Computer Standards & Interfaces 34 (2012) 162–170

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

An analysis of web proxy logs with query distribution pattern approach for search engines

Mona Taghavi [a], Ahmed Patel [b,c,*], Nikita Schmidt [b], Christopher Wills [c], Yiqi Tew [b]

[a] Department of Computer, Science and Research Branch, Islamic Azad University, Tehran, Iran
[b] Department of Computer Science, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (The National University of Malaysia), 43600 Bangi, Selangor Darul Ehsan, Malaysia
[c] Centre for Applied Research in Information Systems, Faculty of Computing Information Systems and Mathematics, Kingston University, Penrhyn Road, Kingston upon Thames KT1 2EE, United Kingdom

ARTICLE INFO

Article history:
Received 1 July 2011
Accepted 18 July 2011
Available online 23 July 2011

Keywords:
Web search services
Search engines
Query analysis
Distributed search engines
Proxy server logs

ABSTRACT

This study presents an analysis of users' queries directed at different search engines to investigate trends and suggest better search engine capabilities. The query distribution among search engines that includes spawning of queries, number of terms per query and query lengths is discussed to highlight the principal factors affecting a user's choice of search engines and evaluate the reasons of varying the length of queries. The results could be used to develop long to short term business plans for search engine service providers to determine whether or not to opt for more focused topic specific search offerings to gain better market share.

1. Introduction

The rapid growth of user accesses to World Wide Web (WWW) and its applications as well as the increase in the amount of content create difficulty for information retrieval that leads to making the task of Web search highly critical in the face of very short response time and near exact search results against user specified queries. Current Web search engines are built to provide answers to all query requests, independent of the special needs of any individual user. Much of it is done by search engines prompting or directing the user to select websites. However, nowadays, search engines attempt to identify some of the user's intentions and suggest more precise or relevant key terms. Although, they have improved a great deal over the years, but their results are still far from perfect.

As a matter of fact, users reveal their private information about their current interests by submitting a search query. Analysis of this information enables search service providers to more or less *precisely* target their search features capabilities to users' needs.

The above mentioned gap and opportunities behind the undiscovered query's patterns motivated this research study has been to provide statistics on numerous aspects of user query behaviour, the distribution of queries over time and changing trends in user behaviour to investigate the problem of how to answer queries efficiently in the current competitive search engines marketplace.

The analysis provided in this study was carried in the context of a distributed search system for the Internet developed by the Adaptive Distributed Search and Advertising (ADSA) research project [1] as part of the advances in Web systems and Web robots/crawlers and aims to design advanced distributed search engines offering high-quality focused topic-specific document databases [2]. From a top-level architectural viewpoint, an ADSA system is a collection of components – Search Engines and Brokers – dispersed across the Internet as shown in Fig. 1 with the following most prominent properties:

- The system supports both document search and placement of advertisements for the purpose of revenue generation.
- Search engines are designed to be topic-specific in order to improve the system's focused target query handling and scalability. Therefore, a number of distributed focused Web robots form a key part of the ADSA system.
- Attribute-value based search facility gives users access to document structure when making search queries.
- Each ADSA system can be independently owned and managed autonomously in a federated cooperative yet competitive business environment.

In general, the distributed search engine systems consist of many search engines acting as one global search system. Each search engine

* Corresponding author at: Kingston University, United Kingdom.
E-mail addresses: mona.taghavi@gmail.com (M. Taghavi), whinchat2010@gmail.com (A. Patel), nikita.schmidt@gmail.com (N. Schmidt), ccwills@kingston.ac.uk (C. Wills), yiqi01@gmail.com (Y. Tew).

0920-5489/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csi.2011.07.001

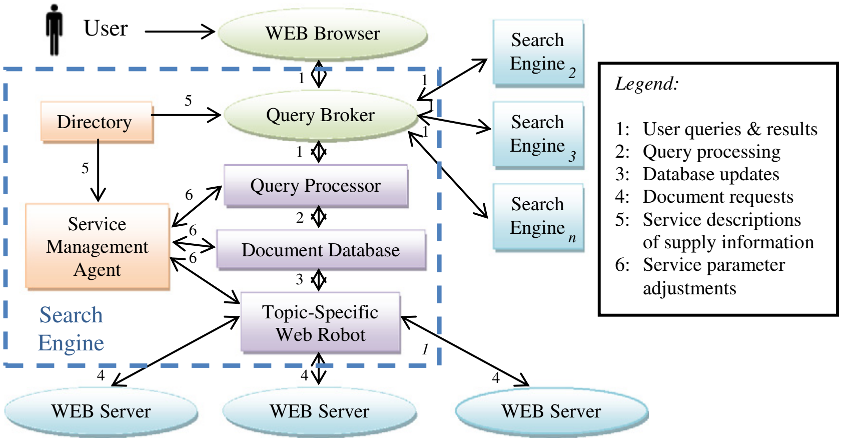


Fig. 1. ADSA architectural system view.

specialises in a selected topic(s) and is independently owned and controlled. Search engines compete for queries in the market by offering selection of topics and service costs through the service level management and agreements [3].

The rest of the paper is organized as follows. Section 2 describes the data collection method employed. Related works and new approaches are reviewed in Section 3. Section 4 analysis search engines and submitted queries comprehensively. Section 5 investigates server usage and Section 6 states the limitations and strengths of the study. Finally, conclusions are drawn in Section 7.

2. The available data and analysis

This research exercise is based on the data collected using Squid [1], a high-performance proxy caching server for Web clients. Web browsers can use the local Squid cache as a proxy HTTP server, reducing access time and bandwidth requirements. A large quantity of these Web proxy logs was provided by the academic research network technical support and management unit at the university computer centre. There were 33,100 files in total, constituting of 43 GB of data from three proxy servers. The data in these logs represent 39,631,832 queries which were directed to over 50 search engines during a period spanning more than 9 months from June 2010 to February 2011.

The log files provided for this work were derived from the files in Squid's *access.log* format. Each log file records requests to a server over a 24-hour period. Only three fields (out of 10 logged) per log entry were made available so as to maintain anonymity of the log files and to protect user privacy of data and information under appropriate data protection and privacy laws and directives:

- URL: The URL (Uniform Resource Locator) is a global address for specifying the location of a resource or a transaction. The user query is embedded in the URL.
- TIMESTAMP: The timestamp indicates when the query is submitted and logged in a UNIX format in millisecond granularity, since the standard epoch of 1 January 1970 UTC (Co-ordinated Universal Time) was introduced.
- ELAPSED TIME: The elapsed time field records how many milliseconds the transaction busied the cache.

3. Related works

The foundation for Web log analysis was initially established by Silverstein in 1998 [4]. He analysed a 280 GB AltaVista log file, consisting of approximately 1 billion entries for search requests over a period of six weeks from 2 August 1998 to 13 September 1998. This consisted of seven fields:

1. *Timestamp*, when a query was submitted in millisecond resolution.
2. A *cookie*, which can be used to say whether two queries came from the same user (this field is blank if the user has disabled cookies).
3. The *query* terms, exactly as submitted by the user.
4. The *result screen*, which is the requested range of search results.
5. Other *user-specified modifiers*, such as a restriction on the result pages' language or date of last modification.
6. *Submission information*, such as whether the query is a simple or advanced query.
7. *Submitter information*, such as the browser the submitter is using and the IP address of the submitting host.

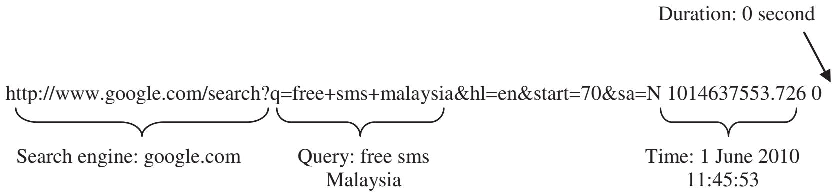
These are more than the aspects of user requests which are recorded in our experimental data set. It bears significant relevance to our study, in that it deals with user query requests in detail, albeit from only one search engine. Another analysis conducted by Zhang et al. [5] contributed significant measurement results on search engine transactional logs in time series. These two results provided useful statistics and suggested user patterns, which may be comparable to the results of our research study. Two obvious advantages of the Web proxy logs used in the present work over those previous studies are the sizeable collection of different search engine queries available and the longer time period which they span. The major findings of the study were that Web users mainly submit short queries and seldom modify the query. Furthermore, users mostly tend to look at the first ten results.

Logs from the Excite search engine have been made available to several different groups for research purposes as part of the Excite study on search engine query handling [6]. The findings from several of these research projects have been summarized by Spink and Xu [6]. The Excite research projects focused on different aspects of user behaviour, including the length of queries over time, a comparison of English and European language queries, multilingual searching, phrase searching and image searching.

Our work differs from these studies in so far as we analysed the distribution of search queries to a large number of different search engines based on the ADSA. Furthermore, we examined query patterns over a period of 9 months and hence established patterns of search over this time period.

3.1. Processing the logs

Fig. 2 shows a typical log entry. The URL field contains the most important data, namely, the base domain and the corresponding query string. The timestamp must be converted from UNIX format to a



```
http://www.google.com/search?q=free+sms+malaysia&hl=en&start=70&sa=N 1014637553.726 0
```

Fig. 2. A typical log entry with features for extraction/conversion highlighted.

human readable format. The elapsed time field usually consists of a “0”, meaning the log entry did not require access to the cache. For items that did access the cache, the elapsed time is given in milliseconds.

The records in the logs were processed to:

1. Extract the search engine and query.
2. Identify, count and remove robot-initiated queries.
3. Remove duplicate queries.

The criteria for this processing were derived from careful analyses of the log entries, combined with data on query formats used with the Analog log file analyzer [7]. The logs do not contain information to allow identical queries from different users to be reliably identified. Thus, we were unable to investigate term frequency matching in any precise way.

The log entries were broken down into *query-related* and *non-query-related* entries. Non-query related entries are those entries with no query functionality. There were also a substantial number of *robot-initiated queries* which were identified and removed wherever possible. Reliable identification of robot-initiated queries was difficult and had to be done by careful examination of the log files to create filters which removed these entries.

Of the 39,631,832 entries analysed, 61.9% of entries were query-related entries. Of these, 10.7% of entries were unique queries (unique URL/query combinations), 26.6% of entries were non-unique queries and 24.6% of entries were robot-initiated queries. Statistics regarding query length and query distribution were calculated from the unique queries.

4. Analysis of search engines and queries

In this section we present a more detailed description and discussion of the changing trends and driving factors of queries in search engines. The practical aspects and the statistics are analysed in order to provide a more comprehensive, comparative portrayal and characterization of the user behaviour from different perspectives:

- The changing trends in this distribution over time.
- The factors affecting this pattern of distribution.
- Regional domains.
- How many times a query contains one term, two terms, three terms, etc.
- The average query length.
- The trend of changing average query length over time.
- How users vary the length of a query depending on the search engine.
- A comparison of English language query length with European language query length.
- The changing trends in the above comparison over time.

The top search engine services combined were responsible for 99.23% of all the queries over the nine month period [8], [9]. Fig. 3 shows the query distribution among top search engines in some geographically separated countries.

As we can observe in Fig. 3, a clear majority of the queries of Internet search were directed to Google search engine because of its dominant position as a world number one leader in the search marketplace compared to its competitors [10]. However, it receives less attention in the USA in comparison with other countries. We also investigate possible reasons for the popularity of Google.com as the favourite search engine among the users who surf the Internet. In the mean time, Yahoo.com and Microsoft Bing are in close competition to engage the next most popular search service in the ranking position.

At present, the position of these three leading global search engine players, Google, Yahoo and Bing, is hard to contest. But, with the use of Web 3.0 or Semantic Web and the upcoming Web 4.0 [11] with much more advanced facilities and options like ontologies and user preferences trends can change. This may be a promising alternative which could offer an opportunity for smaller providers to gain market share through richer search engines and/or specific focused *topic* search engines.

4.1. Query distribution among search engines over time

The distribution of queries to search engines over the 9 month period has been analysed in order to identify usage trends. The analysis results provided some very interesting statistics. For the months of June to December, users did not vary their preference for search engines significantly. There were minor fluctuations in the percentage of queries submitted to the top three search engines, but no consistent pattern could be identified in these fluctuations. The percentage of queries to Google.com remained consistently around the 60% mark for these months. As a matter of fact, the breakdown of query distribution to search engines over the weeks of the year had shown the same consistency in user behaviour. This suggests that most users seldom change their search engine preferences.

Queries to Google fell from 62% in January to 27% in February. Queries to Yahoo and MSN increased by more than twofold as a percentage of the total. This change coincided with a large group of users ceasing to use the proxy servers and it appears that they had a strong preference for Google, causing a significant change in the

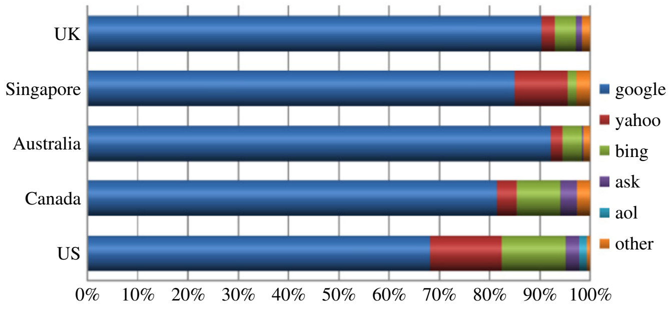


Fig. 3. Top search engines ranked by volume of searches[*] (Experian Hitwise [8] 21/05/2011[†]).

[*] The data samples featured are from the Hitwise Online Competitive Intelligence Service, which bases its daily insights on the online usage and search behaviour of more than 1.5 million Internet users.

[†] UK (Hitwise uk, 2011), Singapore (Hitwise sg, 2011), Australia (Hitwise au, 2011), Canada (Hitwise ca, 2011), US (Hitwise us, 2011).

distribution of queries when their activities were no longer recorded in the logs.

4.2. Factors affecting choice of search engine

The obvious explanation for the overwhelming popularity of Google search engine and the consistency in user behaviour is that users simply use their preferred search engine and that Google occupies this position for a majority of users.

A study carried out in 2001 by Chan et al. [12] and 2009 by White and Dumais [13] surveyed Internet users for their reasons behind search engine choice. The report concluded that search engines that focus on supplying useful search results were ultimately the most popular. On the other hand, the study by Zhaoli et al. [14] proposed that users' intention to use search engines should be separated from users' continuous intention to certain preferred search engines. They emphasized on user habits as an effective factor on users' preference that lead to users' continued intention to choose a particular search engine. However, our results suggest that groups of users may develop a preference for certain search engines, depending on their likes and dislikes as well as subject/topic selection, but this is a question worthy of further in-depth research which is outside the scope of this paper. It is worth pursuing.

4.3. Regional domains

The bulk of regional domains appearing in these logs were either ". my" or ".sg". One such site, AOL.ie, featured in one of the search engines. AOL does not have its own search portal; rather it exports queries to the Google enhanced search engine and imports the results onto its site. This setup is seemingly a good middle ground between quality of search and a feature-loaded site.

There was a small but consistent group of users who, over the period in question, submitted queries to the .my (Malaysia), .th (Thailand), .id (Indonesia), .sg (Singapore) and .ph (Philippines) regional domains of the Yahoo and Google search engines. This suggests that the user demographic consists not only of English speakers, but also of native speakers of these languages.

4.4. Query length

In assessing patterns regarding query length and changing behavioural trends over time, three main areas were focused on. Firstly, the occurrence of different term counts in queries was counted, i.e. how often a query contains one term, two terms, three terms, etc. Secondly, the average terms per query over the entire period was calculated, as well as a monthly average. Thirdly, an analysis of how users vary query length according to the search engine was carried out.

4.4.1. Term count occurrence in queries

On analysis of the query term lengths, the results were found to be contrastingly different to those of the Excite and AltaVista studies. The AltaVista study found that 25.8% of queries contained 2 terms, 25.8% contained just 1 term, and 15% contained 3 terms [4]. The statistics presented in Fig. 4 differ somewhat from these. While 2-term queries were the most common in both studies, AltaVista had a much higher occurrence of single term queries.

31.6% of queries contained more than 3 terms, in contrast to just 12.6% of the AltaVista logs. Ephemeral trends may be responsible for these contrasts, but possibly it is the different user demographic and various search engines under examination in this report. The lion's share of queries comprised one to six terms, with just less than 4% of queries containing more than six terms. The Excite study, in concurrence, found less than 4% of queries to contain more than six terms [6].

Queries with a single term were almost as common as queries with four terms, each accounting for just over 15% of queries. More dominant, however, were queries containing either two or three terms, together accounting for over 53% of queries. In contrast to this, about 62% of queries in the Excite study comprised either one or two terms.

According to the statistics created in this research, the percentage of single term queries has reduced dramatically. This indicates that users are getting fewer satisfactory results with just a single search term which broadens a competitive market for service providers in terms of serving the most suitable search result.

Our study does not provide an analysis of *boolean operator* usage in queries. Due to the variety of search engines analysed, it was technically infeasible to parse the various contrasting operators used by different search engines. However, the use of quotes (denoting term combination strings or phrases in a query) was tracked. It was found that 8.7% of queries contained quotes. The Excite study similarly found that one in sixteen (6.25%) queries contained quotes.

4.4.2. Average query length

The average terms per query in the logs, at 3.08 terms, were remarkably high. The AltaVista study [4] found an average of 2.35 terms, close to the average of 2.4 for Excite [6]. A more recent study done by Zhang et al. [5] has shown the average length of a query is about 2.9, and the query length does not change with the changing of time during the day.

So, query length average of 3.08 is not far-fetched due to the rapid increase in query terms. However, there are several factors giving rise to this. The Excite study reported that the average length of Excite queries has increased steadily over time [15]. In 1996, the average length for US, UK and European users was 1.5 terms. In 1999, the average for US and UK users was 2.6 and for European users 1.9.

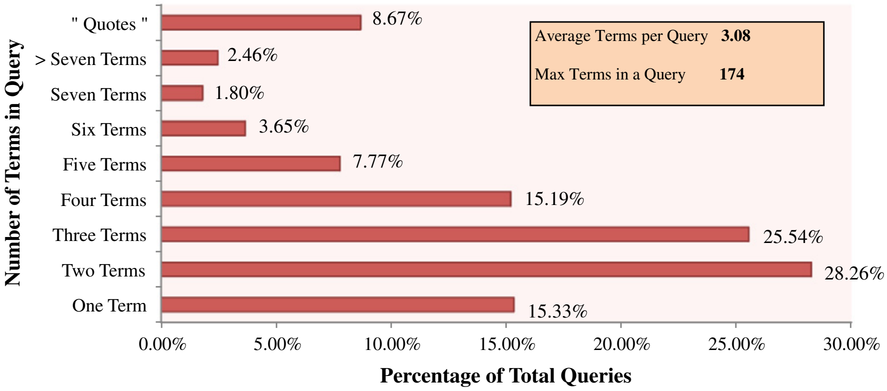


Fig. 4. Query term count.

Similar research is done by Hananzita and Kiran [16] on Malaysian Web search queries. The average length of English language queries has increased quicker than queries in other languages. This can be explained by the fact that the percentage of English language content on the Internet is greater than any other language.

There may also be a simple explanation for the universal increase in query length. As the Web grows larger and more diverse, Web users finds themselves having to refine their searches. This involves constructing more specific queries and, inevitably, adding more terms to a query in order to do so. Therefore, we can conclude that advancing search capabilities among users have increased the sophistication and length of their queries.

In order to distinguish between the average length of Google queries and the average length of other search engines, both were analysed separately. This was necessary because of the overwhelming influence of Google queries on the overall statistics. It was discovered that Google had a much higher average than most other search engines and was thus the cause of an imbalance in the overall average of 3.08 terms per query over the period. The average Google query comprised 3.32 terms, while the average of the outstanding collective was 2.74 terms. This explains the overall average of 3.08 terms.

It is clear that Google users have a tendency towards longer queries. Whether this is related to the information retrieval techniques used by the Google engine, or to behavioural patterns of the users, or to the large amount of data indexed by Google requiring users to be more specific in their queries, is unclear. Regardless, an average of 2.74 terms for the other engines sustains the hypothesis that the average length of queries has increased steadily over time. In order to test this hypothesis, the average query length was examined on a monthly basis.

4.4.3. Average query length per month

Surprisingly, the changing average query length per month in these logs somewhat contradict the hypothesis that the average length of queries is growing steadily over time. Fig. 5 illustrates the moving monthly average of Google query lengths and the aggregate average of other search engine query lengths. The percentage of total queries for that month is also drawn, in order to illustrate that there is little or no correlation between the monthly average terms per query and the amount of queries submitted in a month.

The average query length showed a steady decrease in size from the months of June 2010 to February 2011, with the exception of increases in the months of July 2010 and January 2011. The overall downward trend is evident for both the Google average and the collective average of other engines. This implies that the trend is not confined to a single search engine. The average query length decreased in size by 0.26 for Google and by 0.29 for others over the period.

It is difficult to say if this is an ephemeral trend or whether it reflects the rise and fall on the average over a longer time scale. Nevertheless, it may be concluded that the average query length has risen slowly over time, according to the comparison of the collective average with previous studies.

4.4.4. Average query length per search engine

The question of how users vary the size of a query, relative to the search engine they are using was another area of interest in the logs. Table 1 shows the average query size for a selection of search engines. These are broken down into three categories: popular search engines, foreign language engines and the Ask group of engines. The average query size differed greatly between search engines.

By visually inspecting queries submitted to Ask search engine, it was found that many users formulated their query specifically for the Ask search engine, e.g. “How do I wire a plug?” Consequently, Ask search engine had the highest average terms per query for any search engine in the logs. Since not every query can be formulated in this way, it is conceivable that some users vary their choice of search engine depending on the query they want to submit. Equally then, such users would tailor their queries according to the search engine they are using.

There were several other instances of search engines exhibiting unusual query length averages. *Dictionary.directhit.com* received just 1.78 terms on average. Examination of the type of queries it received, again showed that users of this search engine had a particular type of query in mind, e.g. “correlative”. *Infoplease.com* also received a low average of 1.9 terms per query.

Images.google.com was used consistently throughout the period examined. It received 2.37% of the total queries, and averaged 2.08 terms per query. The Excite study also found that provision of image searching is important to users and that relatively few terms are used in this type of non-text searching.

The aforementioned search engines cater for particularistic queries, i.e. they are designed to handle a specific type of query and return results based on that type of query. Traditional search engines cater for universalistic queries, i.e. they are designed to handle many different query formulations competently. The frequent use of

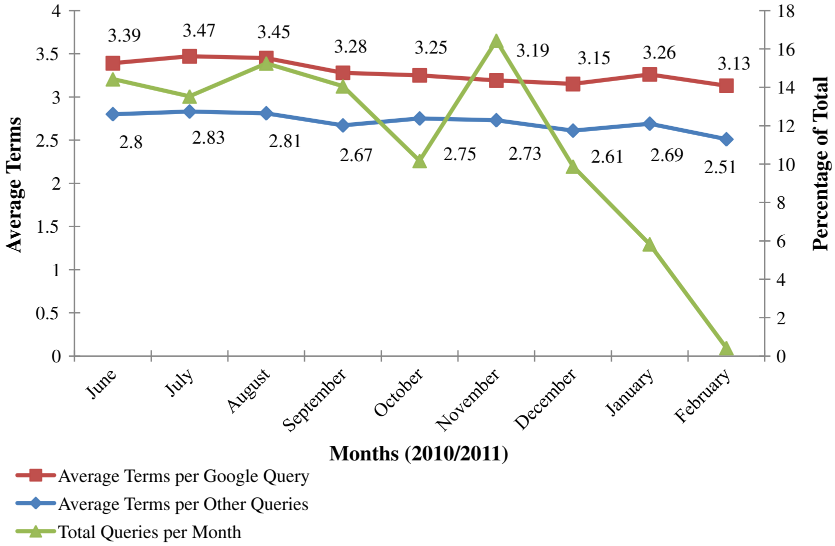


Fig. 5. Average monthly query length.

Table 1
Average query size of selected search engines.

Popular search engines		Non-English language engines		Ask search engine	
google.com	3.32	es.search.yahoo.com	2.83	ask.com	4.86
search.yahoo.com	2.89	fr.search.yahoo.com	2.46	uk.ask.com	3.89
ie.altavista.com	2.76	de.search.yahoo.com	2.03	dictionary.directhit.com received	1.78
bing.com	2.57	it.search.yahoo.com	2.14	infoplease.com	1.9
lycos.co.uk	2.48	kr.search.yahoo.com	1.41	images.google.com	2.08

particularistic search engines in these logs highlights the need for more development of search engine information retrieval. There is a demand for search engines that can accommodate different query formulations in different areas of search, providing results specific to the type of query. In future, search engines with more advanced capabilities than available at present might provide options that allow users to specify the type of results returned, based on the query formulation.

4.5. English and other language queries

Investigating the differences between queries in English and in other languages produced some significant results, which were compared with two other studies providing similar statistics. The Excite study, carried out in 1999, found a distinct contrast between English queries and queries in other European languages. English queries averaged 2.6 terms and European queries averaged 1.9 terms.

Spink et al. [15], collected query data from February to May of 2001. The average length of English queries was also found to be 2.6 terms, while European queries averaged 2.3 terms. As well, Chau et al. [17] conducted an analysis on Chinese search-log data and found that the average length of the Chinese queries was 3.38, which was larger than the mean number of terms in English queries as reported in Excite and AltaVista. They also reported a steady increase in the average length of queries according to previous findings by Pu, et al. [18] with the value of 3.18 terms.

The logs under analysis in this paper showed an average of 2.8 terms for English language queries and an average of 2.45 terms for European queries. Combining the three studies, an emerging trend over time can be seen in Table 2.

The quantity of English language content on the Web is still greater than any other language. Queries attempting to retrieve information in English from the Web tend to comprise more terms due to the greater detail required in the query to retrieve the pertinent results. However, terms and characters are different in each language and cannot be compared directly, but the steady increase in the average length of non-English queries over time indicates that the quantity of non-English language content on the Web has increased substantially in recent years [19].

5. Analysis of server usage

We also analysed requests on four different time scales:

- Hours of the day
- Days of the week
- Weeks of the year
- Calendar months

Table 2
Average length of English and European language queries over time.

	Excite study (1999)	U.S. vs. European study (2001)	Our study (2010/2011)
English language queries	2.6	2.6	2.8
European language queries	1.9	2.3	2.45

These results may be useful to the server administrators in planning downtime.

5.1. Server usage per hour of the day

Fig. 6 illustrates the percentage of requests made to the proxy server per hour of the day. The time is given as server local time. The period from midnight until 08:00 is clearly the least busy. Between 08:00 and 09:00, activity increases significantly, and this pattern of increase continues until around 13:00. The busiest period of the day is between 10:00 and 19:00. The peak hours of usage are between 14:00 and 16:00.

5.2. Server usage per day of the week

The busiest day of the week for the server is Tuesday as shown in see Fig. 7. Over 20% of requests were made on Tuesdays. Wednesdays and Thursdays each recorded 18.7% of server activity. Monday and specifically Friday are the least busy weekdays, whilst the weekend is by far the quietest time.

5.3. Server usage per week of year

The breakdown of usage over weeks of the year reveals an erratic graph in Fig. 8. The logs recorded transactions for 31 weeks from the start of June 2010 to the end of February 2011. In 2010, the National day in Malaysia occurred on 31 August, the last day of the week number 7. This explains why week 8 became the least busy of the entire period examined.

Week 29 (23–29 January) was the second least busy period. It coincides with the change of homepage settings on the network. This suggests that several network changes were in progress during this week and the network was not running at full capacity. Another noticeable period of relative inactivity was week 15 (17–23 October). The erratic pattern of server activity may be explicable not only by user behaviour patterns and public holidays, but also by a slow network, network maintenance or even network failures.

5.4. Server usage per month

The breakdown of server activity per month shows a reasonably steady pattern for the months of June 2010 to February 2011 inclusive as shown in both Figs. 8 and 9.

In October, there was a significant drop in requests to the server, followed by a surge in activity in November, which was the busiest month. Incomplete logs were available for part of January and most of February 2011, thus these two months cannot be considered as a reliable measurement of server activity.

5.5. Access to the cache

The vast majority of log entries contained the digit 0 in the elapsed time field. These entries busied the cache for 0 ms. Entries in the log requiring access to the cache accounted for 16.8% of total log entries. The time required to access the cache differed greatly between

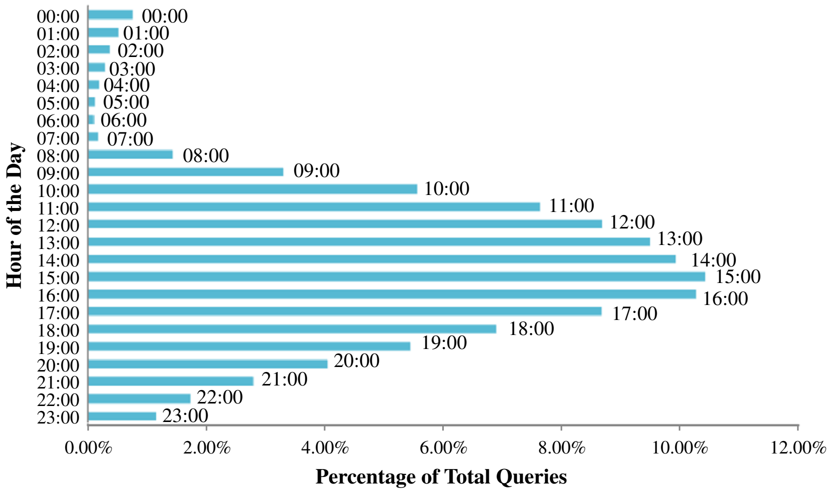


Fig. 6. Server usage per hour of the day.

entries. The factors affecting the elapsed time field include the size of the cached item and the network traffic at the time of access. The average time calculated from the elapsed time field for entries requiring access to the cache was 975.1 ms.

6. Limitations and strengths of the research study and analysis

The available data limited the analysis in a number of ways:

- The analysis presented deals with a small user demographic in a particular geographic location. It is therefore limited in the conclusions drawn.
- The available data set is rather small in comparison to other published studies.
- Owing to the absence of user distinguishing information in the logs, we did not examine query sessions, term frequency or query refinement.
- The analysis does not examine the use of Boolean operators in queries because of the complexity of parsing the various operators used by different search engines.
- Robot-initiated queries are handled in a somewhat simplistic way.

However, we believe the results are of interest because most papers on search engine query analysis focus on queries submitted to a single search engine, or to a few specific search engines. This work analysed the distribution of queries to a large number of different search engines. Some analyses are also limited by the time period they cover. We also identified query patterns over a period of over 9 months.

7. Conclusions

In this research study we attempted to identify key trends in queries to help search engine operators to assess the types of service desired by users in future, and to develop suitable infrastructure to meet future demands. We summarized major trends in queries and our recommendations are as follows.

- The average length of queries examined was somewhat longer than what other similar studies have found. It was concluded that the average length of queries has grown steadily over time. Since users are becoming more specific in their searches, search service providers need to keep adding increasingly advanced service capabilities in this competitive campaign.
- Average length of non-English languages queries had increased more than English queries. Thus, it is more advisable for search engine designers to extract potential search topics by analysing the search terms submitted by users, especially for non-English native users, rather than analyse the available texts online.
- English language queries compared to other natural languages still tend to comprise more terms as exemplified by the user of particular search engines such as Ask and Dictionary. Direct hits indicate that there is a demand for more advanced search engines which cater for

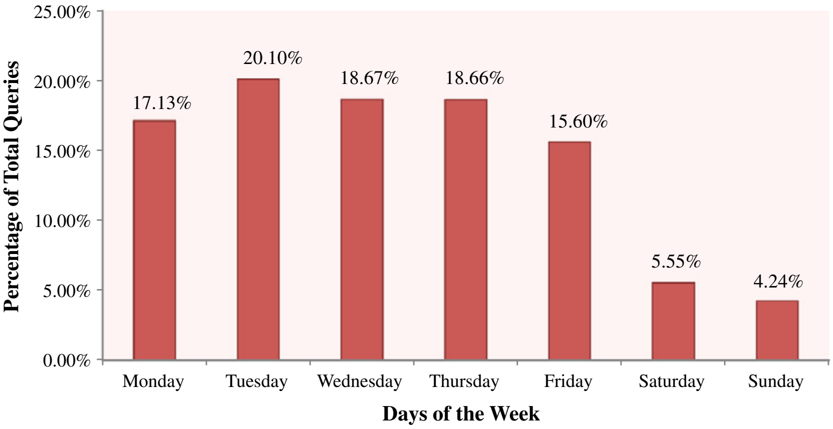


Fig. 7. Breakdown of server activity per day of the week.

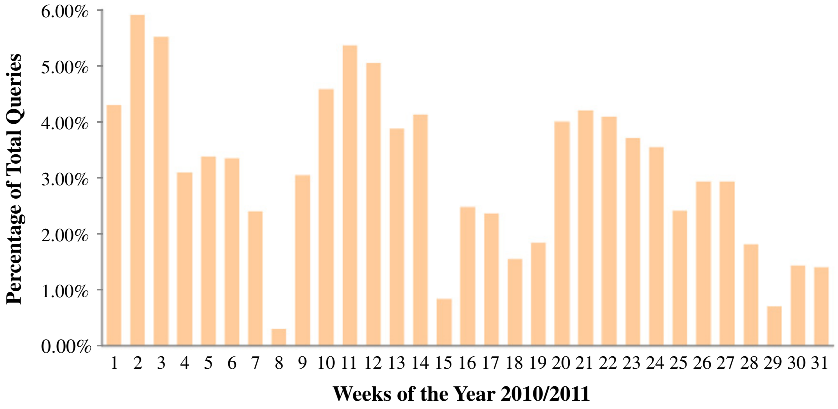


Fig. 8. Server activity per week of the year 2010/2011.

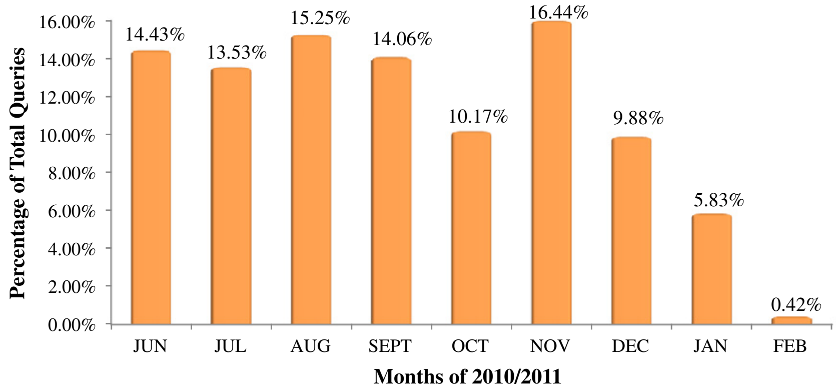


Fig. 9. Server activity per month of the year 2010/2011.

specific types of queries. This trend seems to be developing momentum.

- According to our findings, most users seldom change their search engine preferences. So the service provider who occupies this position would serve a majority of users.

Finally, further research and analysis work is needed, in particular:

- It would be interesting to investigate the issues of user sessions, query refinement and term frequency. This would require logs which include enough information to distinguish between users.
- The issue of how users choose different search engines for different types of search is also a very interesting one and merits future study. A survey of users' search engine preferences and querying habits, similar to that carried out by Chen et al. [12], would be of great value when combined with an analysis of Web proxy logs. The choice of a favoured search engine by groups of users is also worthy of in-depth investigation.
- Official Web searching websites should provide the complete query results on regional web search queries for research and education purposes. These results are suggested to be published in schedule to allow periodical studies for more accurate research results.
- More detailed analysis of robot-initiated queries and an analysis of Boolean operator usage in queries are also candidates for future work.

The analysis and ensuing identified trends from our research may help search engine operators to assess the types of services desired by users in future, and to develop suitable advanced search engines and supporting infrastructure to meet new and innovative demands. This is more likely to happen with the use of mashup meta-applications [20] based on Web 3.0 and Web 4.0 technology which cover mashup development and use of semantic web together with the use of rich ontologies.

Acknowledgement

All trademarks are acknowledged as the property of their respective owners.

References

[1] A. Patel, N. Schmidt, Application of structured document parsing to focused Web crawling, Computer Standards and Interfaces Journal 32 (7) (November 2010) x–y, doi:10.1016/j.csi.2010.08.002.

[2] A. Patel, An adaptive updating topic specific Web search system using T-graph, Journal of Computer Science 6 (4) (2010) 450–456 DOI:10.1.1.165.8503.

[3] A. Patel, M.J. Khan, Evaluation of service management algorithms in a distributed Web search system, Computer Standards & Interfaces 29 (2) (February 2007) 152–160, doi:10.1016/j.csi.2006.03.002.

[4] C. Silverstein, M. Henzinger, H. Marais, M. Moricz, Analysis of a Very Large AltaVista Query Log, Digital SRC Technical Note 1998–014Available from Internet:, ftp://gatekeeper.research.compaq.com/pub/DEC/SRC/technical-notes/SRC-1998-014.pdf1998(visited 13 May 2010).

[5] Y. Zhang, A. Spink, B.J. Jasen, Time series analysis of a Web search engine transaction log, Information Processing & Management 45 (2009) 230–245.

[6] A. Spink, J.L. Xu, Selected results from a large study of Web Searching: the Excite study, Information Research, Vol. 6 No. 1, 2000, Available from Internet:, http://informationr.net/ir/6-1/paper90.html, (visited 15 May 2010).

[7] Analog, The most popular log file analyser in the worldAvailable from Internet:, http://www.analog.cx2005(visited 16 May 2010).

[8] Experian Hitwise, Top websites & search engine analysisAvailable from Internet:, http://www.hitwise.com2011(visited 21 May 2011).

[9] SEO Consultants Directory, Top Ten Search Engine – Top 10 SEsAvailable from Internet:, http://www.seoconsultants.com/search-engines2010(visited 16 May 2010).

[10] V. Bhatiasevi, Y. Chairavutthi, The battle for World Wide Web dominance: in search of network externalities, International Business Management, 5, Medwell Journals, 2011, ISSN: 1993–5250.

[11] C. Marcus, Web 1.0, Web 2.0, Web 3.0 and Web 4.0 explainedRetrieved from:, http://www.marcuscake.com/key-concepts/internet-evolution2011(visited 16 June 2011).

[12] M. Chen, J. Dal Busco, K. Garrett, A. Sinha, A Search Engine UsageAvailable from Internet:, http://courses.ischool.berkeley.edu/i271a/f00/SearchEngine/appendix.htm2001(visited 13 October 2010).

[13] R.W. White, S.T. Dumais, Characterizing and predicting search engine switching behaviour, CIKM'09 Proceeding of the 18th ACM conference on Information and Knowledge Management, 2009, pp. 87–96, doi:10.1145/1645953.1645967.

[14] M. Zhaoli, G. Jiong, L. Guijun, Competition and adoption of search engine software, International Journal of u- and e-Service, Science and Technology 2 (1) (2009).

[15] A. Spink, S. Ozmutlu, H.C. Ozmutlu, B.J. Jansen, US versus European Web Searching Trends", *ACM SIGIR Forum*, Vol. 36 No. 2Available from Internet:, http://www.acm.org/sigir/forum/F2002/spink.pdf2002(visited 18 June 2010).

[16] H. Hananzita, K. Kiran, Malaysian Web search engines: a critical analysis, Malaysian Journal of Library & Information Science, Vol.11, no.1, July 2006, pp. 103–122, Available from Internet:, http://eprints.um.edu.my/282/1/web_search_engines_kiran.pdf, (visited 18 July 2010).

[17] M. Chau, X. Fang, C.C. Yang, Web searching in Chinese: a study of a search engine in Hong Kong, Journal of the American Society for Information Science and Technology 58 (7) (2007) 1044–1054.

[18] H.T. Pu, S.-L. Chuang, C. Yang, Subject categorization of query terms for exploring Web users' Search interests, Journal of the American Society for Information Science and Technology 53 (8) (2002) 617–630.

[19] B. Cory, J. Rosie, R. Moira, The linguistic structure of English Web-search queries, Proceeding of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 1021–1030, DOI=10.1.1.141.5909.

[20] L. Na, A. Patel, R. Latih, C. Wills, Z. Shukur, R. Mulla, A study of mashup as a software application development technique with examples from an end-user programming perspective, Journal of Computer Science 6 (11) (November, 2010) 1406–1415, doi:10.3844/jcssp.2010.1406.1415.

Mona Taghavi, a.k.a. CMT-SWG, received her B.Sc. degree in Information Technology from Parand Islamic Azad University of Iran in 2007. Besides her involvement in several Iranian national ICT research projects, she had worked for an IT consulting and project managing company which was responsible for overseeing and preparing some of the technical reports for the Supreme Council of Information and Communication Technology (SCICT) of Iran programme. Currently, she is pursuing her MSc in Information Systems at Universiti Kebangsaan Malaysia and undertaking research in cooperation with Prof. Dr. Ahmed Patel in advanced secure Web-based information systems and Secure Mobile Agent-based E-Marketplace Systems. She has published 4 papers. She is a reviewer of papers for Computer Standards & Interface Journal.

Ahmed Patel received his MSc and PhD degrees in Computer Science from Trinity College Dublin (TCD) in 1978 and 1984 respectively, specializing in the design, implementation and performance analysis of packet switched networks. He is a Professor in Computer Science at Universiti Kebangsaan Malaysia. He is visiting professor at Kingston University in the UK. He has published over two hundred technical and scientific papers and co-authored several books. He is currently involved in the R&D of cybercrime investigations and forensic computing, intrusion detection & prevention systems, cloud computing autonomic computing, Web search engines, e-commerce and developing a framework and architecture of a comprehensive quality of service facility for networking protocols and advanced services. He is a member of the Editorial Advisory Board of the following International Journals: (i) Computer Standards & Interface, (ii) Information Management & Computer Security and (iii) Cyber Criminology.

Nikita Schmidt received his B.Sc. degree in Mathematics from St-Petersburg University in 1994, and his PhD degree in Computer Science from University College Dublin in 2004. He is a collaborator and visiting researcher at Universiti Kebangsaan Malaysia working with Prof. Dr. Ahmed Patel in the area of Web Technologies, Search Engines and Software Engineering since 2009. Nikita is a software developer in Sensorix LLC in St-Petersburg, Russia, specializing in low-power wireless networks. He is reviewer of papers for Computer Standards & Interface Journal. He has published 20 papers.

Chris Wills joined Kingston University in 1987, before which he worked as a management consultant for a range of clients including the Greater London Council whom he assisted in obtaining funding from the European Union for technology and employment initiatives in London. For a number of years he was the Director of Kingston University's Centre for Applied Research in Information Systems. Chris has managed and undertaken information systems and computing research and consulting projects, on behalf of a range of organizations including the Defence Evaluation Research Agency, the UK's MOD's Tri-Services, the Police Service, the Health Service, the Department for Transport, the International Association of Public Transport (UITP) and The Mass Transit Railway Corporation of Hong Kong. His specialist area of interest is that of software process in mission and safety critical systems and it is in this area of information, computing and communication systems that he has undertaken work for the Royal Navy, scoping the design of warship command and control systems. Chris is a Freeman of the City of London and a Livery Member of the City of London's Worshipful Company of Information Technologists. He has published well over 20 papers. He is active collaborator with Prof. Ahmed Patel on many research topics.

Yiqi Tew received his B.Eng (Hons) Electronics degree in Computer from Multimedia University in 2008, and his Master degree in Computer Science from Universiti Kebangsaan Malaysia in 2011. He is a researcher at University Kebangsaan Malaysia working with Prof. Dr. Ahmed Patel in the area of image processing, search engines and mashup application programming and tools since 2010. Yiqi is software developer in Ziji Teknologi and firmware system developer lead in Almond Technology in Malaysia, specializing in embedded system programming. He has published 4 papers.