Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

# The Search for the Searcher: Understanding User Behaviour through Search Log Analysis

Gemma O'Rourke
13321609

May 14, 2018

A Dissertation submitted in partial fulfilment
of the requirements for the degree of
Magister in Arte Ingeniaria at
Trinity College Dublin
The University of Dublin
with the supervision of
Prof. Seamus Lawless and Dr. Annalina Caputo

Submitted to the University of Dublin, Trinity College, May, 2018

# Declaration

I, Gemma O'Rourke, declare that the following dissertation, except where otherwise stated, is entirely my own work; that it has not previously been submitted as an exercise for a degree, either in Trinity College Dublin, or in any other University; and that the library may lend or copy it or any part thereof on request.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Signed: ____________________ Date: ____________________

# Summary

The objective of this research is to investigate the extent to which search engine logs can provide insight into the domain expertise of the user. A dataset of search logs from the AOL search engine in 2006 will be used to carry out the research. The purpose of this research is not to demonstrate how accurate search log analysis is at defining a user's knowledge level; users are not being surveyed and prior domain knowledge is not known. Rather, the goal is to provide sufficient proof that search logs are sufficient to distinguish users based on some measure of 'knowledge'. This knowledge measure is defined based on what data is available from the search logs, namely the query, the time the query was issued and the URL that was clicked as a result of the query (if any).

Past research has shown that search logs can provide useful distinctions between expert and novice users when the prior domain expertise level is known. Numerous research has focussed on analysing queries, dividing them into informational, navigational and transactional types, investigating reformulation strategies and other basic statistical analysis. However, there are few works which investigate the specific information need or topic associated with a query. This research intends to add to this research gap by investigating possible correlations between query analysis and topic analysis.

To prepare the data for analysis, the queries are categorised into informational and navigational types and the navigational queries are discarded. The set of queries for each user is then split into search sessions, and 100 users with more than 20 sessions are randomly selected for analysis. From the random sample, the topics of each search query and clicked URL are found by querying an online web directory. A complexity score is calculated for each query by finding the average specificity value of all the terms in the query and the maximum reading difficulty of the query. The purpose of this score is to ascertain if complexity tends to increase within a session. A reformulation check is then performed on the queries, comparing a query to the immediately previous query within a session. Each query can only have one reformulation strategy associated with it, multi-reformulations are not performed as part of this research. This is to investigate any correlation between increasing query complexity score and specific reformulation strategies. Similarly, the topic overlap between pairs of queries in a session is calculated, and the same is done for any clicked URLs in that session. The intention is to demonstrate if topics converge to a specific topic in a session, which would indicate that the user is targeting their queries to a particular information need. It is assumed that a single search session is associated with a single information need.

The results show that there is a tendency to increase query complexity within a search session, which means that users increase the specificity of their search terms and use more difficult terms. There is a negative correlation between the rate of increase of complexity and session length. Longer sessions indicate that the user is taking longer to satisfy their information need, which may be a direct result of a slower rise in query complexity but the direction of causality is unclear. There was no correlation observed between query complexity and the number of URL clicks associated with that query. If a positive correlation existed, it could have been concluded that more complex queries yield more relevant results, however this was not the case. Instead, the data appears to suggest that the rate at which the complexity is increased is more important for finding relevant results, assuming that the end of a session means that the user has satisfied their information need.

Query topics also demonstrated a tendency to increase within a search session. A very slight positive correlation was observed between query complexity and topic depth, which suggests that users find more specific, relevant results when they use more specific query terms. However, the correlation is not high enough to definitively prove this. The topic depth of clicked URLs also tended to increase within search sessions, however no correlation was observed between query topic depth and the clicked URL depth. Similarly, topic overlap was more likely to increase than decrease within a search session.

These results demonstrate that search logs can provide interesting insights into user behaviour, in particular that users tend to increase the specificity of their queries towards a specific information need/topic in a session. It can be inferred that a user searching within a shallow sub-topic will have lower complexity queries than a user searching within a deeper sub-topic. Assuming that a deeper sub-topic means a deeper understanding of the high-level topic, it can be concluded that higher values of query complexity indicate higher levels of expertise within a topic. Some limitations discovered in this research highlight scope for future work, in particular the development of a more sophisticated method of topic analysis.

# Acknowledgements

I would like to thank my project supervisors, Seamus Lawless and Annalina Caputo for lending me their support and expertise throughout the course of the project and for being generally patient, helpful and positive.

I also thank my mother, father, sister and boyfriend for their constant support, for keeping me sane and for always believing in me.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

When making decisions about how to best support users in web search, it is extremely helpful to know who they are, why they are searching, and how they interact with the search site. Is the user a novice investigator gradually learning about a topic? Do they refine the precision of their question based on new knowledge? Or are they an expert scholar, rapidly issuing specific, targeted queries as they gather sources for their research? How does the relevance change as the user expertise and interactions change?

The analysis and study of user behaviour on the web has been an on-going area of interest since the dawn of the World Wide Web. Even before this time, investigations were undertaken into how users queried a database, when query reformulation consisted of manually looking up words in a thesaurus. Today, when it comes to issuing queries, a lot of the work is done by the search engine. As a user types a query, suggestions are given to them on what word might come next. After the query is issued, the search engine processes it, and expands the query as it sees fit in the background, without the user's knowledge. When results are returned to the user, they may be ranked according to the search history of the user, with their preferred pages ranking higher.

Users are changing the way they interact with the system in response to these new methods. For example, queries are increasingly issued as linguistic phrases ("what is Obama's first name?") rather than keywords ("Obama first name"). This is mainly due to the rise of voice-to-search technology e.g. Apple's Siri and Amazon's Alexa, where the search engine translates our spoken request to a search query. This means that search queries no longer consist only of keywords, but include punctuation and stop words. As more state-of-the-art technologies and ideas are integrated into search engine algorithms, further changes in user behaviour will surely be observed. These changes are a catalyst for research into what is undoubtedly an exciting, dynamic and highly relevant topic.

There have been many different research projects which have examined how users search, what content they look for, and how the interface and style of webpages affect their search preferences. Much of this analysis can be done by investigating search engine logs which contain a snapshot of the user's behaviour and interest in a topic. Search logs are essentially a record of every query ever searched by a user. They can include data like the time the query was issued, the URL of the clicked result document, and that document's rank in the result list. Typically, the logs would store information about the user, such as their IP address, and login name and email(if the user has an account with the search engine). If search logs are made public, this private data is usually replaced with an anonymous ID.

Eickhoff et al.(1) investigated within-session learning using search logs from a popular search engine. The focus was on procedural and declarative knowledge queries, with metrics like domain count, query complexity and display time defined as synonymous with domain expertise. They were able to predict knowledge acquisition potential of web pages for a particular user. Gadiraju et al.(2) recruited a set of users for a more in-depth analysis of knowledge gain, focusing on informational search only. Users were given a specific information need to search for, and their knowledge of the topic before and after searching was measured. This study was able to show that the average complexity of the queries was positively correlated to the users' knowledge gain. White et al.(3) took an alternative approach to this problem by defining users as expert/non-expert in four broad topics, based on the sites they visited on that topic. Users that visited specialist sites were assumed to be experts. The behaviour of experts vs. non-experts was analysed, with respect to particular features of in-domain search sessions (similar to (1)). Expert's queries were longer and used more specific vocabulary than non-experts and they generally had longer and more diverse search sessions than non-experts. These studies form the foundation for this research project, and some of the evidence discovered in them will be used to formulate the research questions.

## 1.1 Objective & Research Questions

The objective of this project is to investigate whether search logs can provide sufficient data to demonstrate that a user is attempting to increase their domain knowledge; more specifically, to what extent can the analysis of search logs provide insight into how an individual user develops their knowledge of a topic within a search session? A discussion of knowledge and its definition, in the context of this work, can be found in 2.1. It is important to note that 'knowledge' is not defined in the traditional sense for this research, nor is it the intention of this research to evaluate the accuracy of search log analysis at defining a knowledge level. Users are not being assessed or surveyed, so their prior knowledge level is unknown. The data from the search logs will simply be used to get an indication of the user's expertise and show how it affects their search behaviour. The focus will be on the complexity of the query, with the assumption that the more experienced, confident and knowledgeable a user is with a topic, the more complex and specific the queries they issue will be (2). Three research questions will be investigated to achieve the project's objective.

- Q1: Does query complexity increase within a search session and is there a positive correlation between query complexity and topic specificity?
  If a correlation exists, it corroborates the evidence discussed in (2), that complexity is positively correlated with knowledge gain. If it can be shown that a user is increasing their query complexity, then it can be shown that they are gaining knowledge of a topic.
- Q2: Do high-complexity queries lead to more click-throughs?
  If this can be demonstrated, it would indicate that a user understands the need to be specific when issuing queries in order to discover the information they need. The assumption is that more click-throughs mean that the user is discovering relevant documents. Hence, an increase in click-throughs corresponds to a better, more satisfying search for the user.
- Q3: Do the topics of queries/clicked URLs converge to one specific topic within a session?
  A single informational search session is assumed to be associated with a specific information need. If topics converge within a session, it signifies that the user if targeting their queries at a specific topic. Analysing query reformulation in tandem with this can also provide some interesting insights into user behaviour.

The answers to these questions can provide further interesting insights into how users interact with a search engine, and how they simultaneously learn about their information need and how to query a search engine. It may be possible in future work to categorise users based on their search characteristics, deeming them to be "expert" or "novice" in particular topic areas and tailoring their search results based on this.

## 1.2 Choosing a Dataset

The chosen dataset for analysis is the AOL dataset(4). This dataset was chosen from a number of different options, including a Yandex dataset and internal Trinity College datasets. The internal TCD datasets were excluded due to the small number of users and a lack of diversity among users and search queries, as they were datasets from the online library. Yandex is a Russian search engine, and it was excluded due to language difference, as translating queries into English is both noisy and beyond the scope of research. Datasets from TREC and CLEF were also considered, but ultimately it was decided that the data did not accurately represent real-time use of an online search engine. The AOL dataset provided sufficient queries, retrievable content and numerous users, hence it was suited for the purposes of this study. A more detailed look at this dataset can be found in 3.1.

## 1.3 Dissertation Structure

The dissertation will be structured as follows: firstly, the background literature section, focussing on previous research discoveries that will aid in answering the research questions, including past research that was done with the AOL dataset. Then the data preparation techniques are outlined, as well as an in-depth look at the AOL dataset. This section presents some general statistics about the available data, and shows how the data was reduced for the purposes of this study. Following from this, the data analysis is discussed; how the results were obtained and what specific formulae were used. Finally, the results are presented and discussed.

# 2 Literature Review

This section will discuss the background literature to this project. The main focus will be on how other research has tackled the topic of user behaviour on the web, particularly how it defines and evaluates users domain knowledge/expertise. Other related papers will be in the area of query and topic analysis performed on search logs; what methods were used to undertake the research and what was learned from the results. Finally, papers directly related to the AOL search log dataset will be discussed. This project performs analysis on the AOL dataset, so it is beneficial to see what other research has already uncovered about this data.

## 2.1 User Expertise

Knowledge, according to the Oxford English Dictionary(5), is defined as "facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject". Expertise is then defined as "expert skill or knowledge in a particular field". This definition is rather broad, and shows that the word 'knowledge' can vary in meaning depending on the context in which it is used. In the context of search engines, knowledge can mean knowledge of the domain/topic that the user is searching in, but it can also mean experience using a search engine. It is therefore important to explicitly define what knowledge is in terms of the data included in search engine logs.

McAuley and Leskovec (6) focus on the user's experience when it comes to evaluating user expertise. They mention that the word 'expertise' simply refers to "some property of user evolution that is common to all users". This essentially means that there does not need to be a precise or correct definition, once all users are evaluated equally. However, this assumption applied to user expertise in giving online product reviews, which are more verbose than search queries. While this is an interesting concept, it may not apply to search log analysis.

Hölscher and Strube (7) address the problem of distinguishing experience of using a search engine and knowledge of the topic being searched for. They define search experience as "the knowledge and skills necessary to utilize the WWW ... successfully to solve information problems", as distinct from domain-specific background knowledge. They discovered interesting interactions between the two kinds of search knowledge, notably that novices tend to reformulate/reiterate the same queries over and over, and that experts in a topic will click on more links than novices. As well as this, they discovered that users with little domain knowledge used much longer queries, which is rather counter-intuitive. Experts in a domain demonstrated more flexible vocabulary when issuing queries. Similar findings are discussed in (8). While this study demonstrates some useful characteristics of user search behaviour, it is rather dated so the experiment on web expertise may not be applicable to a modern search engine.

Collins-Thompson et al. (9) analyse the user's readability level, and use that measure to customise the returned results. It is safe to assume that readability levels generally increase with age, experience and knowledge. It is unlikely that someone with the reading capability of a 10-year old would be an expert in radiology, for example. The study estimates a user's reading proficiency based on their search history, particularly dwell time on a document. They also mention that query length and the types of websites visited by the user can provide a good indication of readability. Reading level does not define the user's knowledge of a domain, but it can provide a good indication of knowledge so it is a useful factor to consider when analysing search queries.

A second challenge is how to actually measure knowledge; what scale should the measurement have, how do you compare the knowledge levels of two individuals, what makes Person A more "knowledgeable" than Person B? This is made more difficult by the limited information provided by search queries. As previously mentioned, search engine logs are anonymous, so it is very difficult to estimate the prior knowledge level of the user. Search queries tend to be short, directed sequences of words, that at first glance don't seem to provide any distinguishing information about the user issuing them.

Two papers highlighted in Chapter 1 use particular features/metrics of search logs to quantify a user's knowledge level (1, 3). Relevant features include domain count, focus, session length (in terms of pages clicked, queries issued and session duration) and query complexity. Domain count is the number of unique domains that a user has visited in a session. Both studies show that experts tend to visit a wider range of domains than novices. Focus is the overlap of the clicked page topics within a session; the study performed in (1) shows that the topic narrows initially in a session but then broadens again, indicating wider exploration by the user. Query complexity was measured in terms of the 'age of acquisition' of the query terms, and this metric increased steadily throughout a session. (3) showed that experts spend longer in a session, issue more queries and visit more pages in a session when compared to novice users. They reason that experts are more committed to a topic, possibly because the information is very important to their job/studies. These studies show that search logs can indicate knowledge of a user and can accurately distinguish groups of users.

More sophisticated features of search logs can be utilised to measure knowledge gain. Monitoring eye movement (10) can help to show a correlation between cognitive reading effort and domain knowledge. Mouse movement analysis (11) can also show how users make decisions on what they click on, which may correlate to their knowledge level. These advanced techniques are interesting to consider as alternative ways of gauging a user's expertise, however they are not applicable to the dataset used for this project.

Other notable investigations into user expertise include; investigating the effect of domain knowledge on search tactics (12), query reformulation and search term selection (13), and investigating whether user behaviour differs based on the task type (14). In all of these examples, there was a marked difference between the behaviour of an expert user and a novice user. This is positive for the outcome of this research project, as it means that search logs can be sufficient to make useful distinctions between users.

Based on these examples, it is possible to formulate a definition for knowledge in terms of the data made available in the AOL search logs. A user would be considered expert in a topic if they use very specific words when querying that topic, and if they click on specific pages associated with that topic. Within a particular topic, User A is considered more knowledgeable than User B if the majority of these conditions are satisfied:

- The average complexity of User A's queries is greater than User B's, based on the age-of-acquisition and specificity values of the search terms.
- The topic specificity of User A's queries is greater than User B's.
- User A narrows the topic of their queries more quickly than B.
- The topics of the pages clicked by User A are more specific than those of the pages clicked by User B.
- User A has issued more unique queries on the topic than User B.
- User A has clicked more pages on a topic than User B.

## 2.2 Query Analysis

This section will specifically discuss the features of a search query that can be analysed to give some useful information about a user. It is important to bear in mind the practical limitations of using a query to define the knowledge of a user. Swanson's Postulates of Impotence (15) make some relevant remarks about these limitations, particularly the first postulate: “An information need cannot be fully expressed as a search request that is independent of ... context”. He also makes the point that a user must describe what they don’t know in order to find what they don’t know, the so-called ‘paradox’ of Information Retrieval. However, the on-line world has changed significantly since Swanson published his paper, and advanced query analysis methods can now provide a good indication of the mindset of the user, the probable context of their query and the topic they are searching under.

One way of analysing queries is to categorise them into goals/type of query. It is widely agreed that there are three types of query, informational, navigational and transactional (16, 17, 18), although some papers (19) disregard transactional queries due to lack of consensus on what makes a query transactional. Informational queries occur when the user wants to seek information about some topic e.g. ‘irish civil war’, navigational queries occur when a user wants to find a specific website e.g. ‘facebook’ or ‘www.irishtimes.ie’. Transactional queries are usually when a user wants to perform an interaction with a website e.g. ‘free music download’, ‘buy second hand laptop’, which can be difficult to generalise. Most of the research agrees that the majority of search queries are informational in nature, a small amount are navigational (∼10%) and a varying amount are transactional. This is encouraging for the purposes of this research, as informational queries can provide the most insight into the knowledge level of the user. This categorisation is usually performed as a preliminary processing method, before further analysis on the queries.

Query reformulation is an important area of study when understanding user behaviour. Reformulation usually occurs when a user did not find their desired result with their initial query, so they modify the query in various ways before re-issuing. The types of modification preferred by users can tell a lot about how experienced they are at using search engines, and can also indicate to some degree their prior knowledge of the search topic. One particular study (20) defined three broad types of query reformulation; content (change the to meaning of the query), format (reordering words, re-spelling, punctuation) and resource (changing type of information resource e.g. news/image/music video). They discovered eight different reformulation patterns in their data, which may be useful when considering ways in which search engines can adapt to user behaviour. However, the study did not investigate the factors which influence query reformulation, nor the effectiveness of the patterns.

Huang and Efthimiadis (21) conduct a very in-depth study on query reformulation, using the AOL dataset. They combine ideas from eight previous studies to cover 13 different reformulation strategies. Certain strategies were found to be more beneficial to the user than others. For example, acronym expansion resulted in more clicks than reordering words. Based on the effectiveness of reformulation strategies, it can be inferred that a user who tends to use effective reformulation is likely to be more experienced at using search engines.

Of particular interest to this research is the word substitution strategy. This could be rephrased as changing the specificity of a query by replacing ambiguous words with more complex synonyms. It has been demonstrated that increasing both the term specificity (through synonyms) and the overall query specificity (narrowing the topic of the query) leads to more relevant results (22). An older study (8) demonstrated that expert searchers were affected more by lack of subject knowledge, using a thesaurus to look up synonyms more often. However, when they were familiar with the subject, they relied on using their own language to form and re-form queries. As the modern web does not usually involve the use of a thesaurus, it can be assumed that familiarity of a topic will give rise to more term substitutions, at least for an experienced web searcher.

Query complexity is strongly related to specificity. However, there is no definitive method to calculate complexity. Methods vary from using reading difficulty of terms (1), to basic statistics (term count, length, no. of characters) (3, 23), to query syntax (24). However, most of these studies demonstrate that query complexity increases within a session (1) and that more specific information needs lead to longer, more complex queries (25).

In summary, specifically analysing features of search queries can aid in measuring the knowledge level of a user. By only considering informational queries, a lot of 'noisy' queries can be ignored. Knowledge gain can be evaluated by measuring the increase of query complexity within a session. Domain experts can be distinguished from other users by analysing their query reformulation strategies, particularly word substitution.

## 2.3 Topic Analysis

Analysing a user's queries is very useful for understanding their search experience and technique. However, understanding the information need can affect the understanding of search techniques. For example, it can be shown that a user's queries increase in length throughout a search session, and that the complexity of the query terms is also increasing. But why is the user doing this? The answer to this question may lie in the topic of interest or the information need of the user.

Common practice is to define the query topic according to the results returned in a web directory. Using the Open Directory Project (ODP)(26) and Google Directory (27), the likely category of the query can be found. For example, the query "aerosols" has a top category of "Information/Science & Technology". Another study (28) analyses occurrences of term-pairs, with highly correlated pairs indicating the topic of the query. For example, the term-pair 'buffy-vampire' suggests the topic is "Buffy the Vampire Slayer", a popular TV show at the time. The same study also used human-classification to define query topics, which is not feasible for large amounts of data.

Although research has been able to define/infer query topics, the intent has always been to classify the queries by topic or simply to show the most commonly searched for topics. Very little work exists that discusses the idea of narrowing search topics within a session. Eickhoff et al. (1) discuss ‘focus’ and ‘entropy’ features of search queries, using ODP to categorise queries similar to (26). They find that the query topic narrows initially within a session, and then broadens, whereas the entropy (a measure of the diversity of topics in the results) decreases within a session. However, they merely use these findings to demonstrate that the user is learning something in a session, they do not investigate any correlations between focus and entropy, nor do they measure focus/entropy for specific topics.

This research project can fill the gap somewhat by answering the three research questions defined in 1.1. Particularly, linking query complexity with topic specificity and finding a correlation between the complexity and the URL topics can provide some new insights into user behaviour.

## 2.4 AOL Analysis

In this section, research that has been carried out on the AOL dataset will be discussed. The discoveries of these prior works may be useful when considering this project’s objective, and it is also interesting to see the varying approaches taken by other work to the same dataset.

Beitzel et al. (29) have performed various in-depth studies focussed on the area of automatic topical classification of queries. They consider a subset of the available AOL queries, manually classify them, and use them to train a machine learning algorithm to classify the remaining queries. Other work used these classified queries to investigate whether queries can be classified by both topic and user intent (30). Although this research is mainly focussed on how AOL can be used to train machine learning algorithms, some useful observations can be gleaned from it. For example, (30) shows that 99.7% of URL-type queries (queries containing http, www etc.) are navigational. By removing queries such as these, it reduces the noise when evaluating query search topics.

As has been mentioned previously, Huang and Efthimiadis (21) performed an in-depth query reformulation analysis on the AOL dataset. The aim of their study was to better understand how users reformulate their queries and whether certain reformulation strategies are more lucrative (in terms of clicked results) than others. They provide a lot of detail on how each reformulation strategy is detected, which forms a basis for a similar analysis in this research project (see 4.2). They were able to make some interesting discoveries about the effectiveness of the reformulation strategies, notably that the word substitution metric is correlated with different URL clicks and higher ranked results. This suggests that users are narrowing the topic of search to obtain similar, but better results. Other strategies like word addition, spell correction and acronym expansion are also shown to give rise to more clicks, particularly clicks on higher ranked results. These reformulations would be important to consider within the context of this research as well, as they can indicate that the user wishes to delve deeper and get better results for a particular search topic.

This discussion of background work has helped to formulate a definition of user knowledge, which forms the basis of the research questions posed in Section 1.1. Some of the methods used to evaluate query complexity, like age-of-acquisition, will be used in the data analysis, as detailed in Section 4.1. Some statistics outlined in other studies will be exploited to reduce the overall volume of data, as discussed in Section 3.1. As Section 2.3 demonstrated, there is a clear gap in research for analysing how the topics of both search queries and clicked document topics can aid in evaluating a user's expertise. While there are plenty of existing studies on how users reformulate queries, it is difficult to find any research that investigates why, and why certain reformulation strategies are chosen over others. This project hopes to provide some possible reasons why users behave the way they do, linking behaviour with information need.

# 3 Data Preparation

As the volume of data in the AOL dataset is quite vast, it was necessary to prepare and process the data for analysis. This section begins with an overview of the dataset itself, how its structured, and what data can be initially ignored. The approaches used to split the data into search sessions and generate a random sample of users are also outlined, as well as the methods used to categorise the data into topics.

## 3.1 The AOL Dataset

The AOL dataset (4) was released by AOL in August 2006, for use in research. The dataset contained search queries from 657,426 users, collected over a 3-month period from 01 March to 31 May 2006. The dataset is normalised, meaning that only the domain name of the clicked URL is shown and some queries have been removed for privacy reasons. As mentioned in a previous study(21), this can make it difficult to see what the information need of the user actually is, and the progression of their search queries in a search session. Despite this, the dataset was subsequently taken down by AOL due to backlash from the public, particularly users of AOL, as some of the queries contained enough information to personally identify some of the users (31). However, the dataset is still widely available on various mirror sites, hence its use in various research as discussed in Section 1.2. Partly as a consequence of the issues with this data release, it can often be very difficult to access search logs from commercial search engines, as there is a risk that the privacy of the user could be compromised. It would have been ideal to consider more recent data, particularly considering all the technological advances in search engines since 2006, however, this was not possible.

The dataset contains over 36 million lines of data, each one corresponding to a search query issued by the user. Each user is assigned a unique, anonymous user ID. The data is sorted first by user ID, then sequentially by the date and time the query was issued. Each line of data has the User ID, the Search Query and the QueryTime. In the case of a click-through, the line will include the Clicked URL and the Rank of that URL in the result list. For this research, the Rank of the clicked document is not important, so this data will not be used. There are over 3.6 million unique query terms, which form a Zipf distribution as seen in Figure 3.1. The relationship between terms and their frequency is always useful to consider. This clearly shows that the more specific the term, the less frequently it occurs in the data. This is corroborated by Table 3.1, which shows the top 10 most frequently occurring terms in the search queries.

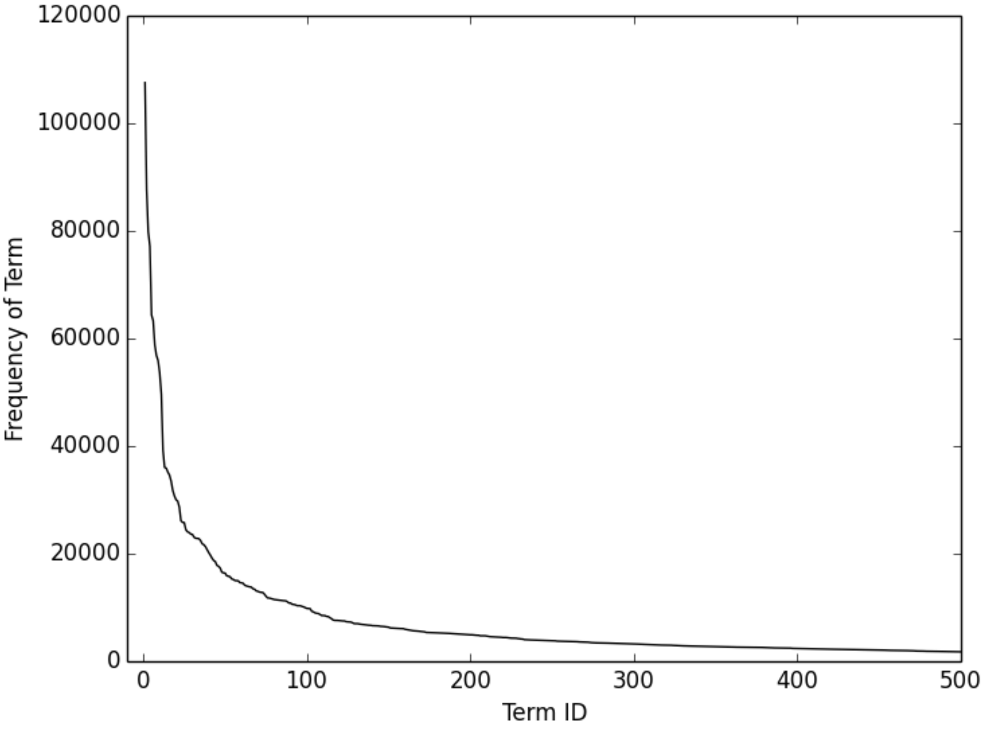

Figure 3.1: Zipf Distribution of query terms

| Term | Frequency |
|---|---|
| free | 445,815 |
| com | 375,963 |
| google | 366,218 |
| new | 268,398 |
| http | 260,863 |
| pictures | 236,865 |
| county | 231,465 |
| yahoo | 220,060 |
| www | 212,680 |
| how | 208,593 |

Table 3.1: Top 10 most commonly occurring query terms

Most of the very commonly occurring terms are stop-words, or general words, which would not be informative or useful when deriving user knowledge or query context. The term frequencies could also be considered relative to user search experience and domain knowledge. If a user tends to use very common terms very frequently, they probably don't know specifics about the domain. Users that use more specific terms are likely both more knowledgeable in the domain and more experienced at using the system as they demonstrate awareness that common terms will not yield good results.

Another interesting aspect of the data is that although the queries and URLs have been normalised, most URLs are only clicked once, as shown in Figure 3.2. This means that analysing the clicked URLs can provide some useful and specific information about the user. Once again, the most commonly occurring URLs are very popular search engines/social network sites of the time, as shown in Table 3.2.

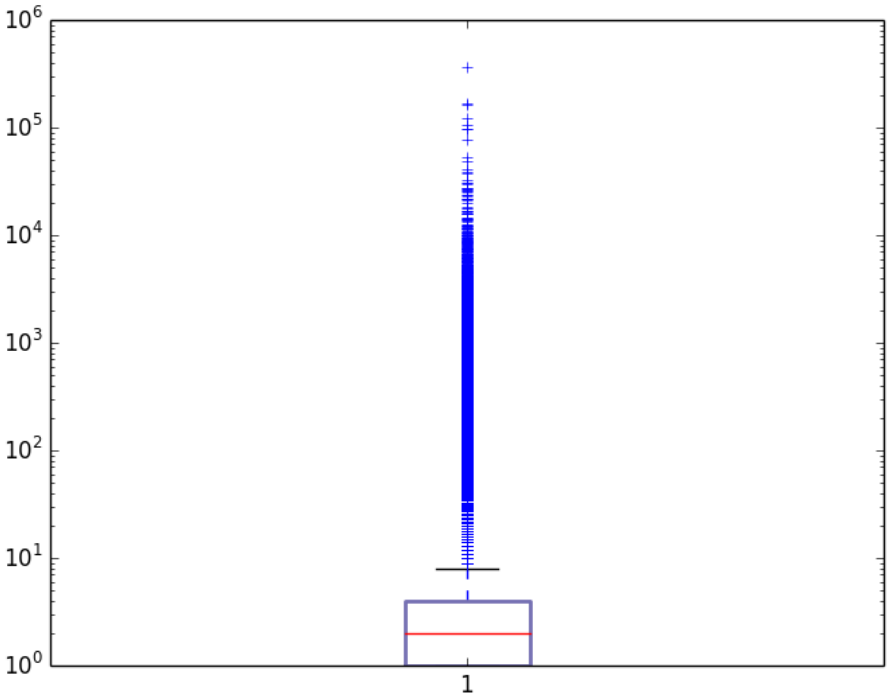

Figure 3.2: Distribution of URL clicks

These common URLs can be ignored for analysis, as they are usually the result of a navigational query. A navigational query, as discussed in Section 2.2, is issued with the intent of finding a specific website, rather than to satisfy an information need.

The paper published with the dataset (4) has some useful statistics on the search queries. These findings can help when formulating the methods of analysis for this research. Firstly, the paper showed that navigational queries make up 21% of the total query frequency. As already discussed in Section 2.4, most URL type queries are navigational, which means that 21% of all queries are irrelevant for this study. The mean number of terms in a query was found to be 3.5, a rather high value for a search engine, if you consider that queries at that

| Term | Frequency |
|---|---|
| www.google.com | 367033 |
| www.myspace.com | 167070 |
| www.yahoo.com | 161082 |
| en.wikipedia.org | 122540 |
| www.amazon.com | 106120 |
| www.imdb.com | 98549 |
| www.mapquest.com | 96136 |
| www.ebay.com | 77947 |
| mail.yahoo.com | 53978 |
| www.bankofamerica.com | 48545 |

Table 3.2: Top 10 most commonly clicked URLs

time were usually 2-3 keywords long. This indicates that there must be significant numbers of queries that are linguistic in nature, having many terms. This is encouraging, as the longer the query, the easier it is to define the topic of the query. It also gives more data to work with when evaluating the query complexity. They also found that a small percentage of users perform the majority of queries, with 41% of users only searching once per day. For the purpose of evaluating knowledge gain within a session, preferred users would have many sessions, with many queries in a session. This could mean that as much as 41% of users can be disregarded for this study. Based on these statistics, the volume of data that is applicable to this research can be vastly reduced. Removing navigational queries also reduces noise when evaluating query complexity and category.

## 3.2 Search Sessions

As this project seeks to demonstrate knowledge gain within a search session, it is first necessary to split the user queries based on session. The approach taken closely follows that of (21), who show that defining a random cut-off time is as effective as automatically detecting session end points. For this research, a session is considered terminated following a 20 minute inactivity period. Queries issued after this time are considered to be the start of a new session. Users are then clustered based on the number of search sessions they had, as shown in Table 3.3.

Users with less than a threshold number of sessions were disregarded for analysis. This threshold number was determined to be 20 sessions, meaning that a user would start a search session every 5 days on average. This session frequency is generally more interesting to consider than users with only 1 or 10 sessions, and provides sufficient data for potential inter-session analysis. By defining this cut-off point, the number of users considered for this study is now 178,146, or 27.1% of the total users.

| Max No. of Sessions | No. of Users |
|---|---|
| One | 97193 |
| Ten | 271976 |
| Twenty | 122962 |
| Thirty | 58089 |
| Fifty | 53033 |
| More | 54164 |

Table 3.3: Initial Search Session Clusters

However, what about the 41% of users that only issue one query per day? Some of these users could satisfy the session requirements, but they would be ineligible for this study, as single-query search sessions cannot be used to demonstrate knowledge gain. As well as this, the main focus of this study is intra-session analysis, so users with only one search query per session will not be useful. Hence, users can be restricted further based on the average number of queries in their sessions. The mean number of queries per session across all users was calculated to be 2.84, with a mode value of 1, which further proves the need for a minimum query threshold value. A threshold value of 3 was chosen, as the rounded mean value; assuming that a session is driven by a specific information need, this threshold should be sufficient to demonstrate different query reformulation strategies and topic-narrowing within sessions. It also reduces the likelihood of empty sessions after parsing out navigational queries. If one query in a session is a URL type, there are still two other queries that can be investigated. After this further refinement, the number of users was reduced to 192,710, a 70.69% reduction and the session clusters were also reduced (see Table 3.4).

| Max No. of Sessions | No. of Users |
|---|---|
| One | 21535 |
| Ten | 65958 |
| Twenty | 35209 |
| Thirty | 19680 |
| Fifty | 21825 |
| More | 28503 |

Table 3.4: Revised Search Session Clusters

## 3.3 Sampling Users

192,710 users still give a large volume of data to process. It was decided to take a random sample of 100 users to focus the analysis on. This is a common approach, similar to that taken by the studies mentioned in Section 2.4, where a small subset of data was analysed. The subset of data considered for this study is not randomly selected, but specifically fine-tuned for the purposes of this research to reduce noise when analysing the data. Selecting a random sample from the subset should accurately represent the subset as a whole, when properly conducted.

The sampling is performed as follows: first, users are clustered according to the number of search sessions they conducted over the three month period, as in Table 3.4. Users are then picked randomly from the applicable clusters according to the distribution of users across those clusters, as in 1; where $C_i$ is the session cluster to select a sample from, $T$ is the total number of users, $n$ is the number of applicable clusters, and $R$ is the number of random users to sample from $C_i$.

$$R = 100\frac{C_i}{T - sum_{k=1}^{n} C_k} \tag{1}$$

The list of users in $C_i$ is then randomly scrambled, and the first $R$ users are taken as the sample from that cluster. This is repeated for all $n$ clusters. Multiplying the cluster ratio by 100 means that the sum of the user samples from all the clusters will be 100.

## 3.4 Query Classification

This is the final stage of data preparation, where noisy queries are removed and the topics/categorises of the queries are finalised.

Queries are classified as informational or navigational simply by checking the query for the presence of certain character sequences like ‘http’ and ‘www’. Top-level domain names were also included in this check, based on the list used in (21) for their URL stripping analysis. They make the point that there are many top-level domains and infinite second-level domains, which cannot all be checked in this way and would require a more sophisticated rule. However, most navigational queries are caught by the ‘www’ and ‘http’ check, which is sufficient for this research. This introduced some problems, as some users queries were almost exclusively structured like a URL e.g. user 4379334. This user had 155 search queries across 30 different sessions. Of these queries, only 34 had click-throughs, and all the queries were one-word, URL type queries e.g. www.smalltravelkeybroadsforsale.com, www.findapianodealerclosetomyhome.com. They clearly have little to no understanding of how search engines operate, or how to properly use them. While this does not mean that the user is not knowledgeable within the domain they are querying, it is difficult to assess what domain that is, particularly since they had very few click-throughs as a result of convoluted queries. Fortunately, no other sample users demonstrated this type of query behaviour, so only user 4379334 was disregarded for analysis.

The classification of transactional queries was ignored on the basis that it is outside the scope of this research. Also, as other papers have discovered (19), the process of defining what a transactional query is can be difficult and prone to error. (30) found that transactional queries make up approximately 15.3% of all queries in AOL. However, it can be argued that a user seeking to purchase an item/download sources etc., can still demonstrate knowledge within that domain. For example, a user looking to 'purchase a laptop' could first consider general sites like 'pcworld.ie' or 'amazon.com'. Their search could then be refined to 'purchase macbook', and the results would be more specific e.g. 'apple.com', 'compub.ie'. This demonstrates topic convergence, and the user is learning about the types of laptop they are interested in. Therefore, transactional queries will be assumed informational for this research.

## 3.5 Categorisation of Queries and URLs

Now that the queries have been finalised, and the navigational queries removed, categorisation into topics of interest can be carried out. As mentioned in Section 2.3, there has been success in using online web directories to figure out the topic of a query. Unfortunately, Google Directory was discontinued in 2011 and the Open Directory Project (ODP) was taken down in 2017. Although archived versions of these sites exist, they cannot be queried and so provide little help. However, a successor version of ODP can be found online, known as Curlie[1].

Curlie returns the top five most likely categories, along with the top twenty sites and their categories, for a query. It is not clear how the top five categories are decided, as they are not consistent and are often completely disjoint from the website categories and the query itself. Hence, the categories of the returned websites were considered instead.

Initially, the most commonly occurring category from the top twenty websites was assumed to be the topic of the search query. However, upon further investigation, it was found that the topic of the highest returned result was usually the most accurate topic for the query. The topic of the top ranked result was then given a higher weight when counting the most common category. In cases where there is a tie, the highest ranking category is taken as the topic. Some queries did not return any results from Curlie, mainly due to misspellings of search terms. As Curlie is not a search engine, it lacks the sophisticated measures to automatically suggest/reformulate a misspelled query. However, this should not adversely affect the results, as the query topic can be inferred somewhat by the other queries in the session.

The same process was initially completed for the clicked URLs in the sample. However, when the categories were generated, many of them seemed completely unrelated to the URL. Upon

[1] Curlie.org: http://curlie.org/

further investigation, it was discovered that Curlie often returned the 'best-match' URL to the queried URL, which in the majority of cases was a completely unrelated URL e.g. Query: 'http://www.cheap-cds.com' Returned URL: 'http://www.cheap-hotel-florence.com/'. The process was then fine-tuned to only return categories of sites that matched the domain name of the queried URL. In addition, in cases where there was an exact match to the queried URL, that category was immediately taken as the topic of the URL. Curlie was able to interpret the majority of the clicked URLs correctly and return the appropriate results. However, problems were encountered as the data is quite dated. Some URLs that existed in 2006 are no longer active and thus are unable to be found by Curlie. In fact, of the clicked URLs of the sample users, only 53.9% were able to be categorised by Curlie.

Another issue was discovered when querying Curlie with very general URLs e.g. 'http://www.google.com'. An appropriate category for Google is 'Computers/Internet/Searching/Search Engines', however most of the results returned by Curlie are in various non-English languages. This is due to the fact that 'google.com' is used by many countries as the default top-level domain, rather than specifying their own top-level domain (like google.ie or google.co.uk). Besides this, as the paths of the URLs were removed during the normalisation process, the specific resource that the user was clicking cannot be determined. When performing the in-depth topical analysis, it bears keeping in mind that such URLs are not very informative or useful, and could cause the results to skew. Returning to the idea demonstrated by Figure 3.2, very common URLs do not distinguish users or topics, and this will be taken into account in the in-depth analysis.

# 4 Data Analysis

Following on from the Data Preparation, this section will outline how the data was analysed to answer the research questions discussed in the Introduction. Firstly, the methods to evaluate the complexity of the search query will be discussed, including methods which were tried initially and then discarded. An overview of the query reformulation analysis will then be presented, and any correlations between reformulation strategies and query complexity will be demonstrated. Finally, the topic analysis of the queries and URLs will be outlined, with specific regard to how the topic convergence and specificity were calculated.

## 4.1 Query Complexity

Two of the research questions of this study address the notion of query complexity. "1. Does query complexity increase within a search session and is there a positive correlation between query complexity and topic specificity?" and "2. Do high-complexity queries lead to more click-throughs?". This section, along with the following, detail how results were generated to provide an answer to these questions. The topic of query analysis was discussed in Section 2.2, giving an overview of the various methods used by other researchers. These methods provide a base from which to form a new method of evaluating complexity.

Initially, it was decided that the query complexity would be measured as some combination of term count, term specificity and noun count. The term count was calculated by simply counting the number of spaces in a query. The idea behind using this metric was that domain experts would issue longer queries, as mentioned in (8). However, this discovery was valid for SQL-type searching. Modern search engines allow for entire linguistic phrases to be issued, so it can be the case that a single-term query can be more specific than a multiple-term query e.g. 'what to cook for someone who is allergic to dairy' vs 'lactose-intolerant recipes'. Due to the uncertainty about what term count actually tells us about the expertise of the user, this metric was disregarded.

A more complex metric to consider is the type of terms used. Although overall term count can tell very little, the counts of nouns, verbs, adjectives etc. might indicate user expertise.

Python's Natural Language Toolkit(NLTK)(32) was used for this analysis. NLTK allows for Part-Of-Speech(POS) tagging of text, taking in a string of words and tagging them according to their grammatical role in the sentence. For example, the search query 'how to define search query complexity?' is tagged as 'how/WRB to/TO define/VB search/NN query/NN complexity/NN'. The nouns in the query give the most information about what topic the user wants to learn about (in this case, search query complexity), and can be assumed to be the keywords of the query. Therefore, a user that tends to use a lot of nouns in their search queries must be more knowledgeable than another user. This assumption is somewhat naive, especially when considering the field of medicine. 'Heart/NN attack/NN' and 'Myocardial/JJ infarction/NN' both mean the same thing, but the first query has more nouns than the second query, meaning that this metric would indicate that user who issued the first query is more expert than the user who issued the second query. This is clearly an incorrect assumption. Based on this example, the noun count of queries was deemed too inconsistent to be used as a reliable measure of complexity.

A dataset that was briefly considered for a readability-type metric was the 'Oxford 3000'[1], a list of the most important words to learn in English. This list is specifically aimed at learners of English, and the 3,000 words are selected based on frequency and also familiarity. The intention is that these words can be used to explain what you mean when you don't know a more specific word. However, the words are not necessarily 'easy' or 'basic' words, as some of them occur rather infrequently in everyday speech e.g. candidate, venture, significant. The intention was to penalise the complexity value of a query if one of the terms occurred in the list. However, as the list contains terms of varying difficulty and specificity, it was decided that an alternative method should be found.

An interesting approach was taken by (1), who used age-of-acquisition of query terms to evaluate complexity. This study had encouraging results, as they were able to demonstrate that query complexity increased within a session, albeit on a different search log dataset. This research used a dataset published by Kuperman et al. (33), who calculated the age-of-acquisition(AoA) value for over 30,000 terms by surveying US residents. The terms are all base words that are used most frequently as nouns, verbs or adjectives. The dataset uses American spelling for the terms, which is appropriate as AOL is an American search engine. The original data lists the AoA value as the average age at which people usually encounter the term for the first time. This value was normalised according to 2; where $nAoA_i$ is the normalised AoA value and $AoA_i$ is the original value.

$$nAoA_i = \frac{AoA_i - max_{k=1}^{n} AoA_k}{max_{k=1}^{n} AoA_k - min_{k=1}^{n} AoA_k} \tag{2}$$

[1] More information: https://www.oxfordlearnersdictionaries.com/about/oxford3000

The query is then split into its terms, and the $NAoA_i$ value is calculated for each term. The maximum AoA value of the query terms is used in the complexity calculation value. This is preferred over the average AoA value, which would be left-skewed, as the majority of words are encountered at younger ages. In the case where a term is not one of the 30,000 base terms, it is given the value 0.4, the average AoA value across all the terms. The higher the AoA value of a query, the more complex the query is assumed to be. This value essentially estimates the age of the user based on what language they use. It is assumed that older users are more likely to be domain experts (very few 10 year-olds are qualified doctors). This also links with the idea of term specificity, as most rarely occurring terms are encountered at an older age, due to assigned school reading, college research, professional documents etc.

When considering semantic meanings of words, the use of a dictionary is imperative. WordNet(34, 35) is an online database of English, and is essentially a tree-like dictionary. Words are grouped according to their meaning, and the semantic relationship between terms is labelled. It can be very useful when calculating metrics like term specificity, which is also referred to as the information content of a term. Basic methods of calculating term or concept specificity involve measuring the depth of the term in the WordNet tree, the number of ancestors of the term or the number of concepts it subsumes(36). These methods can also be combined into one calculation(37). These methods are known as intrinsic calculations of information content, as they depend on the underlying structure of the semantic graph.

An extrinsic approach is taken by Resnik(38), which builds on the intrinsic calculations and is one of the most popular and effective methods of measuring concept specificity. Resnik defines information content as the inverse log of the probability of encountering a term that can be described by that concept. This is more formally stated in 3; where $IC(u)$ is the information content of the concept $u$, $D(u)$ is the set of descendants of $u$, $I(D(u))$ is the set of entities annotated by a descendant and $I(C)$ is the set of entities annotated by a concept defined in $C$.

$$IC(u) = -log(\frac{|I(D(u))|}{|I(C)|}) \tag{3}$$

Essentially, the rarer a term is, the higher its Resnik value, and thus, the higher the specificity of that term. The Semantic Measures Library(SML)[2](39) is a Java library which has methods to calculate both intrinsic and extrinsic measures of term specificity. The terms are required to be POS tagged, which was carried out using NLTK. There were some problems with returning WordNet URIs for plural noun forms and adjectives. The plural

[2] Git repository: https://github.com/sharispe/slib

nouns were stemmed to their base form. However, the adjectives had to be ignored, as SML does not seem to recognise them.

When first used, it was unclear what the Resnik values of the terms signified, as the values had a very wide range. The library contains a normalised Resnik calculation, which returns values between 0 and 1. These values were preferred as they could easily tie-in with the normalised AoA values discussed earlier. The Resnik value for the whole query is taken to be the average of the Resnik values for each term in the query. Conversely to the AoA values, taking the maximum Resnik would actually cause right-skew in the data, as a high number of terms get a score of 1.

The final overall query complexity measure is simply calculated as the average of the Resnik and AoA values of the query, as summarised in 4; where $QC$ is the Query Complexity value, $t_i$ is a term in the query, $nIC(t_i)$ is the normalised Resnik value for term $t_i$, $nAoA(t_i)$ is the normalised age-of-acquisition value for term $t_i$, and $T$ is the total number of terms in the query.

$$QC = mean(max_{i=1}^{n} nIC(t_i), \frac{sum_{i=1}^{n} nAoA(t_i)}{T}) \tag{4}$$

In order to determine whether the query complexity is increasing within a session, the initial idea was to compare the complexity of the first query to the last. However, the complexity may fluctuate within a session as the user reformulates their query. A similar issue arises when comparing the minimum complexity to the maximum. Hence, the slope of the complexity values is taken. A positive slope indicates an increase in complexity and a negative slope indicates a decrease, and higher values of slope indicate a steeper rise/fall in complexity.

## 4.2 Query Reformulation

As discussed in Section 2.2, studying users query reformulation patterns and habits can provide useful insights into their search behaviour. Huang and Efthimiadis(21) carried out an in-depth reformulation analysis on the AOL dataset, which forms the basis for the reformulation strategies considered by this research. The strategies are listed and explained below, along with why they are considered important in the context of user expertise.

1. Word Addition/Subtraction
   This is one of the most common reformulation strategies performed by users, as demonstrated by (21) and Section 5. Word removal is calculated by recursively removing terms from $query_i$ until it matches $query_{i+1}$. Word addition is the reverse of

this method. A user will typically add words if the results they get are too general, and subtract words if the results are too specific. Although this is a common strategy, word addition can indicate that the user is attempting to narrow the scope of their search towards a specific topic or information need. It can also be investigated whether adding more words increases the complexity of the query.

2. Form/Expand Acronym
   This action is performed when the user issues an acronym query and then expands it into the full text e.g. 'pdf' -> 'probability distribution function'. The opposite can also occur, though less frequently. This is simply calculated by checking the initials of the terms in one query and matching them to the acronym in the other query. Expanding an acronym can indicate that the user is searching for a specific company/brand/term that is different from the common use of that acronym, as they didn't get relevant results on the first try. This can indicate that the user has knowledge of specific acronyms in a topic, and thus has high expertise.

3. Word Substitution
   This is perhaps the most important reformulation strategy for this research. Substitution occurs when one or more terms in the original query are replaced with other terms in a later query. The replacement terms are usually closely semantically related to the original terms, e.g. synonyms, hypernyms etc. By analysing the complexity of the reformulated query compared with the original query, it can be demonstrated that the user is narrowing their search further. This method tends to be used by more experienced searchers who may not be experts in the domain, so care will have to be taken when interpreting the use of this strategy.

4. Others
   For the sake of completeness, the other reformulation strategies outlined in (21) are also considered in the search query analysis, see Table 4.1 for the full list. These strategies tend to be simple changes to the query (word re-order, spell correction), and are not very useful for analysing query topic or demonstrating user knowledge. They can indicate lack of searching experience, as word re-order rarely improves search results and spelling correction is often already carried out by the search engine (e.g. Google's 'did you mean?').

The query reformulation strategies were initially defined for pairwise query comparison in (21), as they were performing the analysis on all the data in AOL. When comparing two queries, only one reformulation strategy can apply. Table 4.1 shows the order in which the strategies are checked, according to (21). They also discuss the challenge of multi-reformulation e.g. 'term speficity' -> 'calculating term specificity'. The second query is both adding words and correcting spelling. It is very difficult to detect multi-reformulations autonomously, as results would depend on the order in which the strategies were checked and some reformulation combinations would never occur together(like word reorder and acronym). Hence the reason why the reformulations are checked in a specific order, and the first detected reformulation is taken as the correct label for that query.

|  | Reformulation Strategy |
|---|---|
| 1 | same |
| 2 | word reorder |
| 3 | word addition |
| 4 | word removal |
| 5 | url strip |
| 6 | form acronym |
| 7 | expand acronym |
| 8 | substring |
| 9 | superstring |
| 10 | word substitution |
| 11 | spell correction |

Table 4.1: Reformulation Detection Order

## 4.3 Topic Analysis

This area of analysis mostly concerns the third research question "Do the topics of queries/clicked URLs converge to one specific topic within a session?" This section will discuss how topic specificity is defined, relative to the structure of the category returned by the Curlie web directory, and also how convergence is evaluated.

As already mentioned in Section 3.5, the most frequent returned category for a query is taken as the topic of the query. The category structure is a nested form e.g. 'Computers/Internet/Searching/Search Engine'. The general category is 'Computers', with 'Internet' as the sub-category and so on. In this way, by counting the number of '/' characters in the category, the depth or specificity of the category can be estimated. A problem with this method of evaluation is the 'Regional' categories. These categories tend to be a listing of geographical regions in the world, and thus the depth is very large, despite the category not being very specific e.g. for the URL 'http://www.babynames.org.uk', the returned category is 'Regional/Europe/United_Kingdom/Wales/Society_and_Culture/Genealogy/'. It is obvious that the underlying topic of the URL is 'Society_and_Culture/Genealogy/', which has a depth of 2, rather than the 'Regional' category which has a depth of 6. However, this can be difficult to deal with, as the 'Regional' categories have varying length, depending on the top-level domain of the clicked URL, or the results of a search query. In future analysis, this could be caught by doing some semantic analysis of each nested term in the category and comparing to the search query. If the meaning of the query term is similar to the meaning of the category field, then that field should be taken as the starting point of the query topic. This could be done using WordNet to compare the synsets of the terms, or the Semantic Measures Library, which contains methods to calculate term similarity. However, these measures were not carried out for this research.

Topic convergence refers to the idea that as the user continues to query within a search session, they are directing their queries towards a specific topic or information need. Hence, as the search session continues, the depth of the topics should increase. The overlap of topics can also be evaluated. This can be used to show the correlations between the clicked URL topic and the query topic. When comparing topics between queries, the overlap should increase as the session goes on. The overlap is calculated as follows: If the entire category is the prefix of another category, the overlap is the depth of the first category. If it is not a prefix, the last field in the category is removed, and the check is repeated. This is a similar idea to the substring/superstring reformulation check carried out in (21).

# 5 Results

The final random sample of 100 users had a total of 4,907 sessions and 21,212 queries, with the average number of queries per session calculated as 5.09. The maximum number of queries in a session was 111, and 1,172(5.5%) sessions were single-query search sessions. Of the total queries, 7,284(34.3%) were unique, and 2,743(12.9%) were classed as navigational queries. 12,512 queries had a click-through, with 7,210 unique URLs clicked as a result.

As discussed in Section 4.1, the query complexity values were generated for each query, and then the slope of the values per session was calculated. The mean overall complexity value is 0.417, and the mode complexity is 0.2. The maximum complexity score across all queries is 0.855 for the query "mikado", whereas the minimum complexity score is 0.025 for the query "my". The distribution of the query complexity values is plotted in Figure 5.1. The mode value arises when the query terms receive a Resnik score of 0 (indicating that they are not listed in the WordNet database), and a default age-of-acquisition(AoA) score of 0.4. Equation 4 then gives a complexity value of 0.2 for such queries. This would occur in queries where the term is misspelled ("horsess"), the term is an acronym("wwe"), the term is a

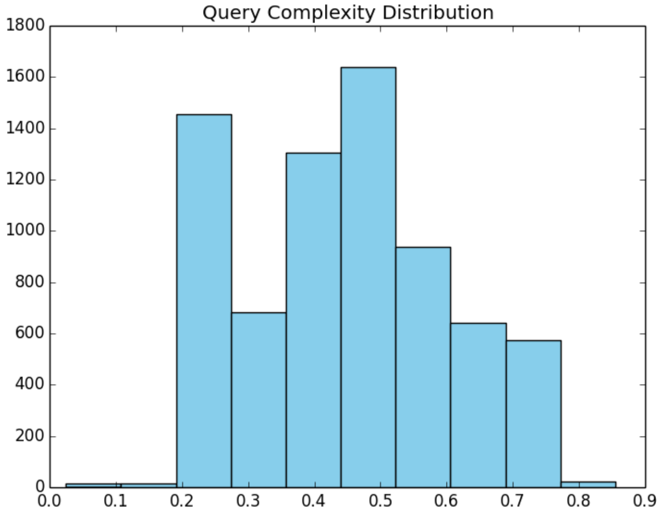

Figure 5.1: Query Complexity Distribution

person's name("rose hulman") or the term is formed of many words strung together("masterelectriciantestonline").

Figure 5.2 shows the plot of the complexity slopes against the session length. The session length is the number of queries in a session, but consecutive repeated queries are not counted. On first glance, this graph appears to be symmetrical about the x-axis, meaning that there are an equal number of sessions which demonstrate increasing and decreasing complexity. However, most sessions actually experience no change in frequency(2,491), with 970 showing a rise in complexity and 703 showing a decline. As Figure 5.3 demonstrates, there are more increasing complexity sessions than decreasing and therefore there is a tendency for users to increase the complexity of their queries within a session. The high number of sessions with no change is mainly due to the number of single-query search sessions (2,322).

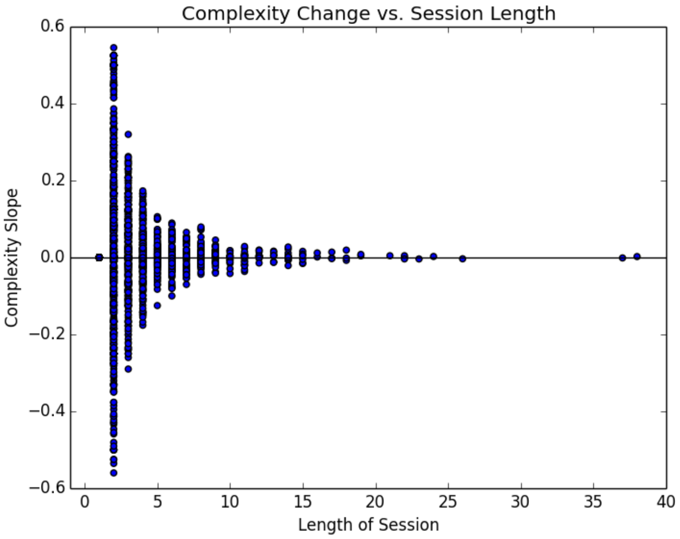

Figure 5.2: Complexity vs Session Length

Shorter sessions show the greatest spread of complexity slope, which is to be expected as demonstrated by Table 5.1. Longer sessions contain more queries and therefore the slope evens out through the session. Looking at the maximum and minimum complexity slope

| Session Length = 2 | | Session Length = 4 | |
|---|---|---|---|
| Query | Complexity | Query | Complexity |
| gamebrew | 0.2 | rihanna | 0.2 |
| bike mania | 0.7056 | rhianna | 0.2 |
| | | rhianna a girl like me | 0.2838 |
| | | van morrison | 0.7 |
| Slope: 0.5056 | | Slope: 0.1583 | |

Table 5.1: Example of Query Slope

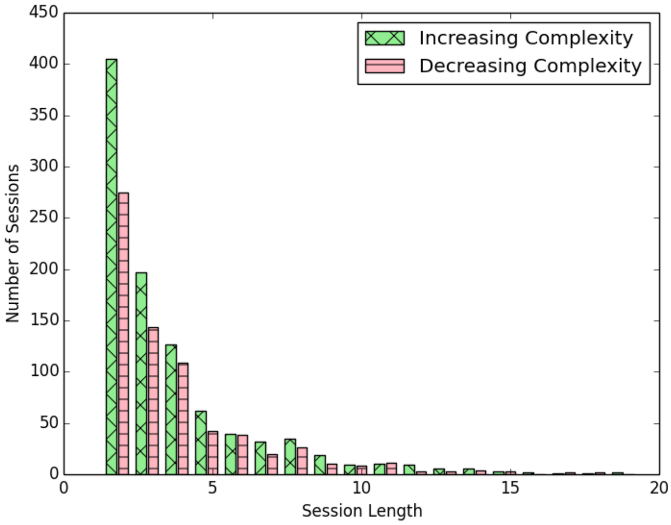

Figure 5.3: Comparing Increasing and Decreasing Complexity Sessions

values for each session length, the shape is similar to that of a decreasing exponential function, which seems to indicate that query complexity values approach zero as the session lengths approach infinity. Hence there is a moderate negative correlation between session length and rate of increase of complexity, with the correlation co-efficient calculated at -0.582.

Similarly, it can be shown that the query topics also tend to increase within a session (see Figure 5.4), though less consistently than the complexity values, with 709 sessions showing an increase and 589 showing a decrease. 2,588 sessions showed no change in topic, and this is likely due to the high number of queries that returned no results from Curlie (1,389). The

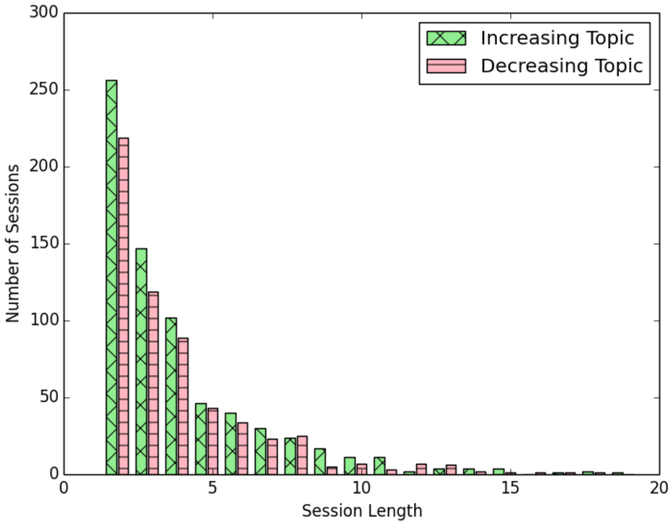

Figure 5.4: Comparing Increasing and Decreasing Query Topic Sessions

correlation coefficient for the rate of increase of query topic depth with session length is -0.570. This seems to indicate that there is a positive correlation between query topic depth and query complexity. However, the correlation coefficient of topic depth with complexity was found to be 0.221, which is too low to conclusively prove that a positive correlation exists. However, it can be said that there is a tendency for higher complexity queries to give deeper topics.

The URL topics show a slightly different pattern (see Figure 5.5), with 367 increasing topic sessions and 342 decreasing. 3,177 sessions showed no change in the URL topic, which is 64.7% of the total session number. Of the 7,210 unique URLs in the sample, only 41.8% returned a result from Curlie. This is likely due to the age of the dataset, as URLs may no longer exist, and partly due to the limitations of Curlie's search effectiveness. This would be a factor in the much higher correlation coefficient of -0.651 between the rate of increase of URL topic depth and session length.

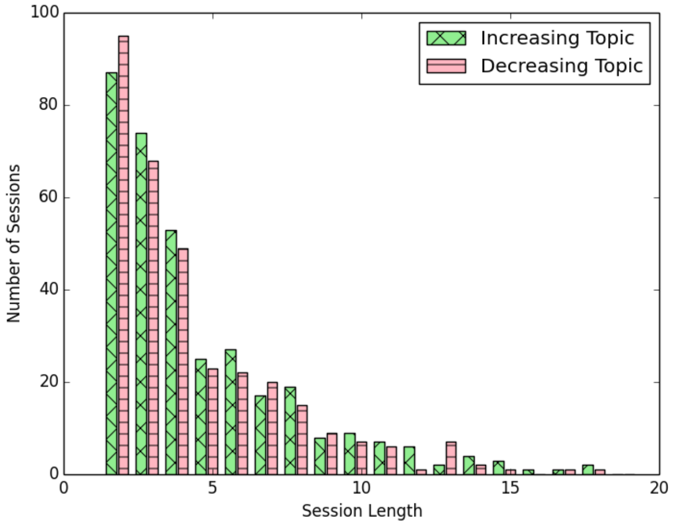

Figure 5.5: Comparing Increasing and Decreasing URL Topic Sessions

Comparing query complexity to URL clicks was rather difficult, as the majority of queries(88.3%) only had one click associated with them. As a result, there is no correlation between complexity and clicks, as shown in Figure 5.6. The graph does show that for queries with clicks higher than 15, the complexity tends to be higher. However, as the majority of queries have between 1 and 5 clicks, the correlation is skewed significantly. Another problem that this graph shows is the number of queries with a default complexity score of 0.2, which would also skew the result.

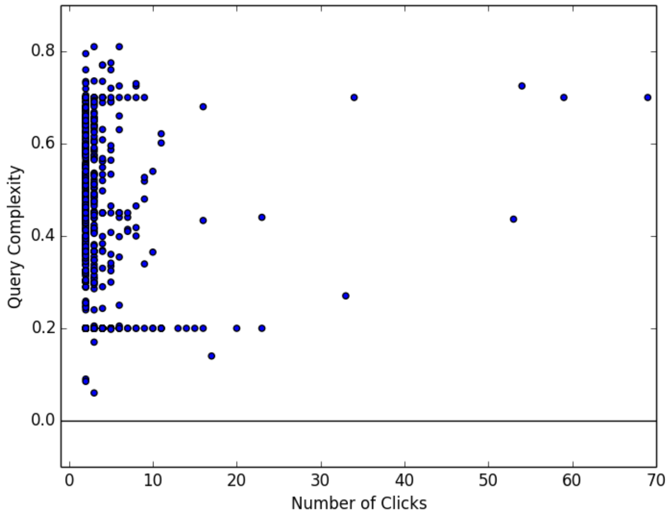

Figure 5.6: Query Complexity vs Query Clicks

The query topic overlap slope was also calculated for each session. The median and mode of the topic overlap were both 0, which is due to the high number of repeated queries in the dataset. The maximum rate of increase was found to be 0.917, with the minimum as -0.05. Once again, similar to the rate of change of complexity, the majority of sessions experienced no change (2,086). However, the number of decreasing sessions was quite small(125). This means that query complexity, topic depth and topic overlap are all more likely to increase during a session than to decrease.

The frequency of reformulation strategies across all queries in the random sample is shown in Figure 5.7. This figure does not show the queries that were classed as 'new' (7,124) or

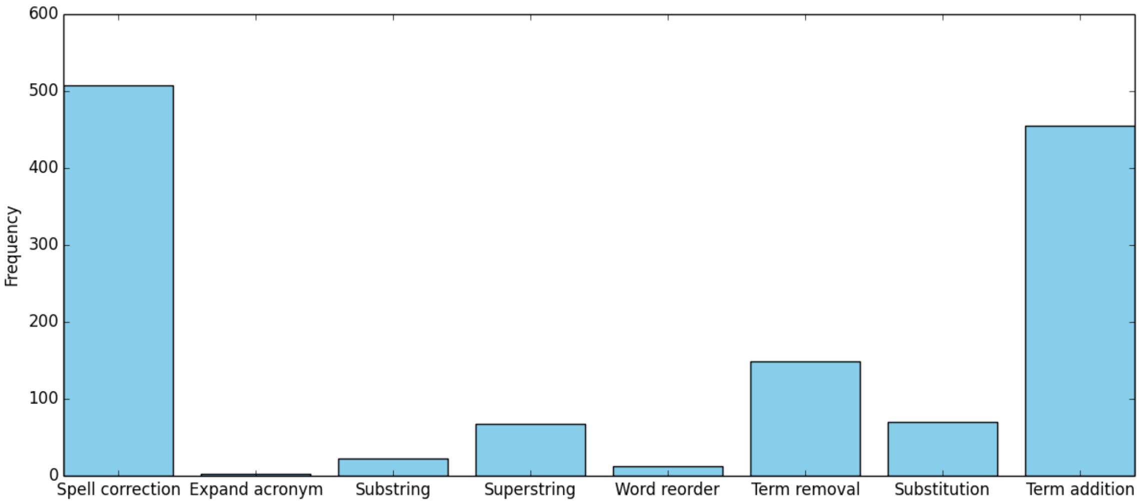

Figure 5.7: Reformulation Strategy Frequency

‘same’ (7,346). As expected, the most common reformulation strategies are spell correction and term addition. As discussed in Section 4.2, reformulations like term addition and substitution can give the most interesting insights into a user’s behaviour. The average complexity score increase for the four most frequent reformulation strategies is shown in Table 5.2. Most of these values are surprising, particularly the substitution values. However, substitution does not necessarily mean that the substituted word is more specific than the original word. Perhaps if the specificity of the substituted words had been analysed with the changes in complexity score, the result would be less surprising. In terms of indicating user expertise, the conclusion is that term addition and spell correction lead to more complex, defined queries, which tend to lead to deeper, more specific topics.

| | Spell Correction | Term Removal | Substitution | Term Addition |
|---|---|---|---|---|
| No. Increasing | 267 | 53 | 21 | 235 |
| No. Decreasing | 64 | 85 | 27 | 185 |
| Average | 0.076 | -0.017 | -0.006 | 0.008 |

Table 5.2: Average Complexity Increase for Reformulation Strategies

## 5.1 Discussion

The results presented above were calculated in order to provide answers to the research questions defined in Section 1.1, and re-stated below:

- Q1: Does query complexity increase within a search session and is there a positive correlation between query complexity and topic specificity?
  The results show that there is a marked tendency for query complexity to increase within a search session. This means that, regardless of domain knowledge, users increase search term specificity to reach their particular information need. There is also a slight positive correlation between query complexity and topic specificity. It can be inferred that a user searching within a shallow sub-topic will have lower complexity queries than a user searching within a deeper sub-topic. Assuming that a deeper sub-topic means a deeper understanding of the high-level topic, it can be concluded that higher values of query complexity indicate higher levels of expertise within a topic.
- Q2: Do high-complexity queries lead to more click-throughs?
  The correlation between complexity score and click-throughs was inconclusive, although queries with very high numbers of clicks tended to have a higher complexity score. However, most queries had between 1 and 5 clicks, with varying complexity. A better approach might have been to look for a correlation between increasing complexity and increasing number of clicks within a session. It could then be shown

that the user was discovering more relevant content as they increased the specificity of their query.

- Q3: Do the topics of queries converge to one specific topic within a session?
This convergence was calculated as the overlap between query topics within a session. The tendency of the overlap was to remain constant throughout a session, though the overlap was more likely to increase than decrease overall. Coupled with the observation from Q1, this means that in general, query complexity, topic depth and topic overlap all tend to increase during a search session. The reformulation analysis provided some unexpected results, with spell correction and term addition being the only strategies that showed an average increase in complexity and topic depth in a session. This suggests that simpler reformulation techniques can provide more dramatic results, though this may be because the original query was of low quality (especially in the case of spell correction).

## 5.2 Limitations

This section discusses the challenges and problems of this research and suggests reasons as to why some of the results did not go as expected. Particular problem areas are the AOL dataset and the use of the web directory Curlie. Suggestions are also provided on ways to improve results, and potential directions of future work in this area.

One of the first apparent issues with this research was the topic analysis of queries and URLs. The lack of prior research into this area made it difficult to know the best approach to take. Further complications arose when it was discovered that the web directories that had been successful at categorising queries in the past had been discontinued. Use of a web directory was a simple and relatively quick way of categorising a large number of search queries. However, there were a lot of issues with inaccurate categories, non-English categories and some queries returning no results at all. Choosing the best topic for a query was also a challenge, as trust had to be placed in the Curlie ranking algorithm to return the best result first. Although the web directory approach is quick and easy, it could also be paired with some semantic analysis of the query to help choose the best category. In the case where queries have no category results, the topic could be inferred based on that queries semantic similarity to another query or based on the category of the clicked URL associated with that query (if any).

The query complexity calculation also had a number of problems. As has already been mentioned, most queries end up with a default complexity score of 0.2. It is clear that the Resnik score and age-of-acquisition value are not sufficient to calculate an accurate complexity score for all queries. Some of the other metrics discussed in Section 4.1, like term count or noun count, could have been taken into account in some way. This would require a lot of testing and combinations of different parameters and variables, and could probably be the main focus of an entire research project.

The AOL dataset provided its own challenges. As has been mentioned in Section 3.1, the clicked URLs were stripped to the base domain, punctuation and capitalisation were removed from the queries, and some queries were removed altogether. This made it more difficult to get good results from Curlie. Another problem was the age of the dataset, as a significant amount of clicked URLs no longer exist. However, as mentioned in Section 3.1, there are not many publicly available search logs, which means that researchers are restricted on choice of dataset. It is possible that a more recent dataset might have introduced fewer problems. Another alternative would be to conduct a controlled experiment, where users volunteer for the study and are assessed before and after their search sessions. This would be useful in showing whether search logs can accurately predict a user's level of expertise.

# 6 Conclusion

This project discussed the use of search engine logs as a means of analysing and understanding user behaviour and domain expertise. The objective of the research was to demonstrate that search logs can provide valuable insights into how a user's behaviour changes within a search session as they work towards a specific information need. The analysis of query complexity and topic depth was performed with this objective in mind. The results demonstrated that query complexity tended to increase within a search session. A positive correlation was observed between the query complexity scores and the depth of the query topics. It was also shown that topic depth and topic overlap are more likely to increase than decrease within a search session. These observations lead to the conclusion that users refine their search queries towards a specific information need within a session. It can also be inferred that users with high query complexity scores for a topic are more likely to be knowledgeable in that topic.

These observations contribute to our understanding of users behaviour, in particular that calculating a complexity score for queries can show the focus of a user towards a particular information need. This research also highlights scope for future work. With the combination of a more sophisticated query complexity equation and semantic-based topic analysis, it may be possible to categorise users into different expertise levels based on their search queries.

# Bibliography

[1] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: a query log analysis of within-session learning. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 223–232. ACM, 2014.

[2] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. Analyzing knowledge gain of users in informational search sessions on the web. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*, pages 2–11. ACM, 2018.

[3] Ryen W White, Susan T Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*, pages 132–141. ACM, 2009.

[4] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *InfoScale*, volume 152, page 1, 2006.

[5] Oxford English Oxford. *Oxford English Dictionary*. Oxford: Oxford University Press, 2009.

[6] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908. ACM, 2013.

[7] Christoph Hölscher and Gerhard Strube. Web search behavior of internet experts and newbies. *Computer networks*, 33(1-6):337–346, 2000.

[8] Ingrid Hsieh-Yee. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*, 44(3):161, 1993.

[9] Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Sebastian De La Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 403–412. ACM, 2011.

[10] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, 49(5):1075–1091, 2013.

[11] Paul E Stillman, Xi Shen, and Melissa J Ferguson. How mouse-tracking can advance social cognitive theory. *Trends In Cognitive Science*, 2018.

[12] Barbara M Wildemuth. The effects of domain knowledge on search tactic formulation. *Journal of the Association for Information Science and Technology*, 55(3):246–258, 2004.

[13] Helene A Hembrooke, Laura A Granka, Geraldine K Gay, and Elizabeth D Liddy. The effects of expertise and feedback on search term selection and subsequent learning. *Journal of the Association for Information Science and Technology*, 56(8):861–871, 2005.

[14] Jingjing Liu, Michael J Cole, Chang Liu, Ralf Bierig, Jacek Gwizdka, Nicholas J Belkin, Jun Zhang, and Xiangmin Zhang. Search behaviors in different task types. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 69–78. ACM, 2010.

[15] Don R Swanson. Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science (1986-1998)*, 39(2):92, 1988.

[16] Daniel E Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM, 2004.

[17] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[18] Bernard J Jansen, Danielle L Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251–1266, 2008.

[19] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 391–400. ACM, 2005.

[20] Soo Young Rieh and Hong Xie. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42(3):751–768, 2006.

[21] Jeff Huang and Efthimis N Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 77–86. ACM, 2009.

[22] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, pages 387–396. ACM, 2006.

[23] Michael Chau, Xiao Fang, and Olivia R Liu Sheng. Analysis of the query logs of a web site search engine. *Journal of the Association for Information Science and Technology*, 56(13):1363–1376, 2005.

[24] Bernard J Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, 28(3):407–432, 2006.

[25] Nina Phan, Peter Bailey, and Ross Wilkinson. Understanding the relationship of information need specificity to search query length. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 709–710. ACM, 2007.

[26] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Query enrichment for web-query classification. *ACM Transactions on Information Systems (TOIS)*, 24(3):320–352, 2006.

[27] Bernard J Jansen, Zhe Liu, Courtney Weaver, Gerry Campbell, and Matthew Gregg. Real time search on the web: Queries, topics, and economic value. *Information Processing & Management*, 47(4):491–506, 2011.

[28] Amanda Spink, Dietmar Wolfram, Major BJ Jansen, and Tefko Saracevic. Searching the web: The public and their queries. *Journal of the Association for Information Science and Technology*, 52(3):226–234, 2001.

[29] Steven M Beitzel, Eric C Jensen, David D Lewis, Abdur Chowdhury, and Ophir Frieder. Automatic classification of web queries using very large unlabeled query logs. *ACM Transactions on Information Systems (TOIS)*, 25(2):9, 2007.

[30] Bernard J Jansen and Danielle Booth. Classifying web queries by topic and user intent. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 4285–4290. ACM, 2010.

[31] Michael Barbaro, Tom Zeller, and Saul Hansell. A face is exposed for aol searcher no. 4417749. *New York Times*, 9(2008):8For, 2006.

[32] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[33] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990, 2012.

[34] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[35] Princeton University. About wordnet. Princeton University, 2018. https://wordnet.princeton.edu.

[36] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, volume 16, page 1089, 2004.

[37] Zili Zhou, Yanna Wang, and Junzhong Gu. A new model of information content for semantic similarity in wordnet. In *Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on*, volume 3, pages 85–89. IEEE, 2008.

[38] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[39] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5):740–742, 2013.