ANALYSIS OF WEB SEARCH QUERIES WITH VERY FEW OR NO RESULTS

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Erdem Sarıgil

September, 2012

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Özgür Ulusoy (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. İbrahim Körpeoğlu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Ahmet Coşar

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

ANALYSIS OF WEB SEARCH QUERIES WITH VERY FEW OR NO RESULTS

Erdem Sarıgil
M.S. in Computer Engineering
Supervisor: Prof. Dr. Özgür Ulusoy
September, 2012

Nowadays search engines have significant impacts on people's life with the rapid growth of World Wide Web. There are billions of web pages that include a huge amount of information. Search engines are indispensable tools for finding information on the Web. Despite the continuous efforts to improve the web search quality, a non-negligible fraction of user queries end up with very few or even no matching results in leading commercial web search engines. In this thesis, we provide the first detailed characterization of such queries based on an analysis of a real-life query log. Our experimental setup allows us to characterize the queries with few/no results and compare the mechanisms employed by the three major search engines to handle them. Furthermore, we build machine learning models for the prediction of query suggestion patterns and no-answer queries.

Keywords: Web search engines, search result quality, query difficulty, query results.

ÖZET

AZ CEVAPLI VEYA CEVAPSIZ İNTERNET ARAMA SORGULARININ ANALİZİ

Erdem Sarıgil
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Özgür Ulusoy
Eylül, 2012

İnternetin büyümesiyle birlikte arama motorları insanların hayatında önemli bir etkiye sahip olmuştur. Günümüzde, büyük miktarda bilgi içeren milyarlarca web sayfası mevcuttur. Arama motorları internetten bilgi edinmek için vazgeçilemez araçlardır. Arama sonucu kalitesini arttırmaya yönelik olarak gösterilen sürekli çabaya rağmen, kullanıcılar arama motorlarında azımsanmayacak oranda az cevaplı ya da cevapsız sorgu sonuçlarıyla karşılaşabilmektedirler. Bu tezde, gerçek sorgu "log"undan alınan bu tarz sorgular için ilk detaylı karakterizasyon analizi sağlanmaktadır. Deneysel düzenimiz üç önemli arama motorunun az cevaplı veya cevapsız sorgularla nasıl başa çıktığını karşılaştırmamıza olanak sağlayacak şekilde kurulmuştur. Ayrıca, sorgu öneri kalıplarının ve cevapsız sorguların tahmini için makine öğrenmesi modelleri geliştirilmiştir.

Anahtar sözcükler: Web arama motorları, arama sonucu kalitesi, sorgu zorluğu, sorgu sonuçları.

To my mother, father and brother

To all loved ones

Anneme, babama ve kardeşime

Tüm sevdiklerime

Acknowledgement

I would like to express my deepest gratitude to my supervisor Prof. Dr. Özgür Ulusoy for his kindness and invaluable guidance during this thesis.

I am grateful to my jury member Assoc. Prof. Dr. İbrahim Körpeoğlu and Assoc. Prof. Dr. Ahmet Coşar for spending their time and effort to read and comment on my thesis.

I would like to thank Dr. İsmail Sengör Altıngövde for his endless support, guidance, and encouragement during this research. Furthermore, I also thank to my colleagues Dr. Berkant Barla Cambazoğlu and Dr. Rıfat Özcan.

I would like to address my thanks to The Scientific and Technological Research Council of Turkey (TÜBİTAK) for their financial support within National Scholarship Programme for MSc Students.

I would like to thank to my office mate Oğuz Yılmaz for his friendship and for the enjoyable office times. He has been very supportive and helpful psychologically during my thesis study. I also would like to thank to Salim Arslan for his friendship. In addition, I would like to thank my friends Burak Aycan, Can Koyuncu, Gökhan Kul and Şadiye Alıcı for their caring friendship.

Last but not least, I would like to thank my family for being with me all the time. Their profound love, tremendous support and motivation led me to where I am today. With very special thanks, I dedicate this thesis to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

After the invention of web at the end of 1980s, search engines have become an issue and it did not take too much time to develop the first generation web search engines [1]. However, the first generation engines performed quite poorly especially when searching long queries. They returned hundreds or thousands of documents containing the keywords of the query that a user searched, but only a few of them were the most relevant documents [2] . However, after the first generation web search engines, the search engine technology has greatly improved, and web search engines have become the main source for reaching information on the web [1].

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engines that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This second form relies much more heavily on the computer itself to do the bulk of the work.

With the rapid growth of World Wide Web, search engines nowadays are significantly impacting people's daily life [3, 4]. Search engines are indispensable tools for finding information on the Web [5]. Nowadays more and more people are using Internet search engine to locate information on the web [6]. Although the web contains huge volumes of data, search engines generally present the most relevant results in less than a second when a user enters a query. Queries that users type as input are taken by search engines and web pages are presented to the users [7]. Of the roughly 2 billion daily web searches made by internet users [8], approximately 28% are modifications to the previous query [9].

A non-negligible portion of web searches end up with very few or even no results. As much as search engine users dislike seeing the message "your search did not match any documents", search engine companies are reluctant to display it. Yet, the users occasionally receive such messages, especially when they are searching for some content in a less common language, an unpopular web page, an infrequent term that for example an unusual file name produced by some malware, or even when their query is unusually long. Search engines try to handle such hard queries by different means. Because they are aware of the risk that every unsatisfied information need increases the fraction of users switching to another search service. The barrier to switching Web search engines is low and multiple engine usage is common. In [10], it is stated that, 70% of Web searchers use multiple search engines. In this thesis, we consider the hardness of a query based on the number of matching results and focus on the queries that can match very few or no results in the web. Queries with large result sets that do not satisfy the users' information need are not in the scope of our study [11].

A particular approach to handle such hard queries is to provide the few available answers for the original user query which is generally fewer than 10 results, while suggesting another query. This suggested query that is potentially a more meaningful version of the query would return larger number of results. These results are provided with a notification such as "do/did you mean". A more aggressive strategy is to directly display the results of the suggested query (at least, mixing these results with the results of the original query), when the search system is more confident about its suggestion. As the daily web users, we occasionally encounter such results displayed by the search engines; however, no search engine exposes a detailed analysis of how and when such mechanisms are applied, and no previous work in the literature discusses these issues in a real and large-scale web search setting.

The first contribution of this thesis is to identify and analyze a large number of hard queries that originally returns very few results when submitted to one of the major search engines. We use the AOL query log [9] to create our own set. Since hard queries are very likely to include spelling errors, search engines typically accompany the original results with some alternative query suggestions or even directly blend the original query result with those of the suggested queries. We will discuss several aspects of these hard queries, retrieved results and suggestions made by the search engines applying both quantitative analyses and user studies on our data (Chapter 4).

Next, we focus on a very specific subset of these hard queries; those that could not be handled even by the above mechanisms and remain unanswered. In this work, we entitle these queries as No Answer Queries (NAQs), and in Chapter 5 we analyze NAQs submitted to a web search engine.

We believe that a characterization about NAQs is important. Because it can fuel the research on solving these queries, eventually leading to improvements on the search quality and user satisfaction. Solving NAQs is a vital issue in today's highly competitive search market, where users frustrated with not finding the requested information may easily switch to another search service. It causes a significant loss in revenues and brand loyalty of a search engine. Indeed, recent studies report that almost half of the users switch between search engines at least once per month [10, 12]. White and Dumais [12] surveyed 488 users regarding their experiences with search engine switching. According to these studies, more than half of the users state the dissatisfaction from search results as the primary reason for switching to another search engine. Hence, characterizing and solving NAQs can provide significant benefits to search engines.

The thesis is organized as follows. In the next chapter of the thesis, we provide a summary of related research about queries from web search engines. In Chapter 3, we present the background information that is essential to understand the concepts discussed in the following chapters. The experimental setup and our analyses about hard queries are provided in Chapter 4. We analyze the No Answer Queries (NAQs) in Chapter 5. In Chapter 6, we present our method for the prediction of query suggestion patterns and NAQs. Chapter 7 states the conclusions drawn from our work and suggests possible directions for future research.

Chapter 2

Related Work

In this chapter, we review previous works that focus on characterizing and classifying queries and then turn our attention to query reformulation. Our discussion of query reformulation covers some works related to query suggestion, query recommendation and spell correction, as they all serve as the means of interacting with the user who is reformulating a query, especially when there is no useful answer or no answer at all. Finally, we discuss some studies that attempt to characterize and solve difficult queries that lead to low user satisfaction, and long queries.

2.1 Query Classification

To the best of our knowledge, there exist very few works that has focused on characterizing or classifying hard queries. However, the literature involves an extensive list of works related to query classification, essentially with the goal of improving retrieval performance. In some studies, queries are classified based on a list of topics or categories [13, 14, 15, 16] or based on user search goals [17]. In web query classification, queries are labeled with a set of topics using a variety of approaches. Bailey et al. [13] categorize these approaches into three groups that are based on the type of information exploited for classification.

In web based method [16, 18], results of queries are retrieved and classified to determine the query topic. Interaction based methods use the click through information. Term matching methods exploit simple lookup of query terms from a set of manually classified queries. The n-gram based matching approach is proposed in [15] because exact matching of query terms is limited to some content.

In particular, two works [13, 16] attempt to classify long and rare web queries. The former [16] exploits the retrieved query results for classification and aims to improve the advertisement selection for rare queries. The latter [13] classifies rare queries by matching them against previously seen classified queries. Furthermore, Downey et al. [19] investigate the characteristics of rare and common queries. It appears that users click fewer results and make more query reformulations for rare queries.

2.2 Query Reformulation

Users reformulate their queries when they are not satisfied with the query results. As mentioned before, out of 2 billion daily web searches made by Internet users [8], approximately 28% are modifications to the previous query [9], which is known as query reformulation or query refinement. In another study, an analysis of a large query log reveals that almost half of the users (46%) reformulate their queries [20]. The reformulation can be performed in several ways. The user might replace one or more terms in the query with others, generalize the query with removing a term or specialize by adding more terms. Huang and Efthimiadis [21] present their taxonomy of user reformulation types and propose a rule based classifier.

Some works on query reformulation focus on offering automatically generated query suggestions to the user who makes the searches on web. Search engines generally show the suggestions on the same page with the search results. Computer-generated suggestions are made by using query substitution [22], query expansion [23] and other refinement techniques [24]. Computer-generated reformulations use implicit relevance feedback from users as a common data source.

Baeza-Yates et al. [25] use query logs to discover new query reformulations, finding similar queries using a cosine function over a term-weighted vector built from the clicked documents.

Spelling correction may be considered as a form of query reformulation. Cucerzan and Brill [7] show that around 10% - 15% of search queries contain spelling errors. The spelling correction is more difficult on the web because of the diversity of terms. Special solutions are required to correct the queries [7, 26]. Our study also confirms this observation that, the existence on NAQs with spelling errors implies that the current spell correction techniques could not adequately handle misspelled queries.

Mei et al. [27] exploit the click-through information in order to suggest semantically related queries. Queries are clustered based on the similarity of clicked documents' content in an alternative approach [5]. Then, a new query is assigned to the most similar query cluster and queries are ranked and suggested to the user in that cluster. White and Marchionini [28] show the effectiveness of a real-time term recommendation system that suggests terms for query expansion while the user enters the query. For query suggestion, Jones et al. [22] investigate word substitutions using query logs. Similarly, Wang and Zhai [29] mine query logs for term associations and propose word-substitution-based query reformulation. In a recent study [30], the anchor text is exploited for query reformulation and several techniques are evaluated using the standard TREC collections.

2.3 Query Suggestion

Query suggestion has been a well-accepted utility used by many search engines to help user explore and express their information need. Many query suggestion approaches have been proposed to address query ambiguity problem in the information retrieval community in recent years. Finding keyword suggestions from the documents retrieved by initial query is a commonly adopted solution. For example, Lam et al. [31] and Xu et al. [32] extract keywords from the top-ranked documents that are regarded as the relevant results of initial query. Daumé and Brill [33] also extract suggestions based on document clusters that have common top-ranked documents. In [25], a query recommendation method based on clickthrough data is proposed.

Wen et al. [34] cluster similar queries by recommending URLs to frequently asked queries of a search engine. Four notions of query distance are used:

- Query distance that is based on keywords or phrases of the query
- Query distance that is based on string matching of keywords
- Query distance that is based on common clicked URLs
- Query distance that is based on the distance of the clicked documents in some pre-defined hierarchy

A method is presented by Fonseca et al. [35] to discover related queries based on association rules. Here queries represent items in traditional association rules. The query log is viewed as a set of transactions, where each transaction represents a session. In a session, a single user submits a sequence of related queries in a time interval. Query expansion is another approach to suggest related queries adopted by search engines [36]. The idea here is to reformulate the query such that it gets closer to the term-weight vector space of the documents that the user is looking for.

Some other works tend to find similar queries from click-through data, which is usually represented as a bipartite graph. Bipartite graph has vertices on one side corresponding to queries and on the other side to clicked URLs. Groups of similar quires, which share a large number of clicked URLs, are obtained through a clustering process over the click-through data. The similar queries are then used as suggestions for each other. For instance, Wen et al. [37] use a density-based clustering method that exploits both query content and click-through information. Beeferman and Berger [38] propose an approach that is represented as a bipartite graph, and applied an agglomerative clustering technique to identify related queries.

2.4 Difficult and Long Queries

The average number of query terms is found to be between 2.35 and 2.60 terms per query based on search logs [39, 40]. In another study, Kamvar et al. [41] report a slightly higher number, 2.93 terms per query. Most of these queries are simple keyword queries. Jansen et al. [42] report that only about 10% of queries contain advanced query operators. On the other hand, there are large regional differences in the use of advanced operators [43]. An analysis by Eastman and Jansen suggested that most of the query operators do not increase the precision of the query, for this reason, they might not be worth the trouble [44]. There has been an increasing interest for understanding and predicting difficult queries. Carmel et al. [11] analyze the reasons for query difficulty. They point out that if the distance of queries and relevant documents from the entire collection is not sufficiently large, then these queries are more difficult to solve. It is shown that the user click behavior is correlated with the query length in that users click lower ranked results more often for long queries compared to the short ones [45]. There is a recent interest in customizing the search for long queries. Kumaran and Carvalho [46] propose to remove extraneous terms in long queries by finding the best subquery that is predicted by the query quality prediction methods. The long query reduction problem is addressed in the context of web search in a recent study [47]. Huston and Croft also focus on verbose queries and report that simply reducing the length of a query by learning and removing stop structures can improve the retrieval performance [48].

Chapter 3

Background

In this chapter, we provide the fundamental information that is needed to understand the concepts discussed in the following chapters. This background information includes the architecture of web search engines, general information about hard queries including No Answer Queries, and suggestion mechanism with patterns that are provided by search engines.

3.1 General Architecture of a Web Search Engine

Search engines are special sites on the Web that are designed to help people to find information stored on other sites. There are some differences in the ways various search engines work, but they all perform three basic tasks:

- They search the internet based on important words.
- They keep an index of the words they find, and where they find them.
- They allow users to look for words or combinations of words found in that index.

At this point it would be beneficial to look how web search engines work. A web search engine is designed to search for information on the World Wide Web. The search results are generally presented in a list of results.

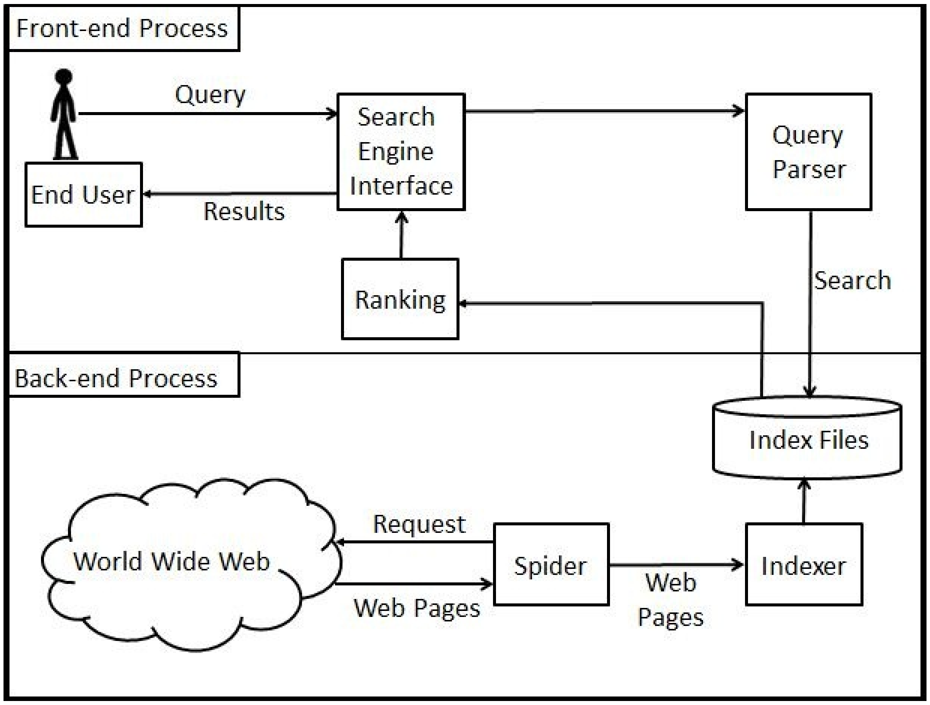

Figure 3.1: The architecture of a web search engine

As shown in Figure 3.1, the architecture of a Web search engine contains a back-end process and a front-end process. In the front-end process, the user enters a query into the search engine interface. The query is parsed into a form that the search engine can understand, and then the engine examines its index. After that it provides a listing of best-matching web pages according to its ranking criteria, usually with a short summary containing the document's title and sometimes parts of the text. In the back-end process, a spider crawls the Web pages from the Internet and the indexer parses the Web pages and stores them into the index files. It is obvious that the three main components of a web search engine are *crawler* that downloads web content continuously, *indexer* that indexes downloaded documents, and *query processor* that submits the queries to the users.

Web searching has become a daily behavior and search engines are used as the main entry point to the web by nearly 70% of the users [17]. We use the results obtained from three different search engines, Bing [49], Google [50] and Yahoo! [51]. For better understanding of the concepts, some related search engine definitions are given in the following:

- **Query**: A query is a form of questioning to obtain information from search engines. A query can consist of an individual word or a combination of more than one word. For instance, "bilkent" is a query and similarly "bilkent university computer engineering" is a query too. Boolean operators can be used to create complex queries. For instance 'AND' operator is used to join all query terms. All the terms joined by it must appear in the pages or documents. Some search engines substitute the operator '+' for the word 'AND'. Similarly, the term or terms following 'NOT' must not appear in the pages or documents. Some search engines substitute the operator '-' for the word 'NOT'.

- **Term**: Each word in a query is called a term. For instance in query "bilkent university computer engineering", there are four terms which are "bilkent", "university", "computer" and "engineering".

- **Result page**: Result page represents the provided results by search engines to the users. In our study it means top-10 results for the query.

- **Result count**: Result count is the number of results returned for the query. For instance Google provides 104000 results for the query "bilkent university computer engineering", while Yahoo! has 255000 and Bing has 222000 results for the same query.

- **Domain**: Web address of a search result is called URL. The main part of the URL is called domain. For instance "http://www.cs.bilkent.edu.tr/∼esarigil/" is a URL and "cs.bilkent.edu.tr" is the domain part of the URL.

3.2 Pattern

Web contains billions of pages and sometimes it is really hard to formulate proper query. Especially when user is looking for information on a topic that he/she does not know too much about. At this point, search engines try to help users to satisfy them. Search engines generally offer search suggestions based upon the words that a user types. Result patterns adopted by search engines can be categorized into four basic types:

- **Pattern 0**: Search engines try to provide direct answers for the submitted query by users. In this type, they do not suggest any alternative query because there are enough result pages for the submitted query. For instance if the query "google" is searched, search engines provide direct result for the query.
- **Pattern 1**: For some queries, search engines return query suggestions but provide results for the original query. For this kind of queries, search engines behave that query might contain typo so they offer alternative queries. For instance, if the query "gogle" is searched, the search engine warns the user with the "do/did you mean" tag and suggests the "google" query. On the other hand, it provides the results for the original query. Because there are results related to "gogle" query, e.g. 'http://www.gogle.es/'. Nevertheless "google" query is more common and popular, so the user is warned to search for it.
- **Pattern 2**: Search engines provide some suggestion and results are related with the suggested query instead of the original query. For instance, for the query "googgle", the search engine warns the user with the "showing result for google" and provides the results for the suggested query. For this kind of queries, search engines are more confident with their suggestions than Pattern 1 and show the results for suggestions. On the other hand, they still have an option to search the original query on the result page. If the user insists to search his/her own query, he/she can select this option.

- **Pattern 3**: In this pattern, no results match with the query. In addition to that, search engines cannot make any suggestion for the query. For instance, for the query "+goglglg", search engines cannot provide any results and cannot suggest any related query.

3.3 Hard Queries

In this study, we consider the hardness of a query based on the number of matching results and focus on queries that can match very few or no results. These queries are generally very likely to include spelling errors and search engines typically accompany the original results with alternative query suggestions. Our definition of hard queries is that the queries that can return very few or no results.

3.4 No Answer Queries

Search engines cannot provide any results for some queries and we name such queries as No Answer Queries (NAQs). We consider NAQs as a subset of hard queries. Especially if a content in a less common language or unpopular website is searched by users, search engines more likely to return no answer. For instance, for the queries "hkl9oi8joo-80yii';p", "hack all1010100100101010101010100.com" and "lkjhghjkkjhgghjkkjhgghjkkjhghjkjhg-ghjkkjhggh jkkjhgghjkkjhgkjhgkhgghj", search engines cannot provide any result or suggestion.

Chapter 4

Analysis of Hard Queries

Search engines take text queries that users type as input, and present users with information of ranked web pages related to users' queries. Although the web contains huge volumes of data; search engines generally present the most relevant results in less than a second when a user enters a query. As we mentioned in Chapter 3, sometimes search engines provide few answers for the user queries which are hard to relate web documents. There are different mechanisms, such as providing results for alternative queries, for such hard queries. In this chapter, we analyze these types of queries and present the result of our experimental study about them.

4.1 Experimental Setup

4.1.1 Dataset

Web search engines record information about the queries that people write, forming what is called a query log. Since a large search engine receives hundreds of millions of queries from millions of users per day, query logs constitute an invaluable resource for understanding the kinds of needs that people have.

In our work, we use the AOL query log [9] which we use to identify and characterize hard queries. Despite the fact that the AOL query log was released in 2006, many recent studies still use this log since it is the largest and most recent publicly available query log. We prefer to use this log for characterization of hard queries since it would enable the other researchers to reproduce our findings. Furthermore, it is very unlikely that a search engine could publish its query log which contains very few or no answer queries because this might expose confidential information about the search engine and it also could show its weaknesses. We use the AOL query log to prevent a potential search engine bias.

For the purposes of this study, we do not use the No Answer Queries (NAQs) in the AOL query log due to the following reasons. First, in this log, result URLs for a query are included only if they are clicked by the users. This means that, it is impossible to distinguish queries without any answer from the queries that return some answers but none of them are clicked. Second, many queries that could not be answered at the time of their original submission may now find answers in the current search engines. Because, the World Wide Web is continuously growing [52]. According to Google, on the average, more than a billion new pages are added to Internet every day [6]. In addition to that, more advanced mechanisms have been adopted by the search providers since the release of this query log.

4.1.2 Dataset Setup

Apparently, it is not possible to retrieve the results of all unique queries in the AOL log because of the limits of search engines. Also, most queries would match a large number of queries which are useless for us. Consequently, we adopt a two-step procedure to identify the hard queries in the AOL log.

In the first step, our goal is to designate a candidate set of hard queries which return very few results or no result when submitted to search engines. Earlier works in the literature suggest that search engine APIs process queries over an index that seems to be smaller than the full web index [53]. So we believe that identifying queries which return no answers from a search engine API could be our starting point. We use a dataset (similar to [54]), where 660K unique AOL queries were issued to the Yahoo! web search API in December 2010. We choose queries which return no answers (around 16K queries) and re-submit them to the API in early July 2011. We distinguish that the number of queries with no answers drops to 11K queries.

As our hard queries are seeded with those that do not retrieve any results from the Yahoo! web search API, we might have a slight bias towards those queries that cannot be resolved by Yahoo! web search API. We randomly selected 6000 singleton queries from the AOL log that are not in our initial 16K queries to investigate this issue. We chose singleton queries from the AOL log because non-singleton queries are most likely to be resolved by all three search engines. We issued these 6K queries to the Google, Yahoo! and Bing, and retrieved the first result pages similar to our 11K candidate hard queries. When we analyze the results, the percentage of NAQs is very similar, as shown in Table 4.1. On the other hand, as expected, the absolute numbers are much smaller. So we believe that the way we create our query set does not introduce a significant bias against any search engine.

In the second step, these 11K candidate hard queries are sent to the three major search engines which are Bing, Google and Yahoo! and the first result pages are retrieved (similar to [53]). We make sure that, for all three search engines, we submit the queries to the U.S. frontends that have the largest index. All non-default search preferences are disabled because of reaching the greatest extent. For Google and Yahoo!, the main search frontend is selected which contains no region extend. On the other hand, United State region (English) is selected for Bing because it is verified that international option have smaller index than this configuration. We collect data two times; first one is in July 2011 and the next one is January 2012. Same query set that is 11K candidate hard queries is used.

4.1.3 Query Set

In this dataset, there are 11673 queries that are identified as hard queries. These 11K queries include 40482 words. Because of the combined terms like ‘3rdgenerationgospelsingers’, the number of words are less than expected. The average number of words for each query is 3.47. This means that each query on the average contains 3-4 words. 77% of the queries contain one keyword, 9% of the queries contain two keywords and 5% of the queries contain three keywords. Further, approximately 91% of the queries contain no more than three keywords. When we look at the number of the characters for our dataset, there are 331291 characters. So the average length of all the queries is 28.4. 73% of the queries are labeled as URI which means resource locator. This means that users try to reach a certain web page. We control all of the 11K queries with a simple code which controls if the query contains some particular markers, such as, ‘http’, ‘www’, ‘.com’, ‘.org’, ‘.info’ etc.

4.2 Experimental Study

4.2.1 Hard Queries with Few Results

In Table 4.1 and Table 4.2, the number of queries that return k or fewer results are reported for the three search engines. The important point here is that, we only consider the number of results retrieved from the original query. For some queries search engines provide query suggestions and their results. In this part we do not take the suggestions into consideration.

Table 4.1: Number of queries that returned k or fewer results for each search engine (only original query results are considered) - July 2011

k	Google	Yahoo	Bing
0	244 (2%)	1791 (15%)	1997 (17%)
≤ 2	1129 (10%)	6368 (55%)	6377 (55%)
≤ 10	3829 (33%)	7366 (63%)	7721 (66%)
≤ 100	7394 (63%)	8960 (73%)	9089 (78%)

Table 4.1 presents the statistics for the retrieved query results collected from the search engines in July 2011. It can be seen from the table that a high fraction of 11K queries that are submitted to search engines return very few or even no results. For the search engines Google, Yahoo! and Bing less than 10 results are returned for 33%, 63% and 66% of the queries, respectively. This can be considered as a remarkably tiny result set since web includes billions of pages. Furthermore, 2% to 17% of these hard queries seem to be actual NAQs.

Table 4.2: Number of queries that returned k or fewer results for each search engine (only original query results are considered) - January 2012

k	Google	Yahoo	Bing
0	106 (1%)	2858 (24%)	3240 (28%)
≤ 2	503 (5%)	6847 (59%)	7316 (63%)
≤ 10	5037 (43%)	8298 (71%)	8554 (73%)
≤ 100	6392 (55%)	9463 (81%)	9384 (80%)

As shown in Table 4.2 similar results were obtained for the queries submitted in January 2012. However this time 1% to 28% of hard queries seem to be actual NAQs. Despite the fact that Google makes some progress and improve its results, the number of hard queries increases for the other two search engines Bing and Yahoo!. The number of NAQs increased by 9% of queries for both engines.

4.2.2 Query Correction for Hard Queries

Table 4.3: Message patterns observed in search engine result pages

Pattern	Search Engine	Message displayed in the search engine result page
0	All	-
1	Bing	Do you mean <suggested query>
	Google	Did you mean: <suggested query>
	Yahoo	Did you mean: <suggested query>
2	Bing	No results found for <original query>. Showing results for <suggested query>.
	Google	Showing results for <suggested query>. Search instead for <original query>.
	Yahoo	We have included <suggested query> results. Show only <original query>.
3	Bing	No results found for <original query>.
	Google	No results found for <original query>.
	Yahoo	We did not find results for: <original query>.

We analyze the retrieved result pages and extract four types of result patterns that are adopted by all three search engines. These result patterns are shown in Table 4.3. In the first pattern, Pattern 0, the answer of the submitted query is shown in the result page. There is not any suggested query in this pattern. As the second pattern, Pattern 1, all three search engines return query suggestions. However the results of the original query are shown. In the third pattern, Pattern 2, all three search engines provide some suggestions/corrections and results are related with the suggested query instead of the original query. Finally, we observe a fourth pattern, Pattern 3, when no results match the query and no suggestion is provided for the original query.

Table 4.4: Number of queries with a certain pattern, observed at each search engine - July 2011

Pattern No - July 2011				
SE	Pattern 0	Pattern 1	Pattern 2	Pattern 3
Google	7267 (62%)	1277 (11%)	2896 (25%)	233 (2%)
Yahoo	2771 (24%)	5340 (46%)	3101 (27%)	461 (4%)
Bing	3519 (30%)	4584 (39%)	3068 (26%)	502 (4%)

Table 4.5: Number of queries with a certain pattern, observed at each search engine - January 2012

Pattern No - January 2012				
SE	Pattern 0	Pattern 1	Pattern 2	Pattern 3
Google	7236 (62%)	1521 (13%)	2813 (24%)	103 (1%)
Yahoo	2411 (21%)	4116 (35%)	4625 (40%)	521 (5%)
Bing	2738 (23%)	3840 (33%)	4313 (37%)	782 (7%)

All three search engines attempt to correct the query terms for most of the hard queries. Search engines handle this issue by either providing query suggestions (Pattern 1) or directly providing the suggested query's results (Pattern 2). In Table 4.4, the numbers of query results falling under each pattern are reported for all three search engines in July 2011. Google provides immediate answers to the majority of hard queries (around 62%) and do not need to provide any suggestions. On the other hand, Bing and Yahoo! try to handle the majority of these queries via Pattern 1 (39% and 46% respectively). Pattern 2 results are very close to each other (25%, 27% and 26%) among all three search engines. Furthermore, 2% to 4% of these hard queries turn out to be actual NAQs. Google provides more results than Bing and Yahoo!. We observe that the number of NAQs for Google is around half of those for Bing and Yahoo!.

In Table 4.5 the numbers of query results falling under each pattern are reported for the three search engines in January 2012. Google still provides immediate answers to the majority of hard queries (around 62%). Despite the fact that in July 2011, Pattern 2 results are close to each other, in January 2012 Pattern 2 results are different from each other. Google provides similar numbers of results but the number of queries with Pattern 2 for Bing and Yahoo! increases (27% to 40 and 26% to 37% respectively). The number of actual NAQs changes in January 2012. The number of NAQs goes down to half for Google (2% to 1%). However, the number of NAQs increases for Bing and Yahoo! (4% to 5% and 4% to 7% respectively).

We observe that all three search engines have similar fraction of queries that result in Pattern 2. A remarkable point is that, search engines Bing and Yahoo! can reduce their NAQs by using Pattern 2. A quick comparison between Table 4.1 → Table 4.4, and between Table 4.2 → Table 4.5 reveal that using Pattern 2 helped several queries in Bing and Yahoo! that originally return no answers.

4.2.3 Pattern Change for Hard Queries

According to Table 4.4 and Table 4.5, some queries are labeled with different patterns in July 2011 and January 2012. For instance, in July 2011, 11% of the queries are labeled as Pattern 1 by Google, but this number increases to 13% in January 2012. So we analyze queries to identify the changes between patterns.

Table 4.6: Number of queries with pattern change between July 2011 and January 2012 for Google

Google		January 2012			
		Pat 0	Pat 1	Pat 2	Pat 3
July 2011	Pat 0	6786 (93%)	270 (4%)	172 (2%)	39 (1%)
	Pat 1	172 (13%)	856 (67%)	244 (19%)	5 (0%)
	Pat 2	136 (5%)	366 (13%)	2393 (83%)	1 (0%)
	Pat 3	142 (61%)	29 (12%)	4 (2%)	58 (25%)

Table 4.6 shows the number of queries with pattern change about queries between July 2011 and January 2012 for Google. According to this, 93% of the queries that are labeled as Pattern 0 by Google in July 2011 are still labeled as Pattern 0. Interestingly, Google provides some results for 1% of the queries in July 2011 but it cannot provide any results for them in January 2012. This can be explained by that related webpages might have been removed from internet. 67% of the queries that are labeled as Pattern 1 in July 2011 are still labeled as Pattern 1 and 83% of the queries that are labeled as Pattern 2 in July 2011 are still labeled as Pattern 2. Google makes some improvements for NAQs in January 2012. It can provide direct answers for 73% of the queries that are labeled as Pattern 3 in July 2011 (Note that 61% with Pattern 0 and 12% with Pattern 1). Remember that, search engines provide direct answers for the queries with Pattern 0 and Pattern 1 and provide their suggestions' results for the queries with Pattern 2. Google still cannot provide any answer only for 25% of the queries that are labeled as Pattern 3 in July 2011.

Table 4.7: Number of queries with pattern change between July 2011 and January 2012 for Yahoo!

Yahoo		January 2012			
		Pat 0	Pat 1	Pat 2	Pat 3
July 2011	Pat 0	1348 (49%)	637 (23%)	656 (24%)	130 (5%)
	Pat 1	771 (14%)	3313 (62%)	1195 (22%)	61 (1%)
	Pat 2	219 (7%)	140 (5%)	2663 (86%)	79 (3%)
	Pat 3	73 (16%)	26 (6%)	111 (24%)	251 (54%)

In Table 4.7 the number of queries with pattern change about queries between July 2011 and January 2012 are listed for Yahoo!. 51% of the queries that are labeled as Pattern 0 in July 2011 are labeled with different patterns in January 2012. Similar to Google, Yahoo! cannot provide any results for 5% of the queries that are labeled as Pattern 0 in July 2011. 62% of the queries that are labeled as Pattern 1 in July 2011 are still labeled as Pattern 1 and 86% of the queries that are labeled as Pattern 2 in July 2011 are still labeled as Pattern 2. In January 2012 Yahoo! provides some answers for 46% of the queries that are NAQs in July

2011.

Table 4.8: Number of queries with pattern change between July 2011 and January 2012 for Bing

Bing		January 2012			
		Pat 0	Pat 1	Pat 2	Pat 3
July 2011	Pat 0	1811 (51%)	641 (18%)	827 (24%)	240 (7%)
	Pat 1	580 (13%)	2994 (65%)	876 (19%)	134 (3%)
	Pat 2	276 (9%)	163 (5%)	2504 (82%)	125 (4%)
	Pat 3	71 (14%)	42 (8%)	106 (21%)	283 (56%)

Table 4.8 shows the number of queries with pattern change about queries between July 2011 and January 2012 for Bing. In January 2012 51% of the queries that are labeled as Pattern 0 by Bing in July 2011 are still labeled as Pattern 0. Interestingly, Bing provides some results for 7% of the queries in July 2011 but it cannot provide any results for them in January 2012. It is a high percentage of queries when compared to Google and Yahoo!. 65% of the queries that are labeled as Pattern 1 in July 2011 are still labeled as Pattern 1 and 82% of the queries that are labeled as Pattern 2 in July 2011 are still labeled as Pattern 2. Bing provides direct answer for 14% of the queries that are labeled as Pattern 2 in July 2011. For 56% of the queries which are NAQs in July 2011 are still labeled as Pattern 3. When we compare all three search engines with the number of pattern change about queries, Google is the most consistent one, especially for queries that are labeled as Pattern 0 in July 2011. In addition to that only Google decreases the number of NAQs in January 2012. Both of Yahoo! and Bing have some problems with handling hard queries when they are compared with Google.

4.2.4 Number of Results

Table 4.9: Number of queries that return k or fewer results for each pattern and search engine (the percentages are computed with respect to all queries with a given pattern and search engine) - July 2011

Number of Results - July 2011				
SE	k	Pattern 0	Pattern 1	Pattern 2
Google	1	277 (4%)	63 (5%)	2 (0%)
	2	267 (4%)	59 (5%)	2 (0%)
	≤ 10	1602 (22%)	435 (34%)	17 (1%)
	≤ 1000	4503 (62%)	751 (59%)	196 (7%)
Yahoo	1	978 (35%)	1019 (19%)	40 (1%)
	2	491 (18%)	586 (11%)	29 (1%)
	≤ 10	1965 (71%)	2608 (49%)	157 (5%)
	≤ 1000	2233 (81%)	3747 (70%)	557 (18%)
Bing	1	1328 (38%)	819 (18%)	52 (2%)
	2	712 (20%)	476 (10%)	44 (1%)
	≤ 10	2785 (79%)	2094 (46%)	189 (6%)
	≤ 1000	3059 (87%)	3099 (68%)	621 (20%)

Table 4.10: Number of queries that return k or fewer results for each pattern and search engine (the percentages are computed with respect to all queries with a given pattern and search engine) - January 2012

Number of Results - January 2012				
SE	k	Pattern 0	Pattern 1	Pattern 2
Google	1	96 (1%)	16 (1%)	1 (0%)
	2	200 (3%)	30 (2%)	0 (0%)
	≤ 10	3012 (42%)	526 (35%)	13 (1%)
	≤ 1000	4414 (61%)	865 (57%)	154 (6%)
Yahoo	1	537 (22%)	1361 (33%)	45 (1%)
	2	202 (8%)	514 (13%)	20 (0%)
	≤ 10	1019 (42%)	2517 (61%)	219 (5%)
	≤ 1000	1494 (62%)	3238 (79%)	690 (15%)
Bing	1	847 (31%)	1179 (31%)	47 (1%)
	2	285 (10%)	457 (12%)	20 (1%)
	≤ 10	1464 (54%)	2236 (58%)	206 (5%)
	≤ 1000	1900 (69%)	2953 (77%)	644 (15%)

We observe that all three search engines can provide some answers to most of the hard queries. It seems worthwhile to analyze the quality of the returned results. We limit our analysis to a comparison of the number of matching results for queries with Pattern 0, 1 and 2 because evaluating 11K query results for all three search engines requires significant human effort. For every pattern and search engine pair, Table 4.9 reports the number of queries which return k or fewer results in July 2011. We observe that queries with Pattern 2 match the largest number of results. On the other hand, queries with Pattern 0 that are directly answers of original queries, match the smallest number of results. For instance, for 79% of queries with Pattern 0 have less than 10 results in Bing but for queries with Pattern 2 this is only 6%. Similarly, Yahoo! returns less than 10 results for 71% of queries with Pattern 0, but only 5% of queries with Pattern 2 return less than 10 results. When we analyze the results for January 2012, the same trend can be seen. Again queries with Pattern 2 match the largest number of results. We can say that search engines are more confident with Pattern 2. This is because, search engines provide results for the suggested queries and these queries are generally more common queries. For example, one of the 11K queries is “tulare outle tmall”. When the original query is searched in Google, only 7 results are returned. On the other hand, when Google searches its own suggestion that is “tulare outlet mall”, it has 122,000 results. In another example, when the original query “www.pueblowio” is searched in Yahoo!, the query has no result. On the other hand, when Yahoo! searches its own suggestion that is “pueblo wio”, it has 120,000 results.

As mentioned above, when Pattern 2 is shown, the search engine is rather confident since the user intention well matches with another query, which can retrieve more results than the original query. In case of Pattern 1, the search engine possibly suggests a better query, but the original query also matches some results. So the results of the original query are revealed. The search engines are either very confident about the results for direct results (Pattern 0) or they simply cannot find a query to recommend and present whatever results the original query matches. For example, in July 2011, Google finds more than 1000 results for 38% of queries which are Pattern 0. Furthermore, Google increases its success rate in January 2012 and finds more than 1000 results for 39% of queries which are Pattern 0. Yahoo! also makes some changes that affect its success rate in a positive way. Yahoo! finds more than 1000 results for 19% of queries with Pattern 0 in July 2011 and finds more than 1000 results for 38% of queries with Pattern 0 in January 2012. The same trend can be seen for Bing. In July 2011 Bing finds more than 1000 results for 13% of queries with Pattern 0, and in January 2012 it finds more than 1000 results for 31% of queries with Pattern 0. As mentioned before, sometimes search engines cannot find a query to recommend and present whatever results the original query matches. For instance, in July 2011, Yahoo! retrieves at most two results for 53% of queries with Pattern 0. On the other hand, Yahoo! improves its results in January 2012. It retrieves at most two results for only 30% of queries with Pattern 0. Similarly, Bing retrieves at most two results for 58% of queries with Pattern 0 in July 2011 and for 41% of queries with Pattern 0 in January 2012. Since they cannot make any suggestion for these queries, only a small number of results are returned.

Table 4.11: Number of queries with Pattern 2 for which the corresponding original query returns k results (the percentages are computed with respect to the corresponding values in Table 4.4) - July 2011

k	Google	Yahoo	Bing
0	11 (0%)	1330 (43%)	1495 (49%)
≤ 2	226 (8%)	2333 (75%)	2540 (83%)
≤ 10	1792 (62%)	2793 (91%)	2842 (93%)
≤ 100	2113 (73%)	2980 (96%)	2931 (96%)

Table 4.12: Number of queries with Pattern 2 for which the corresponding original query returns k results (the percentages are computed with respect to the corresponding values in Table 4.5) - January 2012

k	Google	Yahoo	Bing
0	3 (0%)	2397 (52%)	2458 (57%)
≤ 2	58 (2%)	3710 (80%)	3766 (87%)
≤ 10	1396 (50%)	4241 (92%)	4072 (94%)
≤ 100	1904 (68%)	4491 (97%)	4162 (97%)

For the queries returned with Pattern 2, search engines search for their own suggestions. What happens if the user continues on his/her original queries? To find an answer to this question, we obtained the results of original queries for Pattern 2 which is shown in Table 4.11 for July 2011 and in Table 4.12 for January 2012. Our results show that Bing and Yahoo! have some problems with original queries for Pattern 2. In July 2011, Bing cannot find any results for 49% of queries with Pattern 2 when the original query is searched. Similarly, Yahoo! cannot find any results for 43% of queries with Pattern 2 when the original query is searched. So these queries answered with Pattern 2 are indeed NAQs when the original query is followed. For this reason, presenting the results of the suggested query is mandatory for these queries. In January 2012, percentage of NAQs increases for both Bing and Yahoo!. 52% of the queries with Pattern 2 have no answer when the original query is searched in Yahoo! and 57% of queries with Pattern 2 have no answer when the original query is searched in Bing (Note that, 43% to 52% for Yahoo! and 49% to 57% for Bing). Both Yahoo! and Bing find more than 100 results for only 4% of the queries in July 2011 and 3% of the queries in January 2012 which are Pattern 2. Again both of these search engines return less than 10 results for nearly 92% of queries with Pattern 2. On the other hand, for Google, the situation is different. A significant portion of these queries can retrieve some answers. In July 2011 only 11 (0%) queries and in January 2012 only 3 (0%) queries with Pattern 2 have no answer when the original query is followed. Furthermore, Google finds more than 100 results for 27% of the queries with Pattern 2 in July 2011 and for 32% of the queries with Pattern 2 in January 2012. Anyway, even for Google, which retrieves more results than Bing and Yahoo!, the queries with Pattern 2 match relatively less results than those with Pattern 0 and 1 in both July 2011 and January 2012.

4.2.5 Domain of the Results

For both July 2011 and January 2012, there are more than 250K domains for 11K hard queries in our dataset. Some of the queries have the same domains in two different dates but some do not. In July 2011, there are 68644 unique domains returned as a result of queries. This number increases to 69410 in January 2012. In July 2011, 51% of the unique domains are returned only once. 97% of the unique domains are seen less than 10 times as a result of the queries. In January 2012, however, 46% of the unique domains are returned only once and again 97% of the unique domains are seen less than 10 times as a result of the queries. In Google, all the results for a query are returned from just one domain for 3.6% of the queries in July 2011 and 1.7% of the queries in January 2012. In Yahoo!, these ratios are 21.3% in July 2011 and 18.3% in January 2012 and finally in Bing 23.4% in July 2011 and 20.1% in January 2012. According to these results, we observe that, as time passes the search engines tend to return results from more domains. In Yahoo! and Bing, there are many queries with less than 2 results, so the number of queries that are returned from just one domain is quite high.

Table 4.13: Number of results retrieved from fake result sites for each pattern and search engine - July 2011

Search Engine	Pattern 0	Pattern 1	Pattern 2
Google	38%	19%	1%
Yahoo	13%	5%	1%
Bing	18%	5%	1%

Table 4.11 and Table 4.12 show that the queries with Pattern 2 originally return very few results. When we investigate these queries, we observe that, the web pages returned in the top 10 results simply include a list of the queries in the AOL log for several queries. Of course such a web page cannot be considered as a real answer for the query. While it is impossible to determine all such queries' results manually, we basically listed the domains that appear most frequently in the result of the queries. All of the queries with Pattern 0, Pattern 1 and Pattern 2 are considered. We identify the most frequent five fake domains in July 2011. These five domains seem to be a plain compilation of the AOL queries or URIs in these queries and they are namely, aolscandal.com, aolstalker.com, robtex.com, t35.com and iwant*****.info (sexually explicit part is starred). In Table 4.13, the percentage of answers from these domains in the first result page of the queries with Pattern 0, Pattern 1 and Pattern 2 are presented for each search engine. Note that, the first result page can contain up to the top 10 results for each query. Apparently, a considerable fraction of results for Pattern 0 and Pattern 1 come from a few domains. We observe that, the number of real results for these queries is even smaller than those retrieved from search engines. This observation confirms that our process for identification of hard queries successfully yields those queries that really return few results on the web.

Table 4.14: Number of the most frequent domains - July 2011

aolscandal.com	5729	domains5.cn	980
en.wikipedia.org	5103	answers.yahoo.com	970
aolstalker.com	3973	responsib.hotbox.ru	933
youtube.com	2661	duieigm.t35.com	919
iwant*****.info	2577	gcugeeu.t35.com	864
facebook.com	2485	yelp.com	784
robtex.com	1726	linkedin.com	751
myspace.com	1676	dpuoucu.t35.com	743
membres.multimania.fr	1573	bdiercf.t35.com	719
manta.com	1104	superpages.com	716
local.yahoo.com	1065	jnetxni.t35.com	679
mitglied.multimania.de	1040	blneoda.t35.com	650
piettes.com	1026	bscvsji.t35.com	638
faceconrol.site40.net	1010	aifujpm.t35.com	624
amazon.com	1008	bruyyaq.t35.com	618

Table 4.14 and Table 4.15 show the most frequent domains for the web pages returned in the top 10 results with Pattern 0, Pattern 1 and Pattern 2. In Table 4.14, the most frequent domains in July 2011 are listed. As mentioned above, some of these domains are the list of the queries in the AOL log. Also some of these domains are free hosting websites and they exist no longer. For example: 'membres.multimania.fr', 'mitglied.multimania.de', 'faceconrol.site40.net', '***.t35.com' etc. More than 60% of the most frequent 30 domains are the queries in the AOL log and they cannot be considered as a real answer for the queries. On the other hand some domains, like 'en.wikipedia.org', 'youtube.com' and 'facebook.com' are the most common web sites and they are one of the most frequent domains as expected.

Table 4.15: Number of the most frequent domains - January 2012

aolstalker.com	9452	amazon.com	1029
en.wikipedia.org	5881	local.yahoo.com	1003
search-logs.com	3889	ehow.com	984
youtube.com	3298	superpages.com	919
facebook.com	3155	linkedin.com	874
membres.multimania.fr	2022	engrus.com	829
aolscandal.com	1869	answers.yahoo.com	804
robtex.com	1800	yellowpages.com	793
manta.com	1641	ebay.com	767
myspace.com	1620	zapodlo123.net84.net	751
doryoku.org	1587	harlampiyaiefi.narod.ru	716
decenttools.com	1459	domain-history.info	665
mitglied.multimania.de	1414	avtonomeacgolik.narod.ru	652
yelp.com	1037	izotqelbrovko.narod.ru	652
piettes.com	1030	aboutus.org	649

Table 4.15 shows that some domains that are the list of the queries in the AOL log do no longer exist in January 2012. Especially '***.t35.com' domain no longer appeared as an answer of any queries. Some AOL log queries still exist as result of queries but their numbers have decreased. Because of that, 'aolstalker.com' domain is returned more times than before. For instance, 'aolstalker.com' domain appears 3973 times in July 2011, but 9452 times in January 2012. The number of 'aolscandal.com' domain looks like dropping dramatically but in reality this domain is redirected to 'search-logs.com' site. So, both of them appear 5758 times in January 2012 (note that the 'aolscandal.com' domain appears 5729 times in July 2011).

Table 4.16: Number of the most frequent domains for each search engine - July 2011

Google		Yahoo		Bing	
aolscandal.com	3758	en.wikipedia.org	2013	en.wikipedia.org	2098
aolstalker.com	2673	iwant*****.info	1240	iwant*****.info	1337
youtube.com	1384	facebook.com	846	aolscandal.com	1155
en.wikipedia.org	992	aolscandal.com	816	membres.multimania.fr	805
facebook.com	932	membres.multimania.fr	753	facebook.com	707
responsib.hotbox.ru	896	myspace.com	726	aolstalker.com	664
duieigm.t35.com	880	youtube.com	655	youtube.com	622
gcugeeu.t35.com	864	aolstalker.com	636	myspace.com	618
dpuoucu.t35.com	743	mitglied.multimania.de	546	robtex.com	527
bdiercf.t35.com	719	robtex.com	542	faceconrol.site40.net	506
jnetxni.t35.com	679	faceconrol.site40.net	466	mitglied.multimania.de	490
robtex.com	657	local.yahoo.com	462	manta.com	447
blneoda.t35.com	650	manta.com	458	domains5.cn	429
bscvsji.t35.com	638	answers.yahoo.com	403	local.yahoo.com	424
aifujpm.t35.com	624	domains5.cn	403	answers.yahoo.com	389
bruyyaq.t35.com	618	amazon.com	381	amazon.com	349
eazosli.t35.com	615	superpages.com	355	piettes.com	326
hristoforocecelo.narod.ru	585	linkedin.com	346	linkedin.com	287
hk5.com	581	pageinsider.com	329	superpages.com	282
jafiset.t35.com	509	piettes.com	300	ehow.com	257
orkuoal.t35.com	476	yelp.com	263	usuarios.multimania.es	254
zhulidovaarycos.hotmail.ru	464	ehow.com	258	yelp.com	251
apaauus.t35.com	459	yellowpages.com	253	yellowpages.com	217
poilihn.t35.com	451	aboutus.org	253	aboutus.org	213
njveiqp.t35.com	447	usuarios.multimania.es	247	shop.ebay.com	197
decenttools.com	419	shop.ebay.com	226	imdb.com	183
markosweb.com	414	merchantcircle.com	214	pageinsider.com	182
snriugk.t35.com	405	imdb.com	198	beacuda.t35.com	151
piettes.com	400	alexa.com	190	alexa.com	151
hobapsg.t35.com	398	twitter.com	160	merchantcircle.com	135

Table 4.17: Number of the most frequent domains for each search engine - January 2012

Google		Yahoo		Bing	
aolstalker.com	8593	en.wikipedia.org	2370	en.wikipedia.org	2320
search-logs.com	2812	facebook.com	1217	membres.multimania.fr	1020
aolscandal.com	1867	membres.multimania.fr	993	facebook.com	985
doryoku.org	1581	youtube.com	882	youtube.com	845
youtube.com	1571	manta.com	745	manta.com	676
robtex.com	1374	myspace.com	677	mitglied.multimania.de	675
decenttools.com	1267	mitglied.multimania.de	660	myspace.com	587
en.wikipedia.org	1191	search-logs.com	556	search-logs.com	521
facebook.com	953	ehow.com	431	ehow.com	439
engrus.com	829	local.yahoo.com	429	aolstalker.com	435
harlampiyaiefi.narod.ru	716	superpages.com	426	local.yahoo.com	406
domain-history.info	665	aolstalker.com	424	superpages.com	391
avtonomeacgolik.narod.ru	652	yelp.com	393	yelp.com	373
hristoforocecelo.narod.ru	610	amazon.com	386	zapodlo123.net84.net	353
qilefi.hotmail.ru	607	linkedin.com	374	amazon.com	351
izotqelbrovko.narod.ru	596	ebay.com	372	ebay.com	328
zhulidovaarycos.hotmail.ru	589	zapodlo123.net84.net	341	usuarios.multimania.es	309
piettes.com	455	answers.yahoo.com	334	linkedin.com	307
keyword-selector-tool.com	448	yellowpages.com	326	yellowpages.com	305
jiualdakova.hotmail.ru	439	usuarios.multimania.es	305	piettes.com	299
markosweb.com	415	piettes.com	276	answers.yahoo.com	293
refunded.solo10.com	380	answers.com	219	robtex.com	212
xmarks.com	359	robtex.com	214	answers.com	209
myspace.com	356	aboutus.org	212	aboutus.org	200
whois.domaintools.com	349	imdb.com	206	imdb.com	190
empiritag.com	305	merchantcircle.com	200	dictionary.reference.com	173
hk5.com	298	nextag.com	199	merchantcircle.com	163
amazon.com	292	dictionary.reference.com	168	nextag.com	153
pageinsider.com	289	alexa.com	159	alexa.com	144
yelp.com	271	spoke.com	154	spoke.com	127

We also investigate the number of the most frequent domains for all search engines in July 2011 and in January 2012. We think that it is better to examine each search engine separately. When we look at Table 4.16, it provides the number of the most frequent domains for each search engine in July 2011. Google uses '***.t35.com' domain so much that, 50% of the most frequent 30 domains includes this domain. In addition to that, Google has other domains which are from list of the queries in the AOL log. Nearly 75% of the most frequent 30 domains are the queries in the AOL log and they cannot be considered as a real answer for the queries in Google. This ratio is higher than Yahoo! and Bing. 23% of the most frequent 30 domains are common for all three search engines. We think that, Yahoo! and Bing have similar strategies for returning answers since 97% of the most frequent 30 domains are common for both of them. In addition to that these domains are returned with very close numbers. For instance, 'en.wikipedia.org' domain occurs 2013 times in Yahoo! and 2098 times in Bing. Similarly, 'youtube.com' domain occurs 655 times in Yahoo! and 622 times in Bing. Table 4.17 shows the number of the most frequent domains for each search engine in January 2012. Still a high percentage of the most frequent 30 domains are the queries in the AOL log and they cannot be considered as a real answer for the queries in Google. Here 33% of the most frequent 30 domains are common for all three search engines. The ratio increases since '***.t35.com' domain no longer exists. In addition to that, because of no existence of '***.t35.com' domain, 'aolstalker.com' domain reaches a huge number of occurrence. In July 2011, 'aolstalker.com' domain occurs 2673 times and 8593 times in January 2012. All of the most frequent 30 domains are common for Yahoo! and Bing in January 2012. It shows that, they have adopted very similar strategies.

Table 4.18: Number of the most frequent country extensions

July 2011		January 2012	
no extension	228974	no extension	232739
ru	4079	ru	6157
us	2748	uk	3113
uk	2713	us	2894
fr	1707	fr	2168
de	1653	de	2016

Table 4.19: Number of the most frequent country extensions for each search engine - July 2011

Google		Yahoo		Bing	
no extension	91474	no extension	70278	no extension	67222
ru	3683	us	973	us	902
uk	1281	fr	792	fr	840
us	873	uk	714	uk	718
au	611	de	709	de	620
cn	519	cn	445	cn	476

Table 4.20: Number of the most frequent country extensions for each search engine - January 2012

Google		Yahoo		Bing	
no extension	86141	no extension	75509	no extension	71089
ru	5566	us	1045	fr	1056
uk	1493	fr	1038	us	996
us	853	uk	814	de	806
au	641	de	793	uk	806
ca	437	ca	349	es	347

After investigating the number of the most frequent domains for all search engines, we consider the country extensions for the domains for all queries. For example, ‘membres.multimania.fr’ domain has a country extension as ‘fr’, similarly ‘usuarios.multimania.es’ domain has a country extension as ‘es’. On the other hand, domains like ‘aolscandal.com’, ‘en.wikipedia.org’ and ‘songmeanings.net’ do not have any country extension. Table 4.18 shows that the most frequent country extensions for the web pages returned in the top 10 results with Pattern 0, Pattern 1 and Pattern 2. The most frequent country extension for hard queries is ‘ru’ (Russian web sites) in both July 2011 and January 2012. We need to remind that, in July 2011 228974 results, and in January 2012 232739 results do not have any country extension. The results are as expected since most technological countries have more results than others. Table 4.19 provides the number of the most frequent country extensions for each search engine in July 2011. Interestingly most of the Russian web sites are returned as an answer of queries by Google. Despite the fact that both of Bing and Yahoo! do not return domains with Russian websites, because of Google results ‘ru’ is the most frequent country extension. Table 4.20 shows the number of the most frequent country extensions for each search engine in January 2012. Again Russian websites have a great impact in Google. In January 2012, Yahoo! and Bing results are similar with July 2011 results.

4.2.6 Overlap Between Search Engines

Table 4.21: Number of overlapping queries with each pattern for each search engine (the percentages are computed with the number of corresponding columns' union) - July 2011

Overlapping Queries July 2011				
Search Engine	Pattern 0	Pattern 1	Pattern 2	Pattern 3
Google∩Yahoo∩Bing	1637 (18.86%)	273 (4.11%)	1186 (23.98%)	114 (17.14%)
(Bing∩Yahoo)\Google	754 (8.68%)	3840 (57.87%)	1425 (28.81%)	287 (43.16%)
(Google∩Yahoo)\Bing	241 (2.78%)	152 (2.29%)	158 (3.19%)	0 (0.00%)
(Google∩Bing)\Yahoo	606 (6.98%)	27 (0.41%)	164 (3.32%)	8 (1.20%)
Yahoo\(Google∪Bing)	139 (1.60%)	1075 (16.20%)	332 (6.71%)	60 (9.02%)
Bing\(Google∪Yahoo)	522 (6.01%)	444 (6.69%)	293 (5.92%)	83 (12.48%)
Google\(Yahoo∪Bing)	4783 (55.09%)	825 (12.43%)	1388 (28.06%)	113 (16.99%)
Google∪Yahoo∪Bing	8682	6636	4946	665

Table 4.22: Number of overlapping queries with each pattern for each search engine (the percentages are computed with the number of corresponding columns' union) - January 2012

Overlapping Queries January 2012				
Search Engine	Pattern 0	Pattern 1	Pattern 2	Pattern 3
Google∩Yahoo∩Bing	1237 (14.59%)	271 (4.94%)	1669 (29.00%)	59 (6.76%)
(Bing∩Yahoo)\Google	884 (10.43%)	3425 (62.43%)	2448 (42.54%)	414 (47.42%)
(Google∩Yahoo)\Bing	151 (1.78%)	13 (0.24%)	144 (2.50%)	0 (0.00%)
(Google∩Bing)\Yahoo	400 (4.72%)	11 (0.20%)	66 (1.15%)	1 (0.11%)
Yahoo\(Google∪Bing)	139 (1.64%)	407 (7.42%)	364 (6.32%)	48 (5.50%)
Bing\(Google∪Yahoo)	217 (2.56%)	133 (2.42%)	130 (2.26%)	308 (35.28%)
Google\(Yahoo∪Bing)	5448 (64.28%)	1226 (22.35%)	934 (16.23%)	43 (4.93%)
Google∪Yahoo∪Bing	8476	5486	5755	873

As another analysis, for each pattern, we compute the number of overlapping queries among different search engines. Results are shown in Table 4.21 and Table 4.22. In Table 4.21 the number of the overlapping queries is shown among all three search engines in July 2011. For better understanding, we also report the number of the queries that return a pattern by only a single search engine.

We observe that, the highest agreement among all three search engines is for the queries with Pattern 2. For instance, among 4946 queries which belong to Pattern 2, 1186 are common in all three search engines and it is about 24%. However the agreement among all three search engines is nearly 19% of the queries with Pattern 0, 4% of the queries with Pattern 1 and 17% of the queries with Pattern 3. This shows that, there are some common queries which can be directly answered or cannot be answered by all three search engines. Interestingly, the amount of overlap is a bit low for queries with Pattern 0. Especially Google differs from the others. For instance, 55% of the queries with Pattern 0 are tagged with Pattern 0 by Google, but Bing and Yahoo! tag them with different patterns. The same trend can be seen for Pattern 2. Google again differs from Bing and Yahoo!. 28% of the queries are tagged with Pattern 2 by only Google and 29% of the queries are tagged with Pattern 2 by Yahoo! and Bing, but not in Google. For Pattern 3, Yahoo! and Bing have a similar strategy, 43% of the queries are tagged with Pattern 3 by only Yahoo! and Bing. In Table 4.22, the number of overlapping queries is shown among all three search engines in January 2012. Like the results of July 2011, similar trends can be seen here. Again we observe that, the highest agreement among all three search engines is for the queries with Pattern 2. This time 29% of the queries with Pattern 2 are common in all three search engines. 43% of the queries with Pattern 3 are common in only Yahoo! and Bing. Google again differs from them. The number of common queries with Pattern 2 is increased in January 2012 (from 24% to 29%). The agreement among all three search engines is nearly 15% of the queries with Pattern 0, 5% of the queries with Pattern 1 and 7% of the queries with Pattern 3. The number of common queries decreases in Pattern 0, Pattern 1 and Pattern 3. Despite the fact that the number of queries with Pattern 3 increases, the number of common queries with Pattern 3 decreases. Because the number of NAQs decreases in Google and it makes it harder to find common queries with Pattern 3 among all three search engines. Google again differs from Bing and Yahoo! with Pattern 0. For instance, 64% of the queries with Pattern 0 are tagged with Pattern 0 by Google, but Bing and Yahoo! tag them with different patterns. Again with Pattern 1, 62% of queries with Pattern 1 are tagged as Pattern 1 by Bing and Yahoo! but these queries are tagged with another pattern by Google.

Table 4.23: Number of overlapping query suggestions for queries with Pattern 1 and Pattern 2 for each search engine (the percentages are computed with respect to the corresponding values in Table 4.21) - July 2011

Overlapping Query Suggestions July 2011		
Search Engine	Pattern 1	Pattern 2
Google ∩ Yahoo ∩ Bing	29 (10.62%)	245 (20.66%)
(Bing ∩ Yahoo) \ Google	2834 (73.80%)	1235 (86.67%)
(Google ∩ Yahoo) \ Bing	46 (30.26%)	40 (25.32%)
(Google ∩ Bing) \ Yahoo	2 (7.41%)	16 (9.76%)

Table 4.24: Number of overlapping query suggestions for queries with Pattern 1 and Pattern 2 for each search engine (the percentages are computed with respect to the corresponding values in Table 4.22) - January 2012

Overlapping Query Suggestions January 2012		
Search Engine	Pattern 1	Pattern 2
Google ∩ Yahoo ∩ Bing	3 (1.10%)	417 (24.99%)
(Bing ∩ Yahoo) \ Google	2759 (80.56%)	2007 (81.99%)
(Google ∩ Yahoo) \ Bing	0 (0.00%)	71 (49.31%)
(Google ∩ Bing) \ Yahoo	0 (0.00%)	11 (16.67%)

We also investigate the overlap of suggested queries for Pattern 1 and Pattern 2 between different search engines. As seen in Table 4.23, the overlap for the three search engines is twice larger for queries with Pattern 2 than those in Pattern 1 in July 2011. The overlap between Yahoo! and Bing is very high for queries with Pattern 1 and Pattern 2. 73.8% of the queries with Pattern 1 and 86.7% of the queries with Pattern 2 have the same suggestions for Yahoo! and Bing but Google suggests different alternatives for these queries. Especially Google and Bing have very few queries with the same suggestions. Table 4.24 shows the number of overlapping query suggestions for queries with Pattern 1 and Pattern 2 for each search engine in January 2012. The overlap between all three search engines is around 1% of the queries with Pattern 1. This result is very low when compared with the result of July 2011. Similar to July 2011 results, 25% of the queries with Pattern 2 have the same suggestions in January 2012. This may imply that all of the search engines can detect the user intent better for these queries with Pattern 2. The overlap between Yahoo! and Bing is again very high for queries with Pattern 1 which is 80.6% of the queries and Pattern 2 which is 82% of the queries. One important point is that there is no overlap between Google ↔ Yahoo! and Google ↔ Bing for queries with Pattern 1.

4.2.7 Methods for Generating Suggestions

In Pattern 1 and Pattern 2, search engines suggest an alternative query that modifies the original one. We manually inspected all queries that are answered by either one of these patterns by all three search engines to identify the types of modifications applied on the original query to create a suggestion. In July 2011, 273 queries are labeled as Pattern 1 and 1186 queries are labeled as Pattern 2 by all search engines. We observe that a large amount of the queries entirely or partially include URI. For instance among the 273 queries that are labeled with Pattern 1 by all three search engines, the amount of queries with a URI adds up to 71%. For Pattern 2, the percentage is smaller but still significant: 52% of the 1152 queries contain a URI. In January 2012, 271 queries are labeled as Pattern 1 and 1669 queries are labeled as Pattern 2 by all search engines. The amount of queries with a URI increases to 94% of the 271 queries that are labeled as Pattern 1. Similarly 52% of the 1669 queries with Pattern 2 contain a URI in January 2012. Due to the frequent presence of URIs in queries which are shown above, we present the modifications by search engines for these two types of queries, with and without URIs, separately in Table 4.25. Note that, more than one of these modifications is applied in many cases.

Table 4.25: Most frequent modifications to queries without (M1-M5) and with a URI (M6-M11)

Modifications	Original Query	Suggested Query
M1 Split query string to terms	3rdgenerationgospelsingers	3rd generation gospel singers
M2 Correct typo in a term (insert/delete/replace character)	tadeair compter show	trade air computer show
M3 Combine terms in query string	cup cakes	cupcakes
M4 Add/delete punctuation	childerns hosptial of birmaham	children's hospital of birmingham
M5 Add/delete/replace term	woodiestationwagons	woody station wagons
M6 Split URI to terms	www.eldercare-today.com	elder care today
M7 Correct typo in a term in URI	www.orlandocollages.com	www.orlandocolleges.com
M8 Add/delete/replace term in URI	www.online-houses-for-sale.com	www.online-homes-for-sale.com
M9 Re-order terms in URI	ri-rvs.com	rvs-ri.com
M10 Add/delete punctuation	street-racingvideos.com	street-racing-videos.com
M11 Add/delete/replace domain extension	www.innuendo-music.de	www.innuendo-music.com

Table 4.25 reveals that there are some fundamental differences between the modifications applied to queries with or without URIs. In particular, the most common modifications are adding spaces between the words for queries without URIs and then correcting typos within the terms. On the other hand, only 30% of URI queries involve an obvious typo while the rest do not necessarily contain a spelling mistake in a strict sense. However, they are possibly due to the users, who confused "com" with "biz", or forgot the hyphen between the terms. For this latter class of suggestions, search engines probably use the existence of other closely similar URIs as a clue.

4.2.8 Suggestion Quality

As a complementary experiment, we also investigate the accuracy of the suggestions. To this end, we randomly selected two subsets, each with 100 queries, from the queries that yielded results with Pattern 1 and Pattern 2 from all three search engines. We conducted a user study with 6 participants. The original query and the suggested query are shown to each user and they are asked to decide whether the correction/suggestion makes sense, or not.

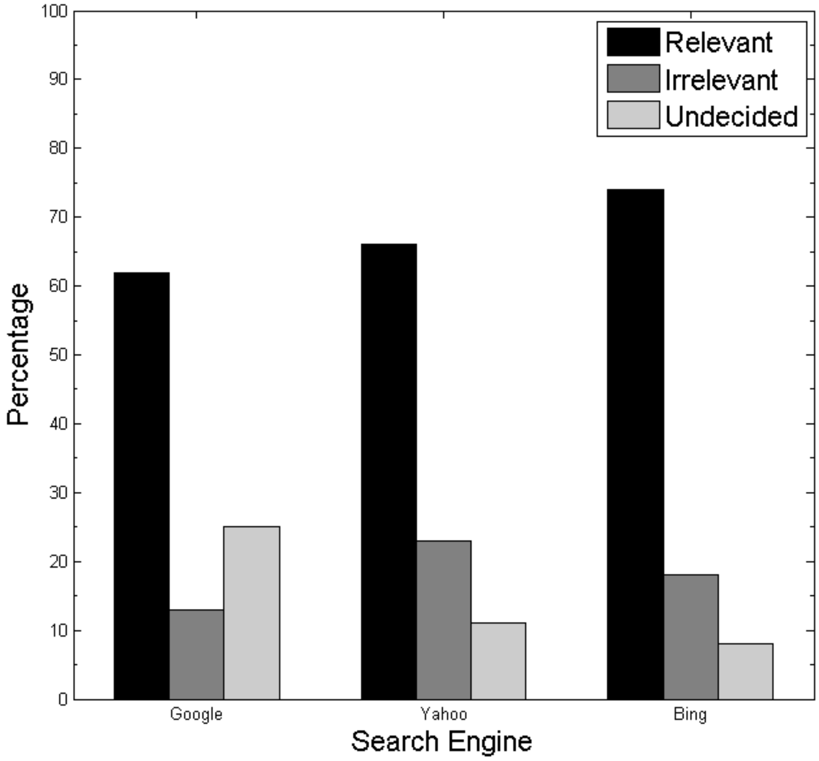

Figure 4.1: Suggestion quality for Pattern 1

In Figure 4.1, the percentages of suggestions labeled as relevant, irrelevant and undecided are shown for Pattern 1 by the judges. The figure shows that the lowest number of irrelevant results is yielded by Google. However, because of the high fraction which is around 25% of suggestions that are left undecided, it is not the best performing search engine. A closer inspection reveals that, Google consistently prefers to provide alternative URI suggestions, whereas the other two search engines Yahoo! and Bing split the URI into terms as a suggestion for most of the cases (for instance, applying M6 in Table 4.25). Participants of the user study could not decide on how good the suggested URI captures the initial intention of the user in a number of cases so Google yields lots of undecided suggestions. For instance, both Yahoo! and Bing suggest 'toledo sona systems' for the query 'toledo.sona-systems.com'. It is labeled as relevant by the judges. On the other hand, Google suggests 'utoledo.sona-systems.com' for the same query and it is labeled as undecided as the judges who do not have any background information to evaluate the correctness of this suggestion. Sometimes search engines offer different suggestions which are also relevant to original query. For instance, for the query 'chemical-records.org', Google suggests 'chemical-records.com' and judges label it as relevant. The other two search engines Yahoo! and Bing suggest 'chemical records' for the same query and it is also labeled as relevant by the judges. Similarly, Google suggests 'www jacquielawson.com renewal.asap' for the query 'wwwjacquielawson.comrenewal.asap' and both of Yahoo! and Bing suggest 'jacquielawson com renewal asap' for the same query but all of the suggestions are labeled as irrelevant by the judges for this suggestions. The figure shows that Bing yields the most number of relevant results which is 74% of the suggestions. 66% of the suggestions are labeled as relevant in Yahoo! and 62% of the suggestions are labeled as relevant in Google. Nevertheless, it can be seen that the fraction of irrelevant suggestions vary between 13% and 23% for the queries with Pattern 1.

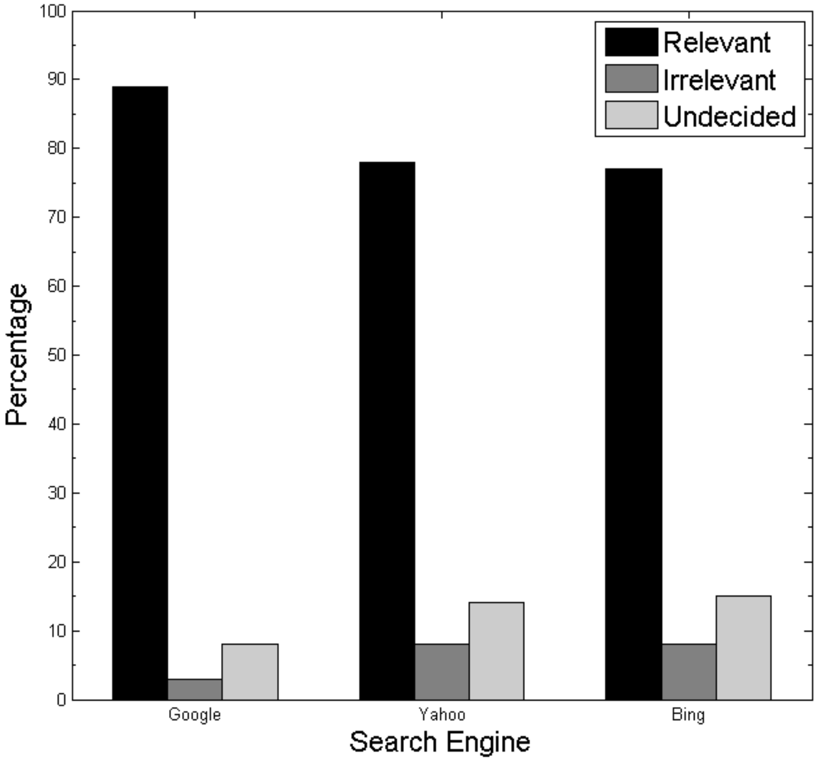

Figure 4.2: Suggestion quality for Pattern 2

In Figure 4.2, the percentages of suggestions labeled as relevant, irrelevant and undecided are shown for Pattern 2 by the judges. For the queries with Pattern 2, all search engines provide a higher fraction of relevant suggestions in comparison to those for Pattern 1 queries. Similarly, the figure shows that the lowest number of irrelevant results which is 3% of the suggestions is yielded by Google. This time Google yields the most number of relevant results which is 89% of the suggestions. 78% of the suggestions are labeled as relevant in Yahoo! and 77% of the suggestions are labeled as relevant in Bing. For instance, for the query 'mycolegeguide.org', Google suggests 'mycollegeguide.org' and the judges label it as relevant. The other two search engines Yahoo! and Bing suggest 'my college guide' for the same query and it is also labeled as relevant by the judges. This is a result that further confirms our intuition discussed before that search engines return results with Pattern 2 only when they are more confident with their suggestions. For example, for none of the queries all three search engines' suggestions are labeled as irrelevant. In this case, the ratio of suggestions labeled as irrelevant is less than 10% for all three search engines. The above findings are also important to comprehend the appropriateness of our query log for such a study. An astute reader could suspect that the large fraction of URIs that appears in our set of hard queries can be caused due to age of our query log. For example, many of the searched URIs might have disappeared within time, yielding no or few results for these queries. However, the modifications as exemplified in 4.25 indicate that URI queries that are handled by Pattern 1 or Pattern 2 essentially include a mistake that in spelling or typing the word exactly as it appears in URI. And apparently this is the main reason for most of these queries to retrieve very few or no results but not the possibility of these sites being disappeared within time.

Chapter 5

Analysis of No Answer Queries

As we mentioned before, search engines cannot provide any results for some queries despite their complex mechanisms. Especially if the users search a content in a less common language or an unpopular web page, search engines have some difficulties to provide results. Similarly, if the query is unusually long or contains an infrequent terms, usually search engines cannot even provide any suggestions. In this chapter, we focus on such kind of queries that we call No Answer Queries (NAQs), and present our experimental result about them.

5.1 Experimental Study of NAQs

5.1.1 Dataset

The last column of Table 4.21 provided in Section 4.2.6 shows the figures for NAQs (Pattern 3 queries) obtained in July 2011. Similarly, the last column of Table 4.22 provided in the same section shows the figures for NAQs in January 2012. The total number of queries that retrieve no answers at all for at least one search engine is 665 in July 2011. This number increases to 873 in January 2012. We use both of these query sets as our NAQ sets and investigate their characteristics in this chapter.

Table 5.1: Example No Answer Queries that are manually selected from the AOL query log

No Answer Queries	Potential reason for not matching any result
zhgadghouchchxjxcxvbnccxcjhshixmnx	Query term does not appear in the vocabulary of the search engine
sprocitletsgrout	Query contains typos that cannot be fixed by the spelling corrector
maazavioavio.com	URI is not discovered by the search engine
healperware tea & coffee pot made in china	No web page contains all of the query terms
- - - - - - -	Query has insufficient information
www.cbcoloradomerrillcorp.net	URI does not exist in the Web

Table 5.1 shows a small number of NAQs that are selected from the NAQ set described above. Potential reasons for being NAQ predicted by us are also given in the table. The root causes are very diverse and too many, so we believe that introducing a classification of NAQs is difficult. In most cases, it is difficult to identify a single reason because most of the times a combination of factors are decisive. Consequently, rather than analyzing the potential reasons, we prefer to provide an analysis on the characteristics of NAQs.

5.1.2 User Study

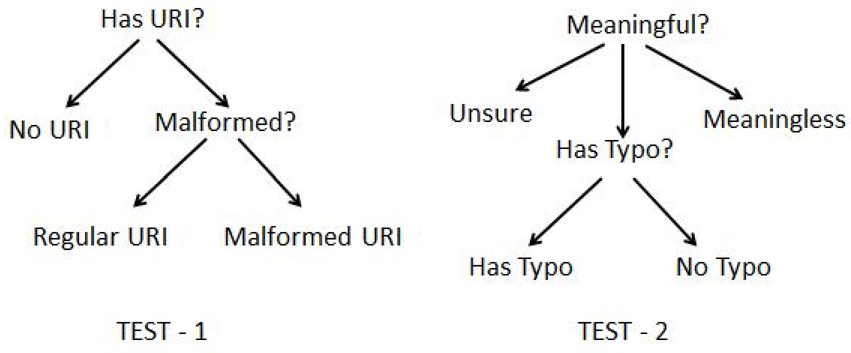

Figure 5.1: The procedure followed by the judges in the user study

We first conduct two user studies over the queries in the NAQs set in July 2011 and January 2012, and label them by four human judges based on two types of tests: URI presence and meaningfulness. Figure 5.1 illustrates the query labeling procedure followed by judges. We separately report the results for each search engine as well as the union and the intersection of their NAQ sets. Our first test is about the presence of URIs. According to our procedure, if a query contains any URI, we check the URI to tag it as "Regular URI" or "Malformed URI". If the URI contains any typo or any error, it is labeled as "Malformed URI" otherwise it is labeled as "Regular URI". If there is no URI part in the query it is labeled as "No URI". For instance 'oolpentricks.comhttp' and 'meridianreource.comwwwmeridianresorse.com' are labeled as "Malformed URI", 'orgbbfl.football.sportsline.com and 'springfieldregionalplanningcommision.com'are labeled as "Regular URI", 'richardsmallwoodporty'and 'hgutugugjjv n n bhv b cv nhv'are labeled as "No URI". Our second test evaluates the meaningfulness of the NAQs. If a query contains a URI, we only consider the remaining part. For instance the query 'westlifeonline.com belly'contains a URI part 'westlifeonline.com'. In this test we consider the remaining part of the query which is 'belly'. If the entire query is a URI, it is labeled as "Only URI" and excluded from this test. If the meaning of the query is not clear to judge, but the NAQ has a potential to have a meaning for the user who issued it, the query is labeled as "Unsure". For example, 'u.s.c.g.c.w.p.g.44 wachusett' is one of the queries which is labeled as "Unsure" by all of the judges. Queries that are clearly meaningless to the judge are labeled as "Meaningless". Queries that are only formed of repetitive key strokes are generally in this class. For example '- - - - - - - - - -', 'fsgdfhgfdg' and 'dgffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff' are labeled as "Meaningless". The remaining queries are considered as meaningful and labeled as "Has Typo" or "No Typo" depending on the presence of a typo. For example, 'richardsmallwoodporty' and 'self condifencense' are labeled as "Has Typo" because of the typo contained. Similarly, 'dvd -r vs dvd r' and 'hack all101011001010101010.com' are labeled as "No Typo". Note that in the second query which is 'hack all101011001010101010.com', we only consider the 'hack' part because it contains the URI part 'all101011001010101010.com'.

5.1.3 Experimental Results of User Study

Table 5.2: Distribution of No Answer Queries based on the presence of a URI - July 2011

	Malformed URI	Regular URI	No URI
Google	28 (17%)	59 (30%)	146 (49%)
Yahoo	132 (78%)	120 (60%)	209 (71%)
Bing	138 (82%)	142 (71%)	222 (75%)
Google ∪ Yahoo ∪ Bing	169 (25%)	200 (30%)	296 (45%)
Google ∩ Yahoo ∩ Bing	16 (2%)	14 (2%)	84 (13%)

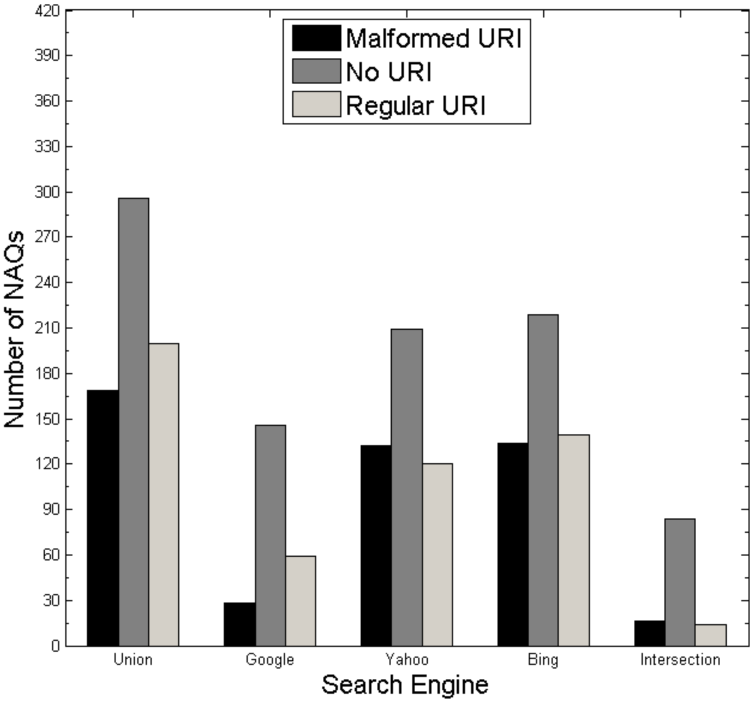

Figure 5.2: Distribution of No Answer Queries based on the presence of a URI - July 2011

Our first test evaluates the presence of URIs in NAQs in both July 2011 and January 2012. Despite the fact that it can be possible to automate this test via a pattern matching technique, we prefer to do it manually. Because it is difficult to automatically catch URIs that contain typos. Figure 5.2 and Table 5.2 show the distribution of NAQs based on the presence of a URI in July 2011. They indicate that about 55% of the NAQs contain at least one URI. About 46% of these contain at least one malformed URI, while the remaining 54% are proper URIs. This shows that about one-third of NAQs aim to retrieve resources that are unknown to or not discoverable by the search engine. It might not be possible to solve these NAQs by any technique that is used now. When we compare the result of the search engines, we observe that Google is significantly better in solving NAQs with malformed URIs. Only 17% of the queries with malformed URIs cannot be solved by Google. Yahoo! cannot find any results for 78% of the queries with malformed URIs and similarly Bing cannot find any results for 79% of such queries. Google is also better in solving NAQs with Regular URIs and No URIs. Google has 30% of the queries with Regular URI and 49% of the queries with No URI. Yahoo! has respectively 60% and 71% of the queries, while Bing has respectively 69% and 74% of the queries with Regular URI and No URI. The number of such NAQs in Google is slightly higher than those present in the intersection set of the three search engines. In the intersection set, 10% of the queries with malformed URIs cannot be solved by any search engine. 7% of the queries with Regular URI and 28% of the queries with No URI also cannot be solved by any search engine. Overall, the size of the intersection is much smaller than the size of the union. It implies that most of the NAQs are solved by at least one search engine.

Table 5.3: Distribution of No Answer Queries based on the presence of a URI - January 2012

	Malformed URI	Regular URI	No URI
Google	17 (9%)	10 (4%)	76 (19%)
Yahoo	127 (67%)	141 (51%)	253 (62%)
Bing	159 (84%)	252 (92%)	371 (91%)
Google ∪ Yahoo ∪ Bing	190 (22%)	275 (31%)	408 (47%)
Google ∩ Yahoo ∩ Bing	4 (1%)	5 (1%)	50 (6%)

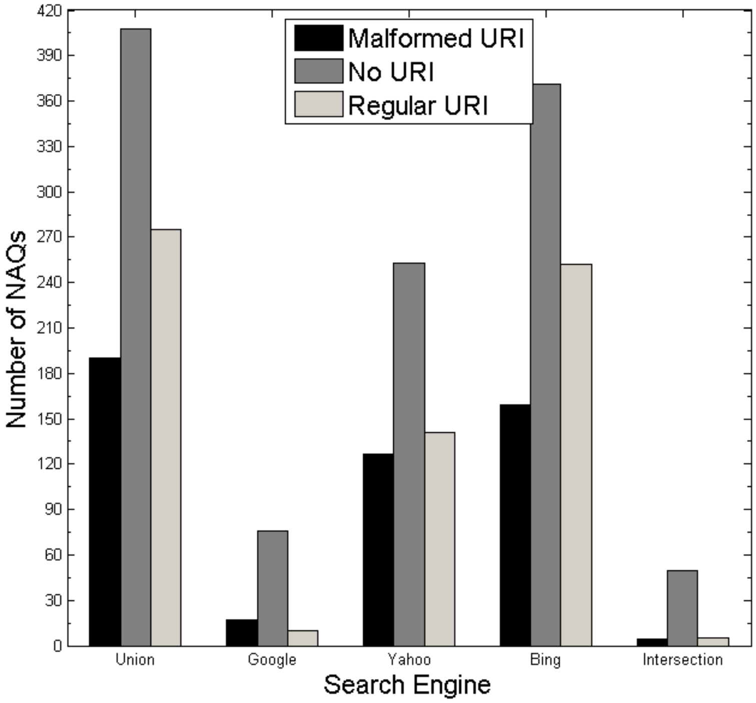

Figure 5.3: Distribution of No Answer Queries based on the presence of a URI - January 2012

Figure 5.3 and Table 5.3 shows the distribution of NAQs based on the presence of a URI in January 2012. Similar to the results obtained in July 2011, 53% of the NAQs contain at least one URI. About 41% of these contain at least one malformed URI while the remaining 59% are proper URIs. When we compare the results of the search engines obtained in January 2012, we observe that Google is still significantly better in solving NAQs with malformed URIs. This time only 9% of the queries with malformed URIs cannot be solved by Google. This is nearly half of that in July 2011. Yahoo! cannot find any results for 67% of the queries with malformed URIs and similarly Bing cannot find any results for 84% of such queries. Yahoo! makes a little improvement with malformed URIs. Google has 4% of the queries with Regular URI and 19% of the queries with No URI. These results are a lot better than the results of July 2011. Yahoo! has 51% of the queries with Regular URI and 62% of the queries with No URI while Bing has 92% of the queries with Regular URI and 91% of the queries with No URI. The results show that Bing has some problems in handling the NAQs. Although Google and Yahoo! have some improvements, Bing has worse results than those observed in July 2011. The number of such NAQs in Google is closer to those present in the intersection set of the three search engines. In the intersection set of the search engines, only 2% of the queries with malformed URIs cannot be solved by any of. Similarly, 2% of the queries with Regular URI and 12% of the queries with No URI also cannot be solved by any search engine. The size of the intersection is much smaller than that of July 2011 and also than the size of the union set. We can say that search engines have improved their methods to handle NAQs, except Bing.

Table 5.4: Distribution of No Answer Queries based on the meaningfulness - July 2011

	Has Typo	Meaningless	No Typo	Unsure	Only URI
Google	42 (44%)	78 (57%)	3 (14%)	30 (34%)	80 (25%)
Yahoo	44 (46%)	118 (86%)	11 (52%)	67 (75%)	221 (68%)
Bing	51 (54%)	121 (88%)	21 (100%)	67 (75%)	232 (72%)
Google ∪ Yahoo ∪ Bing	95 (14%)	137 (21%)	21 (3%)	89 (13%)	323 (49%)
Google ∩ Yahoo ∩ Bing	8 (1%)	64 (10%)	1 (0%)	15 (2%)	26 (4%)

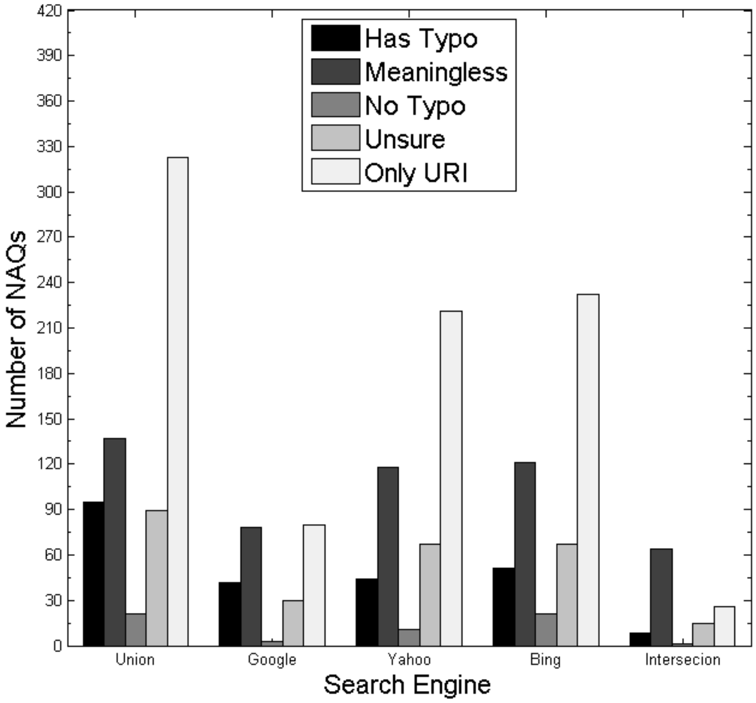

Figure 5.4: Distribution of No Answer Queries based on the meaningfulness - July 2011

Our second test is about the meaningfulness of the NAQs. The results of this test are shown in Figure 5.4 ↔ Table 5.4 and Figure 5.5 ↔ Table 5.5. According to Figure 5.4 and Table 5.4 of July 2011, a considerable portion of the NAQs are labeled as “unsure”, so the numbers reported for the remaining labels can act only as lower bounds. According to these results, only 3% of the NAQs are meaningful and do not contain any typos. It is interesting to note that, we encountered only one such NAQ that is not solved by any search engines. This query is ‘12683990476 http track.airborne.com atrknav.asp shipmentnumber 12683990476’, and none of the three search engines can provide any results for it. 14% of NAQs are also meaningful and contain typos. This means that, at least, four out of every five NAQs that are meaningful contain some typo. 21% of the NAQs do not have any meaning. When we compare the results of the search engines, we observe that Google is again significantly better than the other search engines. Only 14% of the queries which are labeled as No Typo cannot be solved by Google. Yahoo! cannot provide any results for 52% of the queries with No Typo, and Bing cannot provide any results for all of the queries which are labeled as No Typo. Google is better, more than two times compared to Yahoo! and Bing with queries which are labeled as Unsure. Google has 34% of the NAQs with Unsure but each of Yahoo! and Bing has 77% of the NAQs with Unsure. On the average, 47% of the NAQs with Has Typo cannot be solved by any search engine, but only 8% of them cannot be solved by all the three search engine.

Table 5.5: Distribution of No Answer Queries based on the meaningfulness - January 2012

	Has Typo	Meaningless	No Typo	Unsure	Only URI
Google	21 (11%)	51 (29%)	5 (17%)	14 (14%)	12 (3%)
Yahoo	74 (40%)	145 (83%)	18 (62%)	78 (76%)	206 (54%)
Bing	163 (87%)	158 (90%)	21 (72%)	96 (94%)	344 (91%)
Google ∪ Yahoo ∪ Bing	187 (21%)	175 (20%)	29 (3%)	102 (12%)	380 (44%)
Google ∩ Yahoo ∩ Bing	4 (1%)	39 (4%)	1 (0%)	10 (1%)	5 (1%)

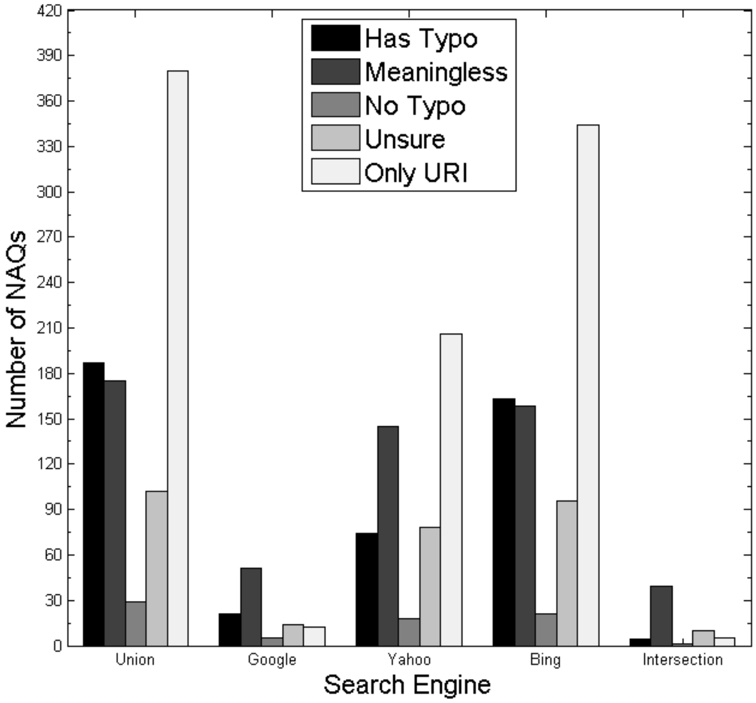

Figure 5.5: Distribution of No Answer Queries based on the meaningfulness - January 2012

Figure 5.5 and Table 5.5 show the distribution of NAQs based on meaningfulness in January 2012. Similar to the results presented for July 2011, only 3% of NAQs are meaningful and do not contain any typos. It is interesting that, we again encountered only one such NAQ that is not solved by any search engines. But this time the query is ‘hack all101010100101010100101100.com’ ’ in January 2012 and none of the three search engines can provide any results for it. However, in July 2011 the search engines provide a total of 5 results for the same query. 21% of the NAQs are also meaningful and contain typos. Similar to the July 2011 results, 20% of NAQs do not have any meaning. When we compare the results of the search engines, we observe that Google is still significantly better than the other search engines. 17% of the queries which are labeled as No Typo cannot be solved by Google. Yahoo! cannot provide any results for 62% of such queries, while Bing cannot provide any results for 72% of them. This shows that Bing makes some improvements compared to July 2011 results. Google is better, more than five times compared Yahoo! and Bing with queries which are labeled as Unsure. Google has 14% of NAQs with Unsure, while Yahoo! has 77% and Bing has 94% of NAQs labeled as Unsure. 11% of the NAQs with Has Typo cannot be solved by Google, while 40% of such queries cannot be solved by Yahoo! and 87% of them cannot be solved by Bing. However, only 2% of the NAQs with Has Typo cannot be solved by all of the three search engines. All these improvements observed in January 2012 show that search engines are adopting more sophisticated techniques to handle NAQs.

5.2 Quantitive Feaures

In Figure 5.6 and Figure 5.7, we display the distribution of NAQs and regular queries as the query length increases. In Figure 5.8 and Figure 5.9 the behavior in terms of the number of characters in the query are shown.

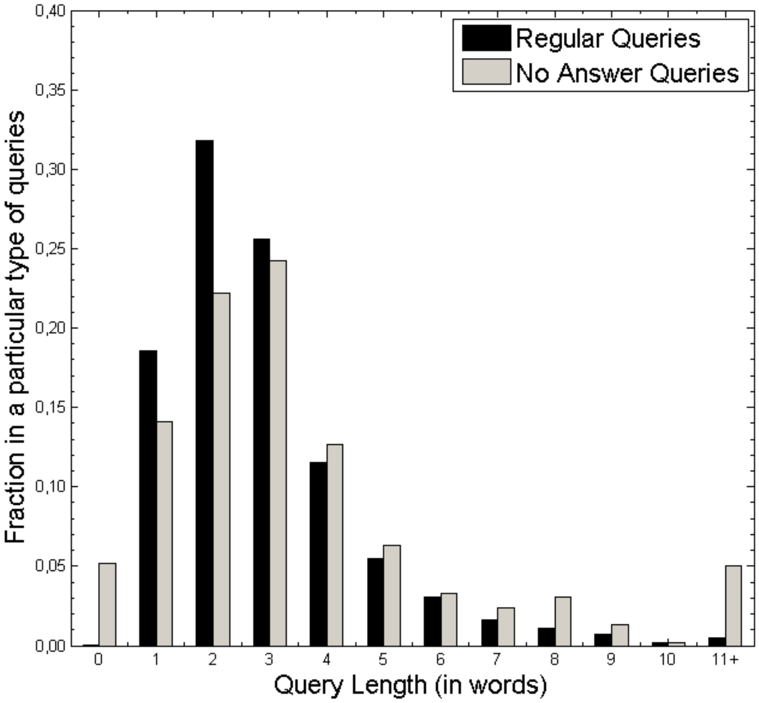

Figure 5.6: Query length (in words) distribution - July 2011

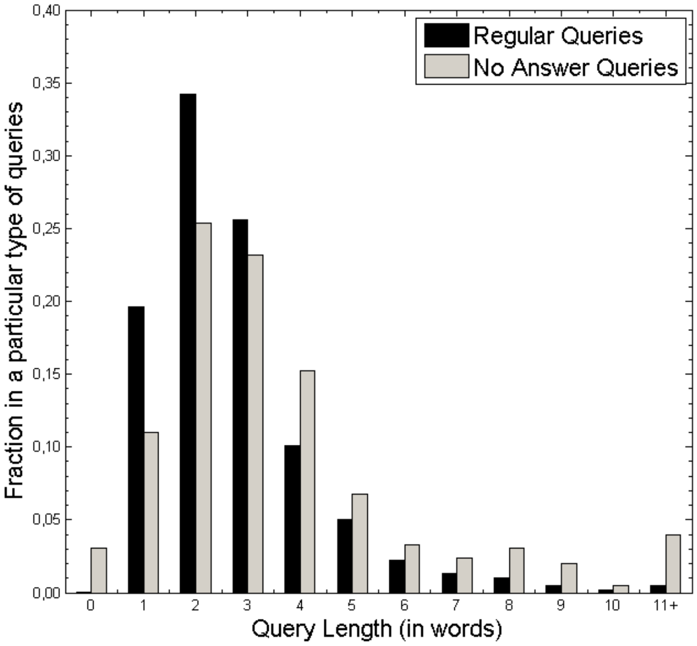

Figure 5.7: Query length (in words) distribution - January 2012

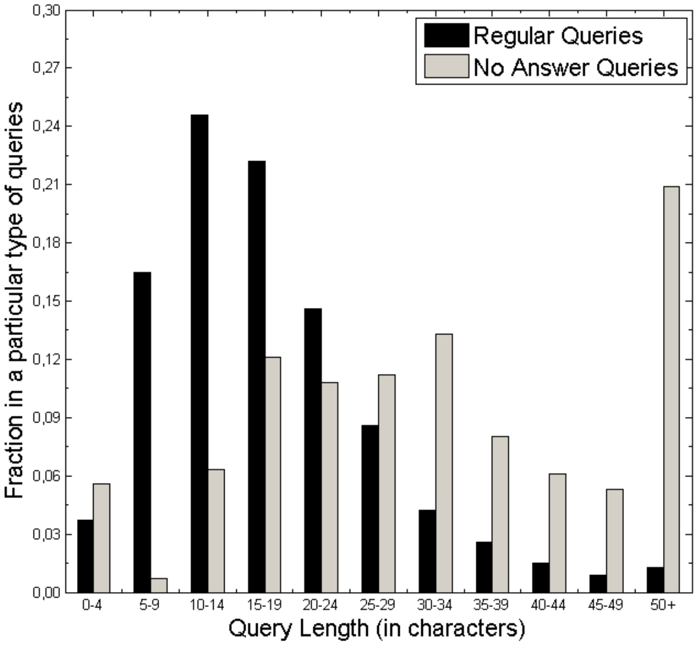

Figure 5.8: Query length (in characters) distribution - July 2011

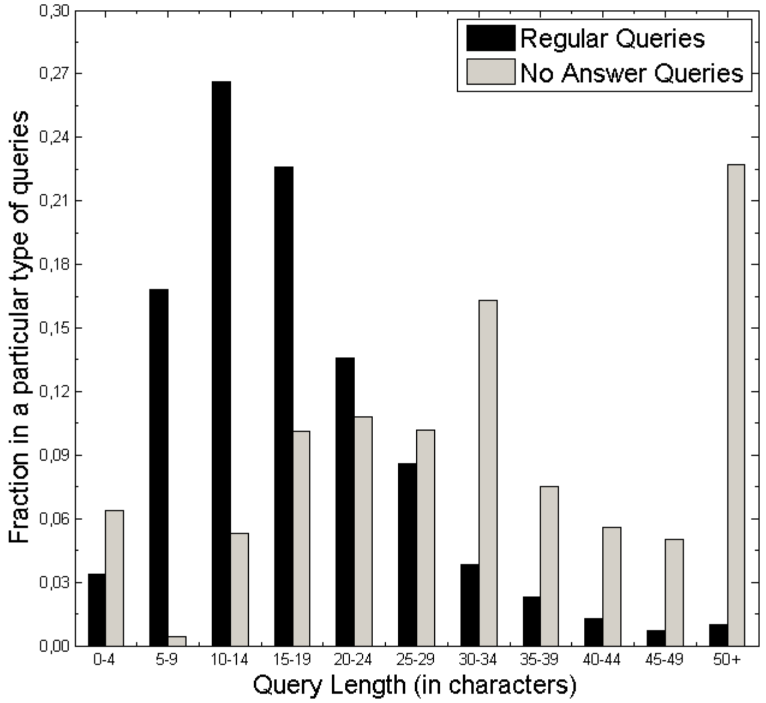

Figure 5.9: Query length (in characters) distribution - January 2012

Figure 5.6 presents the distribution of NAQs and regular queries in July 2011. The fraction of NAQs for queries with one to three terms is lower than those for regular queries. The NAQs are shifted towards longer queries. Overall, this behavior can be explained by two observations. First, it is well known that most regular Web queries include one to three terms, so NAQs are not likely to dominate this range. Second, as there are more terms, it becomes harder to match the query to a document that contains all query terms. The second factor becomes very dominant at large query lengths, which explains the significantly high ratio of NAQs when there are more than 10 terms. Figure 5.8 shows the behavior in terms of the number of characters in the query in July 2011. We observe that the NAQ likelihood is more skewed towards queries with many characters, compared to regular queries.

Figure 5.7 shows the distribution of NAQs and regular queries in January 2012. Similar to the results obtained in July 2011, the NAQs are again shifted towards longer queries. Trends are very similar with the results obtained in July 2011. In Figure 5.9 the behavior in terms of the number of characters in the query is shown in January 2012. Again similar trends can be seen when the results are compared to July 2011 results.

Chapter 6

Prediction of Query Suggestion Patterns and No Answer Queries

As we mentioned in previous chapters, search engines return query suggestions for hard queries. We envision that predicting query suggestion pattern with a model can be beneficial in some use case scenarios, so we build a machine learning model for this purpose. In this chapter, we deal with two interesting prediction tasks that are the prediction of query suggestion patterns of search engines and No Answer Query (NAQ) prediction. As the learner we use Decision Trees [55].

6.1 Predicting Query Suggestion Patterns

6.1.1 Decision Trees

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Each node may have two or more branches. It represents a decision support tool used very often because it is simple to understand and interpret. Decision tree models are commonly used in data mining to examine the data and to induce the tree and its rules that will be used to make predictions. Decision trees offer advantages over other methods of analyzing alternatives. Decision alternatives, possible outcomes, and chance events can be represented schematically. Complex alternatives can be expressed clearly. We can easily modify a decision tree as new information becomes available [56]. We use Weka Data Mining Software in Java [57] tool to implement Decision Tree.

6.1.2 Problem

Our first task is to predict whether a search engine would present its results using Pattern 1 or Pattern 2 (See section 4.2.2) if there is an alternative query with potentially better results. Since we do not have a real ground-truth data on the suggestion patterns, we use the decisions made by the search engines as the ground-truth. Obviously, this is not a perfect formulation for the problem but, at least, it gives us an idea about the difficulty of the prediction problem at hand. Moreover, it gives us a hint about how search engines differ in this decision.

6.1.2.1 Experimental Setting

Table 6.1: The features used by learning model for the pattern prediction problem

Feature	Description
originalResultCount	# of results for the original query
suggestedResultCount	# of results for the suggested query
originalQueryLength	# of characters in the original query
suggestedQueryLength	# of characters in the suggested query
editDistance	the edit distance between the two queries

For each of the three search engines, we build a separate learning model using the features given in Table 6.1. Because of the limited dataset, we can use only these features. We have only the real query, suggested query, number of real query result and number of suggested query result information. So our feature set is quite limited. In July 2011 all three search engines respond with the same pattern that are Pattern 1 and Pattern 2 for 1460 queries in this set. This set contains 274 queries with Pattern 1 and 1186 queries with Pattern 2. In January 2012, 1940 queries are responded with the same pattern that are Pattern 1 and Pattern 2 by all three search engines. This set contains 271 queries with Pattern 1 and 1669 queries with Pattern 2. We perform ten-fold cross validation and all accuracy results are averaged for ten folds. Our aim here is to increase the confidence in predictions since we just rely on search engines suggestions as the ground truth.

6.1.3 Experimental Results

Table 6.2: Prediction performance (percentages for the <actual pattern, predicted pattern> pairs) - July 2011

SE	<P1, P1>	<P1, P2>	<P2, P1>	<P2, P2>
Google	6.2	12.5	15.9	65.3
Yahoo	10.5	8.3	12.0	69.2
Bing	12.5	6.3	14.1	67.1
Average	9.7	9.0	14.0	67.2

Table 6.3: Prediction performance (percentages for the <actual pattern, predicted pattern> pairs) - January 2012

SE	<P1, P1>	<P1, P2>	<P2, P1>	<P2, P2>
Google	5.1	8.9	13.4	72.7
Yahoo	8.4	5.6	12.3	73.8
Bing	8.9	5.1	12.8	73.2
Average	7.5	6.5	12.8	73.2

Table 6.2 shows the prediction performance in July 2011. Four possible combinations of the actual and predicted patterns are listed. In the table, <P1, P2> means that Pattern 1 is preferred by the search engine while the prediction made by the model is Pattern 2. On the average, the prediction accuracy is fairly good

with 76.9%. In Table 6.3 prediction performance of January 2012 is shown. On the average, the prediction accuracy, which is 80.7%, is better than the result of July 2011. Despite the fact that we have limited feature set, the prediction accuracy is fairly good. We observe that our accuracy is lower in predicting the behavior of Google compared to Yahoo! and Bing in both July 2011 and January 2012. This might imply that Google potentially has a more complex decision logic, which cannot be adequately captured by the simple features used by our predictive model.

6.2 Predicting No Answer Queries (NAQs)

We envision that predicting NAQs with a model can be beneficial in some use case scenarios, such as mobile web search or meta search. For example in mobile search, network bandwidths are limited, packet transmission rates are low, and the cost of accessing the Internet is high. In this scenario, a predictive model deployed within a mobile device can warn the user if the query is not likely to return any answers. This can provide significant saving in terms of time, bandwidth usage, power consumption and monetary costs. Another scenario is meta search. In this scenario predicting NAQs can be beneficial in a meta-search engine that forwards queries to component search systems and apply a re-ranking algorithm on the returned results. Every query forwarded to a component system may incur some financial cost to the meta-search engine [58]. In such a scenario, the meta-search system can build a separate NAQ predictor for each search service and can forward the query to only those services that are predicted to return some results. This may reduce the bandwidth usage and financial costs of the system while reducing the load on the search services.

6.2.1 Problem

Table 6.4: The features used by the model for the NAQ prediction problem

Features
#OfWords
#OfCharacters
averageWordLength
#OfPlusSymbols
#OfMinusSymbols
#OfQuotationSymbols
#OfDigits
#OfUpperCases
#OfNotAlphaNumeric
fractionOfDigits
fractionOfUpperCases
fractionOfNotAlphaNumeric

We cast the problem of predicting whether a query will return no results as a classification task. We try to solve it using machine learning techniques. The main goal is to model a set of response variables that are the class of a given query as a function of a set of explanatory variables which are the features associated with the query. In our case, we have a binary classification problem where queries belong to the no answer or regular categories. The set of features used by the learner model is given in Table 6.4. The features' names are self-explanatory. Our dataset contains queries and the number of retrieved results for each query. So, using our limited dataset, we try to find reasonable features which are in Table 6.4.

6.2.2 Experimental Setting

We sample queries from our 11K dataset. There are 665 NAQs and 11008 regular queries in July 2011 and 873 NAQs and 10800 regular queries in January 2012. For training, we down sample regular queries such that the train set contains similar number of NAQs and regular queries to prevent the class imbalance in the training set. While testing the model, we use the original distribution. Due to the high class imbalance in the testing set, we report the performance using the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the true positive rate versus the false positive rate. We also report the Area Under Curve (AUC) as a summary of the performance of the classifier.

6.2.3 Experimental Results

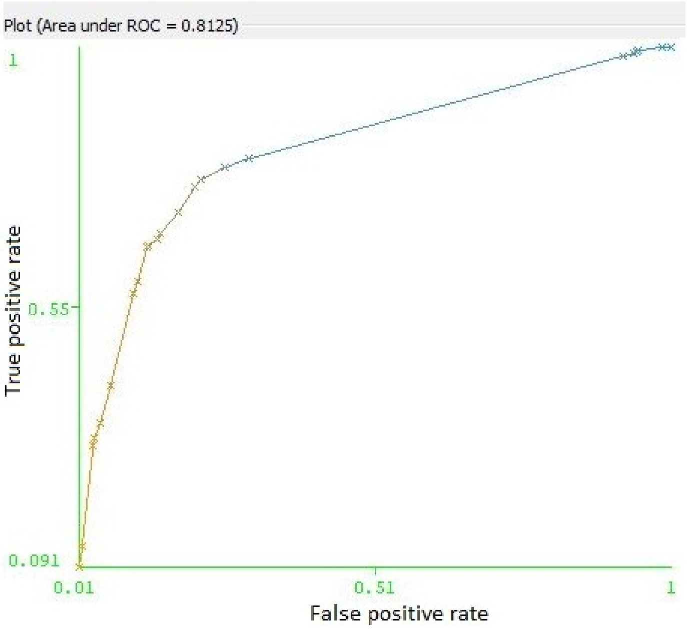

Figure 6.1: ROC curves for the feature set in Table 6.4 - July 2011

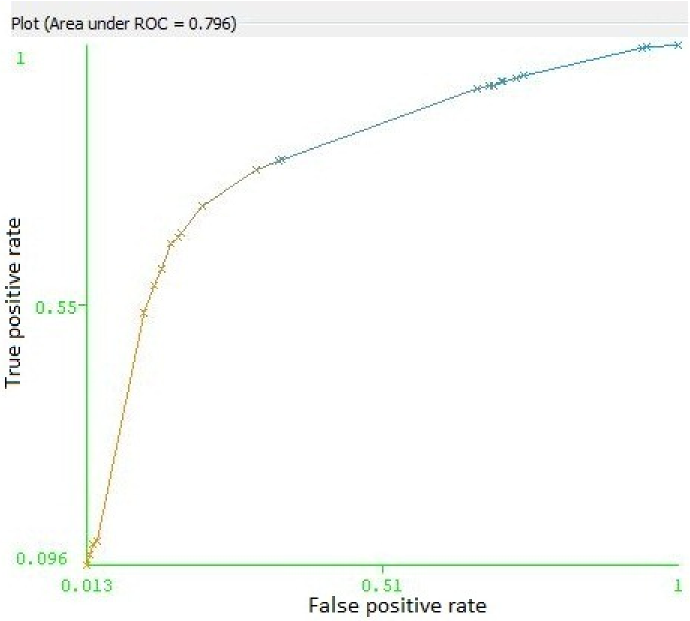

Figure 6.2: ROC curves for the feature set in Table 6.4 - January 2012

We evaluate the performance for the feature set described in Section 6.2.1. The results which belong to July 2011 are summarized in Figure 6.1, which shows the ROC curve. The classifier that uses the features set is able to produce good classification results. AUC is 0.8125 here. If the AUC is close to 0.5, it means that the classifier performs close to a random assignment of classes. So our results are fairly good. In January 2012, AUC is 0.796 which can be seen from Figure 6.2. The result is a bit worse than the result which is taken from July 2011, but it is still fairly good. From both July 2011 and January 2012 results, these are a positive finding, given that the distribution of NAQs in the test set is highly skewed and it implies that the features extracted are useful for classifying NAQs. This assessment is important because such queries are difficult and the search engine might want to be informed so that it can proactively suggest a reformulation of the query to the user.

Chapter 7

Conclusion

In this thesis, we aim to mine web search engine results and understand how web search engines handle the hard queries that can match very few or no results. Throughout the chapters, we first present introductory information about web search engines. Then, we provide a discussion on the related works and the background information. Then, we make an analysis about hard queries and present our experimental study. Following that, we focus on No Answer Queries (NAQs) as a subset of hard queries and lastly, we introduce our machine learning model to predict query suggestion patterns and NAQs.

From a general point of view, we compare the behavior of three search engines against hard queries using query logs obtained at two different times. To the best of our knowledge, there exist very few works that has focused on characterizing or classifying hard queries. We provide a characterization of hard queries that retrieve few or no results in web search engines. After a detailed analysis on how such queries are handled by the search engines, we focus on NAQs. We devise a number of features that can be used to identify such queries. Based on these features, we provide two machine learning models to predict query suggestion patterns of search engines and to predict the NAQs. Our experiments with public query logs show that, although dealing with the NAQs is difficult, their prediction is a relatively easier problem.

We note that most commercial search engines apply some techniques to avoid or solve hard queries. Still, there is a non-negligible volume of such queries, which implies that there is a need for shedding light on the characteristic of them. In the future, we would like to extend this work by providing a detailed analysis of hard queries from a Turkish query log and focus on Turkish NAQs. For this purpose, we are planning to use four search engines Bing, Google, Yandex and Yahoo!. Our future work also involves investigation of techniques that may help generating results for NAQs.

Bibliography

[1] A. Broder, “A taxonomy of web search,” *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

[2] M. Henzinger, “Combinatorial algorithms for web search engines: three success stories,” in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA ’07, pp. 1022–1026, Society for Industrial and Applied Mathematics, 2007.

[3] W. W. Anderson, M. P.; Woessner, *Applied Groundwater Modeling: Simulation of Flow and Advective Transport.* Academic Press, 1992.

[4] D. Fallows, “Search engine users.” `http://www.pewinternet.org/Reports/2005/Search-Engine-Users.aspx`, June 2012.

[5] R. Baeza-Yates, C. Hurtado, and M. Mendoza, “Improving search engines by query clustering,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 12, pp. 1793–1804, 2007.

[6] Q. Chen, M. Li, and M. Zhou, “Improving query spelling correction using web search results,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 181–189, Association for Computational Linguistics, June 2007.

[7] S. Cucerzan and E. Brill, “Spelling correction as an iterative process that exploits the collective knowledge of web users,” in *Proceedings of EMNLP 2004*, pp. 293–300, 2004.

[8] comScore, "Baidu ranked third largest worldwide search property in dec 2007." `http://www.comscore.com/Press_Events/Press_Releases/2008/01/Baidu_Ranked_Third_Largest_World_Wide_Search_Engine/`, June 2012.

[9] G. Pass, A. Chowdhury, and C. Torgeson, "A picture of search," in *Proceedings of the 1st international conference on Scalable information systems*, InfoScale '06, ACM, 2006.

[10] R. W. White, M. Richardson, M. Bilenko, and A. P. Heath, "Enhancing web search by promoting multiple search engine use," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 43–50, ACM, 2008.

[11] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg, "What makes a query difficult?," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 390–397, ACM, 2006.

[12] R. W. White and S. T. Dumais, "Characterizing and predicting search engine switching behavior," in *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pp. 87–96, ACM, 2009.

[13] P. Bailey, R. W. White, H. Liu, and G. Kumaran, "Mining historic query trails to label long and rare search engine queries," *ACM Trans. Web*, vol. 4, no. 4, pp. 15:1–15:27, 2010.

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

[15] S. M. Beitzel, E. C. Jensen, D. D. Lewis, A. Chowdhury, and O. Frieder, "Automatic classification of web queries using very large unlabeled query logs," *ACM Trans. Inf. Syst.*, vol. 25, no. 2, 2007.

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

[17] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of web queries," *Inf. Process. Manage.*, vol. 44, no. 3, pp. 1251–1266, 2008.

[18] E. Gabrilovich, A. Broder, M. Fontoura, A. Joshi, V. Josifovski, L. Riedel, and T. Zhang, "Classifying search queries using the web as a source of knowledge," *ACM Trans. Web*, vol. 3, no. 2, pp. 5:1–5:28, 2009.

[19] D. Downey, S. Dumais, and E. Horvitz, "Heads and tails: studies of web search with common and rare queries," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 847–848, ACM, 2007.

[20] B. J. Jansen, A. Spink, and S. Koshman, "Web searcher interaction with the dogpile.com metasearch engine," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 5, pp. 744–755, 2007.

[21] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," in *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pp. 77–86, ACM, 2009.

[22] R. Jones, B. Rey, O. Madani, and W. Greiner, "Generating query substitutions," in *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pp. 387–396, ACM, 2006.

[23] M. Mitra, A. Singhal, and C. Buckley, "Improving automatic query expansion," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 206–214, ACM, 1998.

[24] R. Kraft and J. Zien, "Mining anchor text for query refinement," in *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pp. 666–674, ACM, 2004.

[25] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query recommendation using query logs in search engines," in *Proceedings of the 2004 international conference on Current Trends in Database Technology*, EDBT'04, pp. 588–596, Springer-Verlag, 2004.

[26] B. Martins and M. J. Silva, "Spelling correction for search engine queries," in *In Proceedings of EsTAL-04, Espaa for Natural Language Processing*, 2004.

[27] Q. Mei, D. Zhou, and K. Church, "Query suggestion using hitting time," in *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pp. 469–478, ACM, 2008.

[28] R. W. White and G. Marchionini, "Examining the effectiveness of real-time query expansion," *Inf. Process. Manage.*, vol. 43, no. 3, pp. 685–704, 2007.

[29] X. Wang and C. Zhai, "Mining term association patterns from search logs for effective query reformulation," in *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pp. 479–488, ACM, 2008.

[30] V. Dang and B. W. Croft, "Query reformulation using anchor text," in *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pp. 41–50, ACM, 2010.

[31] A. M. Lam-Adesina and G. J. F. Jones, "Applying summarization techniques for term selection in relevance feedback," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pp. 1–9, ACM, 2001.

[32] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 4–11, ACM, 1996.

[33] H. Daumé, III and E. Brill, “Web search intent induction via automatic query reformulation,” in *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short ’04, pp. 49–52, Association for Computational Linguistics, 2004.

[34] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang, “Clustering user queries of a search engine,” in *Proceedings of the 10th international conference on World Wide Web*, WWW ’01, pp. 162–168, ACM, 2001.

[35] B. M. Fonseca, P. B. Golgher, E. S. De Moura, B. Pôssas, and N. Ziviani, “Discovering search engine related queries using association rules,” *J. Web Eng.*, vol. 2, no. 4, pp. 215–227, 2003.

[36] J. Xu and W. B. Croft, “Improving the effectiveness of information retrieval with local context analysis,” *ACM Trans. Inf. Syst.*, vol. 18, no. 1, pp. 79–112, 2000.

[37] J.-R. Wen, Q. Li, W.-Y. Ma, and H.-J. Zhang, “A multi-paradigm querying approach for a generic multimedia database management system,” *SIGMOD Rec.*, vol. 32, no. 1, pp. 26–34, 2003.

[38] D. Beeferman and A. Berger, “Agglomerative clustering of a search engine query log,” in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD ’00, pp. 407–416, ACM, 2000.

[39] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, “Real life information retrieval: a study of user queries on the web,” *SIGIR Forum*, vol. 32, no. 1, pp. 5–17, 1998.

[40] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, “Analysis of a very large web search engine query log,” *SIGIR Forum*, vol. 33, no. 1, pp. 6–12, 1999.

[41] M. Kamvar, M. Kellar, R. Patel, and Y. Xu, “Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices,” in *Proceedings of the 18th international conference on World wide web*, WWW ’09, pp. 801–810, ACM, 2009.

[42] B. J. Jansen, A. Spink, and T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the web," *Inf. Process. Manage.*, vol. 36, no. 2, pp. 207–227, 2000.

[43] B. J. Jansen and A. Spink, "How are we searching the world wide web?: a comparison of nine search engine transaction logs," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 248–263, 2006.

[44] C. M. Eastman and B. J. Jansen, "Coverage, relevance, and ranking: The impact of query operators on web search engine results," *ACM Trans. Inf. Syst.*, vol. 21, no. 4, pp. 383–411, 2003.

[45] M. Bendersky and W. B. Croft, "Analysis of long queries in a large scale search log," in *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pp. 8–14, ACM, 2009.

[46] G. Kumaran and V. R. Carvalho, "Reducing long queries using query quality predictors," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 564–571, ACM, 2009.

[47] N. Balasubramanian, G. Kumaran, and V. R. Carvalho, "Exploring reductions for long web queries," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 571–578, ACM, 2010.

[48] S. Huston and W. B. Croft, "Evaluating verbose query processing techniques," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pp. 291–298, ACM, 2010.

[49] Bing, "Search engine - bing." `http://www.bing.com/`, July 2011.

[50] Google, "Search engine - google." `http://www.google.com/`, July 2011.

[51] Yahoo, "Search engine - yahoo." `http://www.yahoo.com`, July 2011.

[52] J. Bar-Ilan and B. Peritz, "The lifespan of informetrics on the web: An eight year study (19982006)," *Scientometrics*, vol. 79, pp. 7–25, 2009.

[53] F. McCown and M. L. Nelson, "Search engines and their public interfaces: which apis are the most synchronized?," in *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pp. 1197–1198, ACM, 2007.

[54] I. S. Altingovde, R. Ozcan, and O. Ulusoy, "Evolution of web search results within years," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 1237–1238, ACM, 2011.

[55] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[56] S. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 21, no. 3, pp. 660 –674, 1991.

[57] R. R. Bouckaert, E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "WEKA–experiences with a java open-source project," *Journal of Machine Learning Research*, vol. 11, pp. 2533–2541, 2010.

[58] B. Chidlovskii and U. M. Borghoff, "Semantic caching of web queries," *The VLDB Journal*, vol. 9, no. 1, pp. 2–17, 2000.