries, have been shown in the literature.

---

1- Named Entity Recognition. This approach is explained in chapter 2.
3.2.2.3 Data Sampling

Sampling the dataset was not a straightforward procedure since a determination of the population, reformulated queries, needs to be made initially. That is to represent a thorough consideration of geo-modification; we have to extract those queries that have been reformulated. Rich & Xie (2006), who studied query reformulation in general, decided to consider sessions that had more than 5 queries and ended up with 313 search sessions. Spink et al. (2000) qualitatively analysed 191 sessions that had two or more queries. Jones et al. (2008), who have studied geo-reformulation, labelled 108 queries which have passed four different filtering algorithms. It should be noted that this aspect of their study was only concerned with reformulating the geo part of the query to another location. Such filtering procedures in these studies were highly sensible, in principle, given the fact that not all the users would have reformulated their queries and when sampling the raw data, there is a high chance of obtaining irrelevant items, namely sessions with one query.

The approach followed in this research is twofold:

**First**: We filtered the raw dataset to extract all the sessions that had at least three queries. The intent behind this was to include the sessions where users persevere in modifying their queries to give an insight, arguably, into the cognitive processes followed by users. This resulted in a new dataset, which we refer to it as “reformulated queries”. Table 9 shows statistical summary of this dataset.

<table>
<thead>
<tr>
<th>Number of queries</th>
<th>Number of sessions</th>
<th>Max. number of queries per session</th>
<th>Min. Number of queries per session</th>
</tr>
</thead>
<tbody>
<tr>
<td>7,511,984</td>
<td>1,502,577</td>
<td>100</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 9 (Reformulated queries)
Section 3: Research Method

**Second:** Given the following formula

\[
* \text{Initial Sample Size} = \frac{Z^2 \times P \times (1 - P)}{C^2}
\]

* Final Sample Size after the correction of finite population

\[
= \frac{\text{Initial Sample Size}}{1 + \frac{\text{Initial Sample Size} - 1}{\text{Population}}}
\]

Where:

Z is the confidence level. Our confidence level is 95%, which corresponds to a value of 1.96.

P is the response distribution. By default it is 0.5.

C is the confidence interval (i.e. the error rate). We have an error rate of 2%.

Population is our reformulated queries. Its value is 7,511,984

We needed a sample size of 2400 queries. These 2400 queries were then randomly extracted from the reformulated queries representing different search sessions. This size was thought to be sufficient given the time limit we had to conduct the study and the fact that we intended to extract all the geo-sessions that had a representative query in the sample. In other words, if we found a query that either explicitly or implicitly referred to a geographic location, we would trace the session related to it and extract it in preparation for studying the reformulation process.

*This formula was calculated using the website [http://www.macorr.com/ss_calculator.htm](http://www.macorr.com/ss_calculator.htm) and rechecked in another website [http://www.raosoft.com/samplesize.html](http://www.raosoft.com/samplesize.html). The former gives a result of 2400 and the latter 2401.*
3.2.3 Data Analysis

This stage was solely driven by the research objectives. In other words, to answer the questions we had raised in chapter 1, it was highly important to analyse the data with them in mind so that we achieve their answers. The remaining of this section is divided according to our objectives.

**Objective 1:** To identify the size of query reformulation in general: Due to the contradictory in the findings published regarding this issue in some cases or the limitation of others resulted from their session identification algorithms, as discussed in our literature review, a reasonable starting point for research in query reformulation is to identify the size of this practise. By macro-analysing the data looking for the number of queries per session, we were able to draw a conclusion of the distribution of sessions’ length. This objective uses the raw dataset.

**Objective 2:** To examine the significance of geographic reformulation in web searching: Our approach at this stage was as follows:

**Firstly:** The aim was to identify queries that had a geo-intent, either explicitly, that is the query has a geographic reference, or implicitly (i.e. the query does not have a geo-component but nevertheless refers to a spatial place). This classification is supported in Gan at al., (2008); Martins et al.,(2006); Yi et al., (2009) and Wang et al.,(2005). For all of them an examination of the query terms, the query session and the session click-history was made to ensure that we did not miss places unknown to the researcher or implicit geo-queries. In cases where neither the session nor the click history provided sufficient information about the query intent, we checked the top retrieved documents for the query from the same search engine “Bing”. This approach was thought to be highly accurate given the fact that we had checked our sample twice for the sake of completeness.

**Secondly:** Bearing in mind that a reformulated query has to be coupled with another query, at least, which shared with it a single information need and that Rieh & Xie (2006) state in their research that user may shift between different topics in one session, we checked if the geo-query had resulted from a reformulation process or not. If not, then the query is not a reformulated query, and thus it is excluded from the geo group.
Chapter 3: Research Method

Thirdly: We traced the sessions of the identified geographic queries for the purpose of analysing their sessions.

Table 10 shows a sample of the geographic queries obtained after the filtering processes.

<table>
<thead>
<tr>
<th>Query</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regency inn</td>
<td>This is an implicit geo-query. By examining the session history, we found that the user is looking for this hotel in the city of Sewanee in Tennessee State.</td>
</tr>
<tr>
<td>Boston Airport</td>
<td>Explicit geographic query.</td>
</tr>
<tr>
<td>iad</td>
<td>Refers to Washington Dulles International Airport. To be sure that this was the intent, the other queries on this session was about Vancouver International Airport and Vancouver weather forecast.</td>
</tr>
<tr>
<td>helen bernhart bakery</td>
<td>A bakery in Portland, Oregon State.</td>
</tr>
</tbody>
</table>

Table 10 (Example of queries tagged)

Objective 3: To examine the difference between geographic and non-geographic searching sessions in terms of their length: For this purpose, we compared the length of geo-sessions identified in objective 2 with the length of non-geo sessions from our 2400 sample by extracting their sessions. Applying the statistical test (T-test) we were able to see whether there was a difference between the two or not.

Objective 4: To determine whether users tend to add, delete or substitute terms when reformulating their geo-needs: In this stage, we applied Spink et al.’s (2000) method by considering the following:

- If the successive query had the same number of words then the change was deemed to be (0 changes).
- If the successive query had (n) number of words more than its former then the change was deemed to be (+n changes) and vice versa.
However, a challenge at this stage was the interruption of a reformulation process by navigational queries. It is acknowledged that searchers might use search engines for the purpose of reaching a particular site (Broder, 2002; Rose & Levinnson, 2004) and thus they may find it easier to type a navigational query to a site (mostly yahoo, MySpace and Google) while they still reformulate their geographic need. By including these queries, it might skew the result since they are mostly one word queris and such activity could happen by other means, for example use of bookmarks. Thus, it was decided to exclude them so that we could get a better understanding of the accurate figures.

**Objective 5:** To investigate geo-reformulation in terms of the actions that users perform on their queries.

Studies in query reformulation have usually dealt with this issue by proposing a number of modification actions that capture the different ways in which searchers reformulate their needs (Riech & Xie, 2006; Spink et al., 2000; Jones et al., 2008; Bruza & Dennis, 1997; Lau & Horvitz, 1999). We adopt Riech & Xie’s (2006) classification loosely in a geographic sense. That is, since there is a consensus in the literature that a geo-query has two elements: the object which searchers are looking for and the location in which they are interested (Gan et al., 2008; Martins et al., 2006). We handled the queries as thought they had these two parts. This segmentation process was done by sketching the query into this statement “I am looking for ‘an object’ in ‘the location name’”, where the object is the non-geo part and the location is the geo part. Of course, not all the queries followed this, but the majority did. By doing this, we were able to examine whether users often reformulate the geo or the non-geo component. Table 11 shows the classification used in this study from three levels. The reason for having these levels is twofold:

- To represent geo-modification from different levels of abstraction.
- To avoid the “long tail” effect by having many specific categories.
Chapter 3: Research Method

Such classification allows us to investigate geo-modification in-depth to explore the following:

- The probability of users moving from one state (i.e. reformulation action) to another by constructing a transition matrix. This approach is supported in Pass et al. (2006) and Spink et al. (2000).
- The probability of applying a specific query reformulation action depending on the waiting time after submitting a query.
- The sequence of reformulation actions that users often follow.
### Level 1: Modification to the query content (“Content Level”)

**Definition:** “Refers to those instances in which users made changes to the meaning of a query”. (Riech & Xie, 2006: 756)

#### Level 2: Modification to the content of the geo element of the query (“Geo Content Modification”).

#### Level 3: Geo Generalisation

**Definition:** This action refers to the cases in which the user either deletes a term from the geo part of the query (e.g. “city of charlotte fire department → charlotte fire department” and “Northampton county unemployment → unemployment”) or changes to a wider location (e.g. “Richmond” → “Texas” and “Detroit” → “Michigan”).

#### Level 3: Geo Specialisation

**Definition:** Searchers specify their search by either adding terms to their geo part (e.g.” Sea Restaurant” → ”Sea Restaurant AND Brooklyn” and “holiday inn” → “holiday inn fair grounds Monroe wa”) or moving from wider location to a more specific one (e.g. “Ohio” → “Cincinnati” and “Illinois”→“Champaign”).

#### Level 3: Change the Geo entity

**Definition:** Refers to the cases in which the user shifts between two locations with the condition that one is not contained in the other. (e.g. “Colorado” → “California” and “San Bernardino” → “New Bedford”).

#### Level 2: Modification to the content of the non-geo element of the query (“Non-geo Content Modification”). In the generalisation and specialisation cases it is similar to the geo definition, but here we were interested in the non-geo part instead of the geo.

#### Level 3: Non-geo Generalisation

E.g. (“rocky mountain newspaper Colorado” → “newspaper Colorado” and “life insurance company of virginia, richmond” → “life of virginia, richmond”).
### Level 3: Non-geo Specialisation

E.g. (“remax altoona,pa” $\rightarrow$ “remax agents in altoona,pa” and “Volkswagen dealers” $\rightarrow$ “Berg Mazda Volkswagen”). The latter is specifying the dealer he/she wants to Berg Mazda which is a car dealer in Arizona State. This user has a follow up query as “Scottsdale Az Volkswagen”, which clearly shows that this is a geographic query. Berg Mazda serves the city specified in the final query “Scottsdale”, which might show that this user is interested in this particular city.

### Level 3: Change the non-geo entity

**Definition:** Refers to those cases when the change made to the non-geo part is neither specialisation nor generalisation but has an overlap in meaning with the previous query. E.g. (“bread stores in bay area, california” $\rightarrow$ “bakeries in bay area, california” and “Directions from Dallas to Navarro” $\rightarrow$ “Mileage from Dallas to Navarro”)

### Level 2: Interrupting the sequence of a reformulation process. (Interruption). The concept of this category was derived from Lau & Horvitz (1999).

### Level 3: Returning to a previously submitted query (Return query)

**Definition:** Refers to those cases when users re-submit a query that has already been investigated within the same session in a non-sequential order. (e.g. “SD State Parks” $\rightarrow$ “Camping-Chamberlain, SD” $\rightarrow$ “SD State Parks”)

### Level 3: Start new search with the same geo part

**Definition:** When the user changes his/her information need but is still searching in the same geographic location.
### Chapter 3: Research Method

<table>
<thead>
<tr>
<th>Level 1: Modification to the query format (“Format Level”)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition:</strong> Include the cases in which users made changes without altering the meaning of the query”. (Riech &amp; Xie, 2006: 756)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2: Modification to the format of the geo element of the query (“Geo format Modification”).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 3: Geo error correction</strong></td>
</tr>
<tr>
<td><strong>Definition:</strong> When users correct spelling errors of the geo-component. E.g (“state of nee york workers compensation” → “state of new york workers compensation”)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3: Geo term variation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition:</strong> This case represents either the switch between an abbreviation and the full name of a location (e.g. “S.F State” → “San Francisco State”), re-ordering the position of the geo-competent in the query (e.g. “used, utility trailer, florida” → “florida, used, utility trailer,”) or changes between upper case and lower case letters (e.g. “EASTON WA” → “Easton WA”)</td>
</tr>
</tbody>
</table>

| Level 2: Modification to the format of the non-geo element of the query (“Non-geo format Modification”). |
**Level 3:** Non-geo error correction

E.g. (“central Florida bed and brefasts” → “central Florida bed and breakfasts”)

**Level 3:** Non-geo term variation

E.g. (e.g. “Hartley iowa obit” → “Hartley iowa obituaries” and “camera stores somerset county nj” → “camera shops somerset county nj”)

**Level 2:** User of search operators (“Operator usage Modification”).

**Definition:** This case refers to the situation when the searcher reformulates the query by using a search operator (e.g. “aqua restaurant San Francisco” → “aqua restaurant AND San Francisco”).

**Level 1:** Modification to the query resource (“Resource Level”)

**Definition:** Refers to the situation in which users specify the type of document they are interested in (e.g. map or picture) or the domain of their query (e.g. tx.us and .gov)

**Level 2:** Type of resource. E.g. (“Salt Lake City” → “Salt Lake City + Map”).

**Level 2:** Domain suffix. E.g. (“melia cozumel” → “Melia Cozumel.com”). A resort in Cozumel, Mexico.

Table 11 (The classification used in analysing the data)
Chapter 3: Research Methodology

It should be noted that some of the queries might have two reformulation actions since the user, for example, may generalise the geographic location and specialise the non-geo. Therefore, we assigned a weight for each action depending on the total number of changes that the user had made to the query according to the following formula:

\[ \text{Weight} = 1 / \text{the total number of changes performed on the query} \]

The construction of the transition matrix:

When designing the transition matrix, we took into consideration the current state of the query, that is the reformulation action that had been performed on the query, and the next state the user had moved to. We did not consider the previous state. This approach was followed in Pass et al. (2006) and Spink et al. (2000). The following example illustrates this idea:

<table>
<thead>
<tr>
<th>No.</th>
<th>Query</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gas Motor Bicycles</td>
<td>Initial query</td>
</tr>
<tr>
<td>2</td>
<td>Gas Bicycles</td>
<td>Non-geo Generalization</td>
</tr>
<tr>
<td>3</td>
<td>Gas Bicycles Los Angeles</td>
<td>Geo Specialisation</td>
</tr>
</tbody>
</table>

Transition states of these queries:

1. Initial query → Non-geo Generalization.
2. Non-geo Generalization → Geo Specialisation.

Objective 6: To explore geo-reformulation trends

Riech & Xie (2006) argue that a major limitation of research on query reformulation is the consideration of the changes in successive queries only without inferring the whole process. They propose eight different trends in user reformulation, as shown in chapter 2. However, their trends are rather broad and the interest in this study is to focus on geographic patterns; hence, we qualitatively analysed the geo sessions looking for particular trends that the session, as a whole, is representing which could be seen as an application of open coding in the grounded theory approach, which is defined as “the analytic process through which concepts are identified and their properties and dimensions are discovered in data” (Strauss & Corbin, 1998: 101).
4.1 Introduction

This chapter details the results of this research along with their implications and position with the existing literature. Overall, it is structured as follows: Firstly, the results of analysing the dataset as a whole are reported in order to examine the length of searching sessions; secondly, the findings of the sampled data analysis are presented for the purpose of investigating the following general aspects: the size of geographic reformulation; the mechanism by which users reformulate their queries either in terms of changes in successive queries length or by applying different modification actions to their needs; the complexity and common structures of geographic modifications; and, finally, trends in geo-rewritings.

4.2 Analysis of sessions length

Despite the limitations of the figures presented in the literature, the general consensus is on the shortness of searching sessions. Our results concur with this since we found that 62.07% of such use cases contained only one query, whereas 17.82% of them had two queries. The remaining proportion was distributed in a “long tail” shape towards a higher number of queries. Figure 1 illustrates these statistical results diagrammatically.

![Figure 1: Sessions length distribution](image-url)
Interestingly, the mean session length before and after filtering the data from agent queries was nearly the same, with about 1.99 and 1.98, respectively. This quite small number is comparable with Silverstein et al.’s (1998) mean length of 2.02, although they used a different session identification method and did not clean non-human queries. However, it should be noted that our figures might be at the maximum level given that Zhang & Moffat (2006), who studied the same query log, reported 1.55 as the average sessions’ length, probably using an over-filtering approach to clean the queries. Abstractly, these results convey the tendency to have short web searching sessions and trigger two questions, at least, which need to be clarified.

Firstly, it might be the case that search engines are successful in returning a precise result (high precision) that satisfies users’ needs. To examine this, the clicks data of those sessions were investigated. 69.57% of those who submitted one query did view a result (3,226,007 out of 4,637,061 queries), which might partially support the former claim. However, perhaps the remaining 30.43% of users were satisfied with the snippets provided or they moved to use another searching facility or perhaps they were not earnest in their needs. Secondly, since the majority only submitted a single query, and assuming that the search engines were performing well, it is of interest to see why the others have reformulated their queries. Jansen et al. (2000: 213) raise this issue and state that “they could represent sub-populations of web users with more experience or higher motivation”. We add to their interpretation that it is important to study query reformulation in terms of the topical domains, such as in our study (geographic scope). This would inform us of the difficulties in locating particular types of needs, leading to improvement in handling them or it would provide interpretations of the use of modifications.

Considering the behavioural angle of this result and compared with the early studies published in this field, for example Silverstein et al. (1998) and Jansen et al. (2000), it does appear that searchers’ behaviour is static despite the fact that search engines have improved significantly during this period. Spink et al. (2001: 227) support this claim by saying that “The Internet changes fast. In contrast, people, their information needs, and behavior do not”.
4.3.1 The size of geographic reformulation on the web

The analysis of the 2400 reformulated queries revealed that 471 (19.625%) of them were geographic queries. These queries represented 471 different sessions. That is, nearly one in five users who reformulated their needs where looking for geographically related information. Given the diversity of topics searched for on the web and that our data were extracted from a general purpose search engine, such a figure is rather high, indicating the huge amount of geographic interest on web searching, as has been shown in Sanderson & Kohler (2004) and, meanwhile, questioning the reasons for having this percentage of modification on them. Whether geographically related information is difficult to locate on the web or there are other reasons to have this amount of modification on them is explored in the following sections of this report.

4.3.2 Analysis of Geo-sessions length

Our aim at this stage was to examine whether there was a difference between the length of geo-sessions and the non-geo sessions. The mean length of geo-sessions was 7.55(1), which is very close to the non-geo sessions figure (7.49 queries per session). In order to test this statistically, we applied a two-tailed T-test to the relevant records of these two groups assuming that they had unequal variance. The result showed that there was no statistical significant difference between the two groups ($p=0.89$). This illustrates that geo-sessions do not differ from the general trend in terms of their length despite the difference in their queries length, as have been shown in Sanderson & Kohler (2004) and Jones et al. (2006). However, it is important to say that these figures were calculated in terms of the reformulated sessions that had at least 3 queries, meaning that these numbers do not represent geo and non-geo searching episodes in general but rather give an indication that users who modify their geographically related information do not submit more or fewer queries than the average users who do the same.

---

1- We have included the navigational and repeated queries in this figure.
4.3.3 Geo-modification:

A view on changes in successive queries’ length

The 471 geo-sessions were analysed to shed light on the difference between a modified geographic query and its former in terms of the query length. Table 12 shows the result obtained.

<table>
<thead>
<tr>
<th>Number of Words</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Modification by adding words</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>427</td>
<td>21.49 %</td>
</tr>
<tr>
<td>2</td>
<td>231</td>
<td>11.63 %</td>
</tr>
<tr>
<td>3</td>
<td>76</td>
<td>3.82 %</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>1.21 %</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0.45 %</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>0.25 %</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0.15 %</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0.1 %</td>
</tr>
<tr>
<td><strong>Total queries</strong></td>
<td><strong>777</strong></td>
<td><strong>39.10 %</strong></td>
</tr>
</tbody>
</table>

| **Modification by deleting words** | | |
| -1                             | 316       | 15.9 %     |
| -2                             | 190       | 9.56 %     |
| -3                             | 67        | 3.37 %     |
| -4                             | 28        | 1.41 %     |
| -5                             | 14        | 0.7 %      |
| -6                             | 4         | 0.2 %      |
| -7                             | 1         | 0.05 %     |
| -8                             | 2         | 0.1 %      |
| **Total queries** | **622** | **31.29 %** |

| **Modification by substitution** | | |
| 0                             | 588       | 29.59 %    |
| **Overall** | **1987** | **99.98 %** |

Table 12: Geo-modification – changes in successive queries length-

From the above, it seems that the scale tilts fairly in favour of adding more terms to the geo-query rather than subtracting from it, with a decrement trend on both sides, addition and deletion, towards a higher number of terms. However, a closer look at the previous table reveals that words substitution is the most frequent single reformulation pattern, with a proportion of 588 occurrences among 1987 reformulated queries. These figures are consistent with the generic ones published in Spink et al. (2000), although they reveal that 34.67% of modifications are words replacement.

---

1- The total number of the queries is 2458. This figure does not include the initial queries.
2- Does not count to 100% due to rounding errors.
Chapter 4: Results & Discussion

From a developer’s point of view, such figures might raise the need for modifying the way search engines propose reformulated queries to balance and pay attention to words substitution approaches rather than focusing on adding extra words. This claim could support the work of Jones et al. (2006), who proposed a similar method. However, from a geographical perspective, care should be taken in replacing the geo-part of the query since previous research has shown that automatically reformulated geo-queries are deemed to be irrelevant when the geo-part is changed (Jones et al., 2006).

4.3.4 Geo-Modification: A depiction of users actions

For the sake of clarity, the results of this section are presented in two parts. Firstly, we would provide an overview of the modifications made to the geo-queries, followed by a detailed analysis of the actions performed by the searchers.

4.3.4.1 Overview analysis

The goal in this phase was to examine geo-reformulation in terms of Rieh & Xie’s (2006) three main facets: content, format and resource. The results show that the vast majority of geo-rewritings affect the content of the query more than the other two factors. Specifically, 83.2% of reformulation is made on the content compared with 12.7% and 1.53% for the format and resource, respectively. These figures illustrate a slight difference between geo modifications and the generic findings published in Riech & Xie (2006). A comparison between them is presented in table 13.

<table>
<thead>
<tr>
<th>Facet of modification</th>
<th>This study</th>
<th>Riech &amp; Xie (2006)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Percentage</td>
</tr>
<tr>
<td>Content</td>
<td>1653.17</td>
<td>83.2 %</td>
</tr>
<tr>
<td>Format</td>
<td>252.33</td>
<td>12.7 %</td>
</tr>
<tr>
<td>Resource</td>
<td>30.5</td>
<td>1.53 %</td>
</tr>
<tr>
<td>Unknown(^1)</td>
<td>51</td>
<td>2.57 %</td>
</tr>
<tr>
<td>Total</td>
<td>1987</td>
<td>100 %</td>
</tr>
</tbody>
</table>

Table 13: Geo-modification—an overview analysis-

\(^1\) We calculate the frequency depending on the weight of each modification action in order to fairly distribute the figure between the three facets providing that each query may have two or three modifications.

\(^2\) This group represents the malformed queries.
The difference can be seen in the relatively high percentage of content modification of the geographic queries compared with the average in Rieh & Xie (2006), whereas format and resource changes were slightly lower. However, it should be noted that it is expected to have a lower amount of resource reformulations since Live search engines provide other tools to search for different kind of resources (e.g. images and maps) and the data we have analysed was for web documents. Nevertheless, a question arising from the above figures is whether a content modification would lead the user to click on a result more than a format or resource action.

Under half of those who reformulated their geo-query content viewed at least one result (46.87%), while 36.72% of format modification resulted in a click. The highest percentage was for resource modifications (57.38%); this was expected to be higher than the others since nearly all of its rewritings are for maps and images of popular places, which it is thought that the search engine can handle effectively. However, in terms of the content and format, these figures do not seem to be encouraging and give an indication that most users’ modifications may not be successful if we measure them by the clicks data, which is not always the case but it might be the only measurement of successful queries in the data provided in standard query logs. In general, users’ inability to successfully formulate their queries has long been discussed in the IR literature (Gauch & Smith, 1991) and has provided this result; it does appear that such a problem still exists, at least from a geographical perspective.

Overall, the aforementioned results might reveal the uncertainty that searchers have when looking for online geographic information. In other words, since they change the meaning of the query rather than its format, it could be an indicator of difficulty in formulating their geographic needs. This claim could be of value if we consider the two parts of the query: the geo and the non-geo components, in which users are abstractly handling two different entities. On the other hand, it might be caused by their exposure to the results or snippets obtained initially and thus initiating a cycle of formulating more accurate queries that best represent their needs, or perhaps it is a process of learning more about a particular problem. When we tested whether those 471 users had clicked on a result after submitting their initial query, we found that 46.70% of them did. Clearly, this percentage might not be enough to support one of the former claims rather than
Chapter 4: Results & Discussion

another, but it could be unsatisfactory enough to say that geo-queries might not be initially formulated in a proper manner, leading users to modify them.

4.3.4.2 Detailed Analysis

A thorough investigation into the 471 geographic sessions was conducted using the modification scheme explained in chapter 3. The results were as follows:

- Searchers were found to modify the geo content of their queries almost equally as they do with the non-geo component. Specifically, 34.03% of content modification was on the geo-part of the query compared with 35.47% on the non-geo. In contrast, they seemed to change the format of the non-geo part (57.33% of format reformulation) more than the geographic one (30.84%). However, based on the fact that the vast majority of modifications were on the content of the query, these results could partially reject the claim that “when people search for local information or services, they are often fairly confident about the appropriate geo terms. Thus, when users modify their queries, they are more often modify the non-geo terms” (Gan et al., 2008: 56).

- Among geo-content modifications, nearly half of the users were specialising their geographic references compared to 25.36% for generalisation and 24.83% for changing the search location.

- The percentage of changing the search location in terms of the overall figures was 7.03%. The method we have used to identify this action is to consider a change between two geographic entities that had a spatial relationship, that is one is contained in the other, to be either specialisation or generalisation. If we include these cases as instances of changing the geo-entity, the figure would be 8.55%.

- An interesting finding is that nearly one in four queries modifications represents an interruption of either starting a new search or returning to a previously submitted query. In fact, the highest percentage of the interruptions is re-visits to previous queries, with a percentage of 34.6% in this category. When we investigated the clicks stream of those who
Chapter 4: Results & Discussion

returned, we found that only 21.2% of them had clicked on a result. Such a percentage leads to questions over the reasons for this behaviour.

- A negligible number of modifications on the geo-part were for correcting spelling errors, indicating that users mostly type the correctly spelled location. Meanwhile, the results did not show a noticeable figure in regard to changes made on the lexical structure of the geo-part (e.g. use of abbreviations instead of full names or vice versa).

Table 14 summarises the findings of this stage.
## Chapter 4: Results & Discussion

<table>
<thead>
<tr>
<th>Facet</th>
<th>Sub-category</th>
<th>Action</th>
<th>Frequency</th>
<th>Percentage within the sub category</th>
<th>percentage within the facet</th>
<th>Overall percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td>Geo Content Modification</td>
<td>Change the Geo entity</td>
<td>139.67</td>
<td>24.83%</td>
<td>8.45%</td>
<td>7.03%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Geo Generalisation</td>
<td>142.67</td>
<td>25.36%</td>
<td>8.63%</td>
<td>7.18%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Geo Specialisation</td>
<td>280.17</td>
<td>49.81%</td>
<td>16.95%</td>
<td>14.10%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>562.51</td>
<td>100.00%</td>
<td>34.03%</td>
<td>28.31%</td>
</tr>
<tr>
<td><strong>Non-geo Content Modification</strong></td>
<td></td>
<td>Change the non-geo entity</td>
<td>187</td>
<td>31.89%</td>
<td>11.31%</td>
<td>9.41%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-geo Generalisation</td>
<td>160.83</td>
<td>27.43%</td>
<td>9.73%</td>
<td>8.09%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-geo Specialisation</td>
<td>238.5</td>
<td>40.68%</td>
<td>14.43%</td>
<td>12%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>586.33</td>
<td>100.00%</td>
<td>35.47%</td>
<td>30%</td>
</tr>
<tr>
<td><strong>Interruption</strong></td>
<td></td>
<td>Return query</td>
<td>174.5</td>
<td>34.60%</td>
<td>10.56%</td>
<td>8.78%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Start new search with other geo part</td>
<td>113.5</td>
<td>22.50%</td>
<td>6.87%</td>
<td>5.71%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Start new search with the same geo part</td>
<td>92.83</td>
<td>18.41%</td>
<td>5.62%</td>
<td>4.67%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Start new search without geo-part</td>
<td>123.5</td>
<td>24.49%</td>
<td>7.47%</td>
<td>6.22%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>504.33</td>
<td>100.00%</td>
<td>30.52%</td>
<td>25.38%</td>
</tr>
</tbody>
</table>
Chapter 4: Results & Discussion

- The total number of queries is 1987. However it does not count to this number due to rounding errors in the frequency.

<table>
<thead>
<tr>
<th>Format</th>
<th>Geo format Modification</th>
<th>Geo error correction</th>
<th>Geo term variation</th>
<th>Total</th>
<th>100.00%</th>
<th>30.84%</th>
<th>3.92%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>20.56%</td>
<td>6.34%</td>
<td>0.81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>61.83</td>
<td>79.44%</td>
<td>24.50%</td>
<td>3.11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Total</strong></td>
<td><strong>77.83</strong></td>
<td><strong>100.00%</strong></td>
<td><strong>30.84%</strong></td>
<td><strong>3.92%</strong></td>
<td></td>
</tr>
<tr>
<td>Non-geo format Modification</td>
<td>Non-geo error correction</td>
<td>29.83</td>
<td>20.62%</td>
<td>11.82%</td>
<td>1.50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-geo term variation</td>
<td>114.83</td>
<td>79.38%</td>
<td>45.51%</td>
<td>5.78%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>144.66</strong></td>
<td><strong>100.00%</strong></td>
<td><strong>57.33%</strong></td>
<td><strong>7.28%</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operator usage</td>
<td>29.83</td>
<td>100.00%</td>
<td>11.82%</td>
<td>1.50%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resource</td>
<td>Domain Suffix</td>
<td>9</td>
<td>29.51%</td>
<td>29.51%</td>
<td>0.45%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type of Resource</td>
<td>21.5</td>
<td>70.49%</td>
<td>70.49%</td>
<td>1.08%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>51</td>
<td>100%</td>
<td>100%</td>
<td>2.57%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Total number of queries</strong></td>
<td>1986.99</td>
<td></td>
<td><strong>Total percentage</strong></td>
<td><strong>99.99%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 14: Geo-modification—detailed analysis-
The previous table reveals that users have a tendency to specify their queries based on the fact that about one in four modifications actions are for specialisation, with the geo part having 14.10% of them. Such behaviour could be attributed to the user attempting to disambiguate one place from another in order to solve the problem known as “referent ambiguity”. This claim is supported in Jones et al. (2008). On the other hand, it might be a case of narrowing the search geographic location to specific geo-entity that the searcher is interested in, for example searching about Los Angeles and then modifying it to Hollywood, which clearly shows the specific geo-intent within Los Angeles city.

Another prominent fact was the relatively unexpected figure of return queries with the lower percentage of clicks. Although the literature reveals that users often repeat their queries (Bruza & Dennis, 1997), we included only the non-sequential repeated queries. In other words, it seems that users, after a number of modifications on their geographic query, decide to go back to a previously submitted query. This action might be a result of a cognitive process in which the user realises that a previous query is most relevant to his/her information need or perhaps it could be an indicator of failure in their successive reformulated requests. Another possible interpretation, given the percentage of clicks they made, is that they may have seen an interesting snippet to re-examine. Regardless of what made them to perform such an action, this result certainly supports the claim that appropriate catching would enhance the search process (Bruza & Dennis, 1997).

### 4.3.5 Transition Analysis

It is believed that exploring the connection between two successive queries reformulation actions is important either for drawing a complete picture of users’ modifications behaviour or for developing reformulation tools that can exploit this knowledge in predicting the user’s next move. To do this, we have adopted the notation of transition matrix. We have included in this analysis only the transitions between the content and format facets. The resource was eliminated because of its negligible role in geo-modification. This phase led to the following observations:
Chapter 4: Results & Discussion

- A significant percentage of modifications on the initial query were made on either its geo or non-geo content (63.8%), with the geo-part having the highest proportion (34.61%). Through deeper investigation into this, it was found that specialisation actions were responsible for over half of these rewritings in either the geo or the non-geo component. This might support the assumption that users start their search with a general query, especially in terms of the geo-component, and then try to limit it. This observation is supported in Jansen et al. (2009) in regard to general queries.

- 73.62% of users who generalised their queries were likely to perform a specialisation on their successive query. In specific, when they generalised their geo-part, they often specialised it afterwards (24.09%), and the same behaviour was applicable to the non-geo component. Conceptually, this finding shows that users occasionally treated the geographic query as having two separate components, given the fact that their next move was on the same part that they were modifying. Such a result might be exploited in designing automatic modification tools to tailor the suggested queries according to the query component that the searcher has just modified.

- In contrast to the inverse relationship between generalisation and specialisation explained above, users were likely to change the part of the query that they had just altered. In other words, 45.32% of those who had changed the geo-element of the query had already done the same in the preceding request. This finding is thought to be of high value, especially in regard to the geo-component since it is needed to anticipate when the user is likely to seek a change so that an automatic reformulation tool can suggest alternative places. By drawing on Jones et al.’s (2006) previously mentioned claim about the risk of changing the query location automatically, we could argue that this finding might mitigate such a risk since we know when such an action is likely to be relevant.

- Nearly one in three users who interrupted the sequence of their queries by searching for a different information need without specifying the location of interest did follow this by determining the location that they wanted to search in. Most of
these specialisation cases were found to be about the geographic place that the user was searching after before the interruption. Thus, it might be optimal to treat any interruption of this kind as having the geographic scope represented in the previous query, even if the user did not explicitly ask for it, so that the number of modifications that he/she is likely to perform can be reduced, especially when the timeframe between the two queries is limited.

- It is evident that when users traced back a query from their search sequence, they were likely to continue in doing this as 39.51% of “return query” modification actions were followed by the same behaviour. This interesting figure could be attributed to at least two interpretations. It might be the case that searchers use the back button on their browser to move to a particular point in their search and from there to another preceding one or they maybe use their browsing history to navigate through their search footprints.

The application of transition a matrix on query reformulation was thought to be an efficacious approach in discovering hidden relationships between the sequences of rewrites although in some cases it is difficult to anticipate users’ actions given the diversity of ways they can modify their search with. However, it is thought that aggregating such knowledge with other resources such as mining the logs to discover possible rewrites can improve the way geographically related queries are handled by the search engine since we can anticipate to some extent what the user is likely to do next. Table 15 illustrates the findings of this stage.
## Table 15: Transition matrix between the modification actions

<table>
<thead>
<tr>
<th>From</th>
<th>Generalisation</th>
<th>Specialisation</th>
<th>Change</th>
<th>New search (Same place)</th>
<th>New search (Other place)</th>
<th>New search (Without place)</th>
<th>Return query</th>
<th>Error correction</th>
<th>Term variation</th>
<th>Operator usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Geo %</td>
<td>Non-geo %</td>
<td>Geo %</td>
<td>Non-geo %</td>
<td>Geo %</td>
<td>Non-geo %</td>
<td>Geo %</td>
<td>Non-geo %</td>
<td>Geo %</td>
<td>Non-geo %</td>
</tr>
<tr>
<td>Initial query</td>
<td>8.02</td>
<td>6.37</td>
<td>19.12</td>
<td>14.91</td>
<td>7.47</td>
<td>7.91</td>
<td>3.52</td>
<td>8.79</td>
<td>5.71</td>
<td>0</td>
</tr>
<tr>
<td>Generalisation</td>
<td>5.81</td>
<td>11.94</td>
<td>24.09</td>
<td>8.39</td>
<td>2.9</td>
<td>12.9</td>
<td>3.87</td>
<td>3.23</td>
<td>8.39</td>
<td>5.16</td>
</tr>
<tr>
<td>Non-geo</td>
<td>7.81</td>
<td>8.44</td>
<td>13.71</td>
<td>27.43</td>
<td>6.96</td>
<td>2.74</td>
<td>1.9</td>
<td>3.16</td>
<td>4.43</td>
<td>8.23</td>
</tr>
<tr>
<td>Specialisation</td>
<td>10.7</td>
<td>12.06</td>
<td>9.51</td>
<td>10.76</td>
<td>5.22</td>
<td>7.65</td>
<td>5.6</td>
<td>4.85</td>
<td>6.53</td>
<td>11.19</td>
</tr>
<tr>
<td>Non-geo</td>
<td>7.65</td>
<td>16.45</td>
<td>11.66</td>
<td>6.37</td>
<td>2.79</td>
<td>12.88</td>
<td>3.58</td>
<td>5.58</td>
<td>6.44</td>
<td>7.3</td>
</tr>
<tr>
<td>Change</td>
<td>7.02</td>
<td>0.83</td>
<td>7.44</td>
<td>11.85</td>
<td>45.32</td>
<td>2.89</td>
<td>2.48</td>
<td>5.79</td>
<td>3.31</td>
<td>4.13</td>
</tr>
<tr>
<td>New search (Same place)</td>
<td>3.83</td>
<td>11.48</td>
<td>4.92</td>
<td>16.12</td>
<td>2.46</td>
<td>17.76</td>
<td>18.03</td>
<td>3.28</td>
<td>11.48</td>
<td>4.92</td>
</tr>
<tr>
<td>New search (Other place)</td>
<td>9.77</td>
<td>5.17</td>
<td>5.17</td>
<td>14.94</td>
<td>6.7</td>
<td>7.28</td>
<td>4.6</td>
<td>14.94</td>
<td>9.2</td>
<td>3.45</td>
</tr>
<tr>
<td>New search (Without place)</td>
<td>0</td>
<td>2.02</td>
<td>32.32</td>
<td>9.09</td>
<td>0</td>
<td>4.55</td>
<td>12.12</td>
<td>6.06</td>
<td>17.17</td>
<td>5.05</td>
</tr>
<tr>
<td>Return query</td>
<td>3.85</td>
<td>6.64</td>
<td>11.89</td>
<td>7.69</td>
<td>1.05</td>
<td>7.34</td>
<td>5.24</td>
<td>3.85</td>
<td>4.9</td>
<td>39.51</td>
</tr>
<tr>
<td>Error correction</td>
<td>3.57</td>
<td>3.57</td>
<td>10.71</td>
<td>17.86</td>
<td>35.71</td>
<td>0</td>
<td>0</td>
<td>7.14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Term variation</td>
<td>8.7</td>
<td>7.97</td>
<td>15.22</td>
<td>4.35</td>
<td>2.17</td>
<td>2.17</td>
<td>8.7</td>
<td>8.7</td>
<td>4.35</td>
<td>4.35</td>
</tr>
<tr>
<td></td>
<td>6.45</td>
<td>6.45</td>
<td>10.75</td>
<td>5.91</td>
<td>3.76</td>
<td>10.75</td>
<td>4.3</td>
<td>7.53</td>
<td>5.38</td>
<td>16.13</td>
</tr>
<tr>
<td></td>
<td>4.69</td>
<td>15.63</td>
<td>10.94</td>
<td>14.06</td>
<td>0</td>
<td>9.38</td>
<td>3.13</td>
<td>6.25</td>
<td>6.25</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table description:**

- Rows represent the state from which the user is moving (current state) and columns represent “the next expected state”.
- Values in bold means the highest probability.
- Resource actions were eliminated from the table given their insignificant role in geo-modification and thus it shall be noted that the values for each row do not amount to 100% but all of them weight for over 94%.
4.3.6 Analysis of correlated modifications actions

The aim of this analysis was to investigate the complexity of modifications actions that users simultaneously performed on their geographic query. The results showed that only 17.11% of reformulated queries were having two or three parallel modifications. When we investigated this behaviour, the mostly occurring pairs of actions were unexpectedly found to have an inverse relationship. That is, when users generalised the geo-part of the query, they tended to specialise the non-geo component of it if they would perform a parallel move and vice versa. For instance, one of the users made the query “century furniture” and then modified it to “North Carolina furniture manufacturers”. In the previous case the user generalised his/her non-geo part by moving from a single defined manufacturer to a rather broad category but with specifying the location of interest. However, given the fact that the vast majority of queries were modified using a single action at each time, a reasonable conclusion to be drawn from this is that searchers usually rewrite their geographic information needs in a simple way by performing a single modification action on one of the query components.

4.3.7 Time based analysis

It was found that the likelihood of modifying a geographic query is at its peak in the first minute after its submission. This illustrates that most of the users reformulated their requests within seconds although there were occurrences for noticeable amount of modifications after the first minute. Such a finding might indicate that users often modify their queries because they are not satisfied with the result obtained since it is argued by Lau & Horvitz (1999) that longer waiting times might mean that users have found relevant documents or at least other paths to fulfil their needs. Zhang & Moffat (2006) came to the same finding in regard to the interval between two successive general queries.

A closer investigation into this reveals that a significant amount of the geo-rewritings actions that do not alter the information need happen in this time span. In other words, users do not seem to start a new search or change the geo or the non-geo components of their queries as often as the other actions in the initial 60 seconds, which is an expected behaviour, especially in regards to a new search given
that such an action might have a relatively heavy cognitive load on the user more than simpler actions, such as generalisations or specialisations, since it may take longer to formulate the new information need in the mind of the user than adding or deleting terms to the query. This finding is consistent with Lau & Horvitz’s (1999) figures which draw similar facts. Moreover, it was noted that the vast majority of the error corrections for both the geo and the non-geo part of the query were conducted instantly. Also, it was found that most of the return queries appear in the first seconds indicating that users may decide to return back to a previous query without examining the results of their current one.

### 4.3.8 Geo sessions reformulation pattern

This analysis aims to examine whether users follow a common pattern of modifications in their geo-queries. For example, if the user modified the initial query by specialising the location of interest and then changed it to another location, the search pattern would be “start \(\rightarrow\) geo content modification \(\rightarrow\) end of the search”, providing that the two modifications actions (i.e. geo specialisation and change geo-entity) were from the “geo content modification” category. The sequence of such actions was not considered because of the specific nature of these actions for which it is unlikely to have a common pattern. Figure 2 shows the 5 most occurring patterns.

![Geo sessions reformulation pattern](image)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Percentage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP1</td>
<td>7.86%</td>
<td>Start (\rightarrow) Geo content modification (\rightarrow) End of the search.</td>
</tr>
<tr>
<td>SP2</td>
<td>5.52%</td>
<td>Start (\rightarrow) Interruption (\rightarrow) End of the search.</td>
</tr>
<tr>
<td>SP3</td>
<td>4.25%</td>
<td>Start (\rightarrow) Non-geo content modification (\rightarrow) End of the search.</td>
</tr>
<tr>
<td>SP4</td>
<td>3.4%</td>
<td>Start (\rightarrow) Geo content modification (\rightarrow) Interruption (\rightarrow) End of the search.</td>
</tr>
<tr>
<td>SP5</td>
<td>2.97%</td>
<td>Start (\rightarrow) Geo content modification (\rightarrow) Non-geo content modification (\rightarrow) End.</td>
</tr>
<tr>
<td>Others</td>
<td>76%</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4: Results and Discussion

It was found that the majority of users modified their geo-sessions in quite diverse ways. In fact 225 sessions out of the 471 had distinctive behaviour. This might be attributed to the length of the session since as its size enlarges it is unlikely to have a shared reformulation path. However, a noticeable finding was that one of the common patterns occurred when users interrupted their search just after submitting the initial query by starting a new one. Some might argue that this indicates a failure in defining the session boundaries since a session ideally consists of a single information need. On the other hand, the inclusion of such behaviour might reveal interesting facts about users’ searching activities, especially when the timeframe between the two queries is short. In our study, we found that users occasionally started a new search either within the same location expressed in the initial query or within other locations, which indicates a need for studying the history of the session as a whole so that a deeper understanding of the reformulation characteristics can be reached. This supports Rieh & Xie’s (2006) previously mentioned limitation of query reformulation studies.

4.3.9 Trends in geographic query reformulation:

The geo-sessions were analysed for the purpose of identifying trends in geographic rewritings. This was achieved through an inspection of the query sequence within each search episode. As a result, four different patterns emerged which can, to some extent, provide a valid interpretation of the question “why do users reformulate their geographic query?”. These are presented in table 16. However, given the qualitative nature of this analysis, it shall be noted that the results obtained are essentially subjective and might not match the user’s actual intent, although we were able to draw a conclusion about most of the sessions.

<table>
<thead>
<tr>
<th>Trend</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal approach</td>
<td>318</td>
<td>67.52%</td>
</tr>
<tr>
<td>Multi-needs approach</td>
<td>62</td>
<td>13.16%</td>
</tr>
<tr>
<td>Multi-places approach</td>
<td>43</td>
<td>9.13%</td>
</tr>
<tr>
<td>Hybrid approach</td>
<td>35</td>
<td>7.43%</td>
</tr>
<tr>
<td>Unclassifiable1</td>
<td>13</td>
<td>2.76%</td>
</tr>
<tr>
<td>Total</td>
<td>471</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 16: Trends in geographic reformulation

1- These are sessions which contain either malformed queries or placeless queries.
Chapter 4: Results and Discussion

In the following a detailed discussion about these trends is presented:

- **Normal approach**: This pattern refers to those sessions in which the user is looking for a single information need in one location (i.e. classical session). As shown above, nearly two-thirds of the 471 users had difficulty in locating their single geographic need. For example, a user looking for life insurance company in Richmond, Virginia reformulated his/her query as follows:

  “life of Virginia insurance → life insurance company of Virginia → life insurance company of Virginia, Richmond → life of Virginia, Richmond → life insurance, Richmond Virginia”.

  The dominant modification actions in this category were found to be generalisations and specialisations for both the geo and the non-geo component, as in the example shown above. Overall, this finding does support our former claim that in most of the cases geo-queries were not formulated properly leading users to rewrite them.

- **Multi-needs approach**: It refers to those cases in which the user searches for different information needs but within the same location. Users following this approach might be interested either in exploring a place for tourism purposes or to acquire general knowledge about a specific geographic entity or perhaps they simply limit their daily information needs in a local sense. Regardless of what interpretation is valid, this category does show that not all of the users reformulate their geographic query either because of inaccurate formulation of the initial query or irrelevant retrieved documents by the search engine. Examples of this category are presented below:

  1- “ks travel → may events in ks → ks farms → ks farms to visit → ks bed and breakfast”. (1)

  2- “Corsica → map of Corsica → flag of Corsica → dress style of Corsica”.

  3- “Wells Fargo → Wells Fargo Dallas Locations → Forest Cove Apartments Dallas” (2)

---

1- KS is the abbreviation of Kansas state.
2- Wells Fargo is a bank in the US.
Chapter 4: Results and Discussion

- **Multi-places approach:** refers to those cases when the user searches for a single information need in multiple places. Such behaviour might be attributed to users comparing a service in two or more places or perhaps a willingness to travel to a nearby location in cases when they could not locate their need in the original one. The following examples illustrate this point:
  1. “Mortgage Lending → Mortgage Lending Herndon → Mortgage Lending Missoula”.
  2. “ford of Grapevine → ford of Fort Worth → ford of Richardson → ford of Plano → ford of Duncanville”

- **Hybrid approach:** refers to the situation in which the user mixes between two of the former approaches in one session. An example of this is presented below:
  1. “Pennsylvania personal license plates → penna personal license plates → penna personal plates → New York marathon → New Jersey marathon”.
  2. “Charlotte fire department → Rock Hill fire department → Rock Hill,sc fire department → US LEC → US LEC Charlotte, nc”.
4.4 Results limitations

It is possible that the dates on which the research data was gathered might have affected the significant of the result. That is, the query logs represented entries for July, month which possibly means an increase in the number of geographic queries since it is about the summer and travel and tourism activities could be at their peak. Thus it is thought that we might have over-estimated the proportion of geo-queries and a study of other months of the year may reach different conclusions. Meanwhile, the data was collected from the US site which means that the user group is this research is US residents but not web users in general. Having said that, it might not be a major limitation since geographic search is usually provided on the basis of a single country.

In addition to the previous limitations, query logs studies, in general, have long been criticised for their lack of recording users’ real intentions. Grimes et al. (2007) argue that such data does not provide sufficient information to answer questions as to why users are following a particular approach and any attempt to infer their behaviour is this limited. This might indicate the need for combining other methods of investigations, such as surveys, with the analysis of query logs in order to measure the qualitative aspects of web searching behaviour in a comprehensive manner.
5.1 Conclusion

Local queries are found to represent a significant share of query reformulation. One can relate this fact to the claim that “search engines do not analyze the geographical locality of queries and users, and hence often produce sub-optimal results” (Gravano et al., 2003: 325). If we consider this as the only reason for this behaviour, spatial ranking algorithms for geo-queries might be the optimal solution (Jones et al., 2008). However as evidenced in this study, this is not always the case. Users might reformulate their geo-queries either to look for the same information need in multiple locations or by switching between different needs in one place. This clearly indicates that a reformulation of this kind is not, necessarily, a result of ineffective retrieval methods but of special interests of those users. They might do so for purposes of travel arrangements, general knowledge, comparison between services and preparations for special events. On the other hand, this might be resulted from users’ inability to formulate geo-queries in an effective manner. Such a problem is not a new concept in IR and poses a significant challenge to search engines. It was evidenced that a great deal of reformulation actions performed on the initial queries were to specialise their content, indicating that users may tend to start their search with broad and less representative keywords on which search engines have limited control of but to propose alternative rewritings by developing novel reformulation tools.

In order to design such tools, it is necessary to understand the special characteristics of each type of queries, namely local and global. We have covered the local ones and found the importance of handling the two components, namely the location identifier and the information need, of such queries carefully. In fact, separating these two elements might be challenging in the first instance given that users do not follow a unified structure in their geo-queries, but nevertheless is crucial in order to best answer these kinds of requests or to suggest relevant rewritings. By considering these two independent elements, it is found that users mostly change their meanings by means of generalization, specialisation or replacing them with alternative options. These actions are found to result in an increase in the number of these queries terms although substituting and deleting terms are common. Most of these actions are performed within seconds from
submitting the original query illustrating the impatience of web searchers in examining further results or pages. In addition, when examining the sequence of these actions, it was found that users reformulate their queries in quite diverse patterns but it was possible to anticipate, to some extent, users’ next reformulation action building upon their current state.

Overall, this research provides the basic but essential knowledge that a geographic reformulation tool can act upon. It is apparent that an understanding of the query topic and the reformulation characteristics of it is vital in order to improve the efficiency and effectiveness of search engines. User have always and might will in the future submit short and quite vague queries which puts a heavy burden on the search engines sides in inferring their intent but nevertheless they leave a trace data behind in which analysing them is certainly a novel approach in exploring their behaviour.

5.2 Recommendations for Future work:

It was found that users occasionally reformulate their queries by changing the location of their search. It would be interesting to calculate the distance between their modified locations and the direction to which they reformulate their queries. This can be achieved through computing the distance between the two longitudes and latitudes and then identifying the bearing degree so that the direction can be identified. This is thought to be helpful in determining the boundaries in which users express a geographic flexibility so that the search engine can expand the geographic scope depending on it. An attempt to achieve this was made in this research but because of the limited size of the data sample it is believed that such results were not be representative.

- It would be of interest to conduct similar study on the query logs of local search tools. These queries are all, possibly, about geography and might yield further facts.

Word Count: 19,277


Bibliography


