informationresearch

Vol. 12 No. 1, October 2006

Contents | Author index | Subject index | Search | Home

Query transformations and their role in Web searching by the general public

Martin Whittle, Barry Eaglestone, Nigel Ford, Valerie J. Gillet and Andrew Madden
Department of Information Studies, University of Sheffield, Regent Court, 210 Portobello Street, Sheffield, S1 4DY, United Kingdom

Abstract

Introduction. *This paper reports preliminary research in a primarily experimental study of how the general public search for information on the Web. The focus is on the query transformation patterns that characterise searching.*
Method. *In this work, we have used transaction logs from the Excite search engine to develop methods for analysing query transformations that should aid the analysis of our ongoing experimental work. Our methods involve the use of similarity techniques to link queries with the most similar previous query in a train. The resulting query transformations are represented as a list of codes representing a whole search.*
Analysis. *It is shown how query transformation sequences can be represented as graphical networks and some basic statistical results are shown. A correlation analysis is performed to examine the co-occurrence of Boolean and quotation mark changes with the syntactic changes.*
Results. *A frequency analysis of the occurrence of query transformation codes is presented. The connectivity of graphs obtained from the query transformation is investigated and found to follow an exponential scaling law. The correlation analysis reveals a number of patterns that provide some interesting insights into Web searching by the general public.*
Conclusion. *We have developed analytical methods based on query similarity that can be applied to our current experimental work with volunteer subjects. The results of these will form part of a database with the aim of developing an improved understanding of how the public search the Web.*

change font

Introduction

This paper reports on preliminary research in an ongoing study of how, and how effectively, members of the general public search for information on the Web. The focus is *query transformation* patterns that occur in a search session. Thus, we hope to gain insights into essential generic strategies that may result in more, and less, effective Web-searching. Our long-term aim is to build an evidence-based model of effective searching, based on statistical and qualitative data, to inform the design of training and of intelligent adaptive search interfaces. In this initial phase, transaction logs from the Excite search engine dating from 2001 have been used to develop methods for analysing query transformations. Although not current, the Excite logs represent a typical record of queries from the time with many of the syntactic changes that will be of interest to us. Also, they are well known and extensively studied. The activity logs thus present us with core data for development of analytical methods that can be used later in conjunction with qualitative results in our own empirical studies with volunteer members of the general public. This paper describes the research method that we have developed through our study of these logs, and also presents analyses of the logs which provide new insights into Web searching by the general public.

Related work

A significant body of research into Web searching exists (for a review, see: Spink & Jansen 2004) and potential value is demonstrated by Jansen and McNeese (2005), whose empirically study has identified predictable search states when searchers are more likely to be receptive to assistance, which can be informed by knowledge of effective Web searching strategies. Also, such interventions can be supported by search-pattern classifier algorithms, such as those of Ozmutlu and Cavdur (2005).

Most studies have been quantitative, using transaction logs (e.g., Jansen *et al.* 1998; 2000; Huberman *et al.* 1998) or on-line surveys (e.g., Broder 2002; Spink *et al.* 1999). Some have investigated searching by the general public, including query reformulations (e.g., Spink *et al.* 2000 and 2001). From these, various characteristics have emerged. For example, transaction log analyses show that most users (51.8% in 1998, 51.2% in 2002 - Jansen *et al.* 2005) only enter one or two search terms; and most only look at one screen of results (85.2% in 1998, 72.8% in 2002; Jansen et al 2005). Changes are occurring rapidly as search engines improve and searchers become more adept at using them (Jansen *et al.* 2005). In Europe, a study has demonstrated a decline in query length and a reduction in session time between 2001 and 2002 (Jansen and Spink 2005). Indeed, analyses suggest that little is gained by increasing the complexity of a search (Jansen 2000). More recently an analysis of search engine transaction logs over time and geographical region (Jansen and Spink 2006) confirmed the predominance of simple queries, but also identified geographical and search engine specific differences. Specifically, use of query operators differs across search engines and greater searching competence was found to exist in the USA compared with Europe. In general, their study found that searchers now tend to reformulate queries rather than examining results pages beyond the first page, and consequently searchers view fewer results pages. Also, the nature of search topics is changing in balance (from entertainment more to information and commerce).

Such studies, with exceptions (e.g. Nicholas *et al.*, 2003), used anonymous Web search engine logs, which provide no data relating to relevance, personal characteristics of searchers, or qualitative insights into searchers' thinking behind their observable search behaviour. Consequently, only limited explanation is provided for the patterns of Web information seeking identified. Nevertheless, studies of this kind have sometimes been able to infer useful information about user behaviour (Chau *et al.* 2005) with consequent recommendations for search engine and Website design. Several recent studies have focused on automated topic identification and an examination of diurnal changes as a means of optimising search engine response (Ozmutlu 2006; Ozmutlu *et al.* 2004; Ozmutlu and Cavdur 2005).

A potentially powerful but largely unexploited tool in Web searching research is the use of abstract models to represent semantics of the search process. Two approaches, often combined, are synthesis of statistics-based models and of semantically enriched data models. Examples of the former include Bayesian networks to model successive search queries (Lau and Horvitz 1999) and Markov models to predict Web users' next moves (Zukerman *et al.* 1999). In contrast, Wildemuth (2004) uses zero-order state transition matrices and maximal repeating patterns (MRP) (Siochi and Ehrich 1991) to characterise search tactics. An example of the data modelling approach is the use of graphs to represent sequences of search actions and interactions, finding both combined and individual effects on search strategies and effectiveness of domain and Web experience (Hölscher and Strube 2000). They used graphical representations to summarise the action sequences of the participants in their study, and as one of a number of means of comparing the behaviour of experts and novices.

Studies to date have provided only limited qualitative understanding of Web searching behaviour, with very little focus on the general public. We have noted various approaches to modelling Web searching behaviour, but feel that the potential of data modelling approaches for developing semantically rich models is under-exploited.

A study of Web searching by the general public: data modelling issues

Research aims

The preliminary phase of our research into Web searching reported here is part of a larger study. The study aims to extend previous research by seeking greater understanding of *real* Web-based searching, and of searching by the general public. Specifically, we want to examine searching in natural contexts in relation to the searchers' genuine information needs, as opposed to experimental situations and researcher-assigned search topics. Our research is designed to employs a balance of quantitative and qualitative approaches to develop semantically rich models of the ways in which searchers transform their queries during a session.

In the initial phase reported here we have addressed the research questions: to what extent can different types of query transformation be identified by the automatic analysis of search engine logs and can the resulting transformations be used to identify any significant patterns or relationships relating to Web use? This initial work involved the analysis of an Excite search engine transaction log dating from 2001 containing 1,025,910 queries (previously studied by Spink *et al.* 2001). For operational reasons our results were limited to those 1,025,838 queries with fewer than 256 characters. These logs presented us with core data for the development of analytical methods that can be used later in conjunction with qualitative data collected through our own experiments with volunteer members of the general public.

Research method

Space limits allow only an overview the method of research used in this initial phase. Our method of research has been to inductively derive a taxonomy of syntactic query transformations. We have done this through the development and elaboration of a program to analyse and encode the Excite logs. The program transforms logs into a string of codes designed to be converted to graphical form in which nodes and edges respectively represent queries and query transformations. Pseudo-code for the heart of the program is presented in the Appendix. The advantages of this representation are that, it captures the dynamic nature of Web searching by providing a history of the changes the searcher makes, there is potential for elaborating the model to represent semantic explanations as well as syntactic query transformations, and the similarities between this representation and that of chemical compounds should allow us to exploit data mining techniques developed for chemoinformatics research to identify characteristic search patterns.

Our main assumption is that the target query can be treated as a modification of the most similar preceding query rather than any other in the list. If none of the earlier queries bears sufficient similarity to the current one then we have a criterion for defining an entirely new query that is syntactically distinct from the rest of the session. Modification could be by addition or subtraction of words, by a change in word order, a change in spelling, word endings or prefixes etc. Accordingly, the transformation algorithm scans those sequences of queries within the log that form a session; that is, queries submitted by a single user. For each query within a session, the most similar previous query in that session is identified, and the syntactical transformation from that previous query to the current query is analysed and coded. Thus, the key aspects of the analysis algorithm are the query similarity test and the method of analysing and encoding query transformations. Both are being developed inductively and are thus evolving to derive models of searching within a session that appear better to capture the semantics of a Web search. Initially, in the absence of qualitative data, we are seeking to identify only syntactical query similarities and transformations.

Query comparison

Our approach to *query comparison* is based on the use of similarity methods (Willett *et al.* 1998) to compare individual words and whole queries. Query comparison is based on word content using a routine that effectively recognises words with modified spellings that have essentially the same meaning as similar. This method allows users also to specify the sensitivity of query comparison by setting a word similarity threshold (WST) and query modification threshold (QMT). These, respectively, are based on the similarity of words in the queries being compared and the amount of modification needed to transform one into the other. Values of WST = 0.4 and QMT = 0.3 were found to give an optimum balance between discrimination and acceptance when compared with human decisions for queries, based on a sample of results from the logs. Although these values appear to be about the most favourable the choice is largely a case of judgment.

Query transformation analysis

The query transformation classification developed describes the key multiple alterations that may occur in a single query transformation. Transformations frequently occur in combination with a Boolean operator or with the addition of quotation marks that transform a string of words into a phrase. In addition, we found it useful to introduce a code to register any long time delays between queries as this gives some indication of a user's working practice. Accordingly, our coding system for query transformation involves up to four symbols. A primary code denotes the main transformation type and this may be supplemented by up to three supplementary codes. Two of these prefix the primary code and indicate changes involving a Boolean term and those involving quotation mark (") changes. A third symbol, the delay code "_", may be added to the end of the query transform description if there has been no activity for a prescribed time. In this work we have chosen this to be sixty minutes. The primary and supplementary codes used to denote these changes are listed in the Appendix. Using these, the transformation between the queries

- Query 1: isic conference
- Query 2: "isic conference" +Sydney

would be represented as BQC(1), since the second query is a conjoint modification of the first and also includes quotation marks and a Boolean symbol. The number in parenthesis indicates the most similar previously occurring query in the train, which in this case is trivially 1. If the logs showed that there was no further activity for at least one hour, a delay term would be added and the code would become BQC (1)_.

The routines outlined in the Appendix apply a series of tests to the pairs of queries found to be most similar in a session and code each transformation accordingly. Previous coding schemes have been limited to sequential transformations and have used a smaller set, e.g. Spink *et al.* (2000). The categories can also be related to those discussed by He *et al.* (2002). Their notion of *generalisation* is equivalent to a disjoint modification and *specialisation* is equivalent to a conjoint modification in our coding system. However, the inclusion of phrase marks and Boolean terms also specialise a search and were not considered by these authors. These authors also discuss *reformulation*" which is essentially split into four categories, S, s, W, and w, in our scheme.

An example is given in Figure 1, in which we represent a query session (by user 74) concerning nurse training courses. A table of the queries involved is given in Table 1.

qid0	Query
1	nursing careers
2	nurse training
3	nurse training programs in baltimore
4	nurse training programs in baltimore city
5	undergraduate nurse training programs in baltimore city, maryland
6-20	paid undergraduate nurse training programs in baltimore city maryland
21	undergraduate nurse training programs in baltimore city, maryland
22	nursing schools in baltimore in baltimore city maryland
23	undergraduate nursing schools in baltimore city maryland
24	free nursing schools in baltimore city maryland
25	free undergraduate schools in baltimore city maryland
26	free undergraduate nursind schools in baltimore city maryland
27	free undergraduate nursing schools in baltimore city maryland
28	paid undergraduate nursing schools in baltimore city maryland
29	paid undergraduate nursing schools in baltimore city maryland

Table 1: Queries from the Excite log for user id 74 used to construct the graph shown in Figure 1. The first column gives a relative query identification number, qid0, and queries 6-20 were textually identical.

Figure 1 shows the initial and final queries, together with the string of intermediate query transformation types. The transformations are represented both as a string of codes and also as a graph. All of the codes used are defined in Tables 1 and 2. Note that the same query can occur many times within the one session (in the example, the 5th and 21st iterations of the query are equivalent) and undergo different transformation. Consequently, the graph can be cyclic, as a transformation may produce a query used earlier in the session.

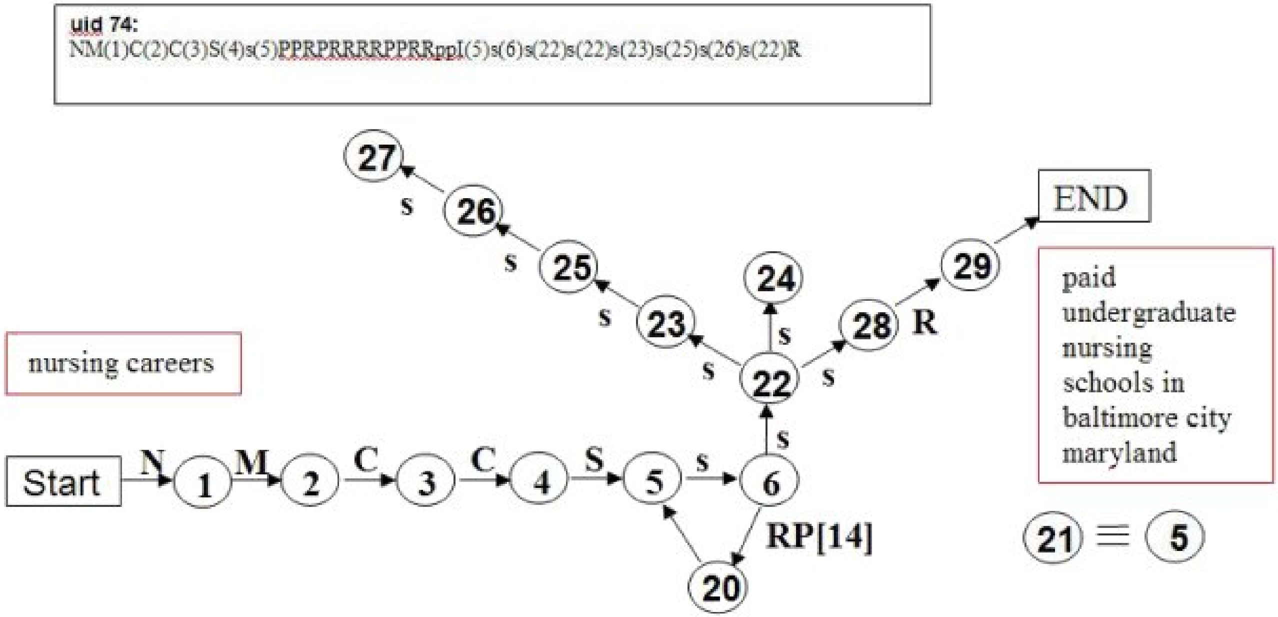
uid 74:
NM(1)C(2)C(3)S(4)s(5)PPRPRRRRPPRRppI(5)s(6)s(22)s(22)s(23)s(25)s(26)s(22)R
Start
N
1
M
2
C
3
C
4
S
5
s
6
RP[14]
20
s
22
s
23
s
25
s
26
s
27
s
24
s
28
R
29
END
nursing careers
paid undergraduate nursing schools in baltimore city maryland
21 ≡ 5

Figure 1: Example query transformation graph for user id 74 from the Excite log.
Note: The numbers assigned to each node signify a distinct query in a path starting and ending with those given explicitly. The string of codes shown in the box represent the transformations between maximally similar queries (Appendix) as the user modifies a search. These codes are also used to label the edges of the graph. In this case RP[14] is a condensed notation meaning a sequence of fourteen page viewing and repetition events relating to queries 6-20 in Table 1.

The string and graphical representations illustrated in Figure 1 are motivated by the use of similar constructions in the field of chemoinformatics. The strategy is to utilise data mining techniques developed in that field to determine characteristic search development patterns, thus building models of the different ways in which the general public conduct their searches and the associated effectiveness of these approaches. Having established a quantitative methodology based on a large sample of typical queries, the research has now progressed into a second phase, in which we are conducting our own empirical studies with volunteer members of the general public to validate and elaborate the models. The data collected in this phase are richer than is provided by transaction logs, since they includes complete screen records supplemented by qualitative data collected through interviews, talk-aloud protocols and questionnaires. Thus, the data provide a full record of all human-computer interactions, including, but not limited to, dialogue with the search engines used. In addition, the qualitative data provides insights into the semantics of and motivation for the search behaviour and its effectiveness. The richer data are being used to develop richer models of the search process, since the qualitative data allows us to also determine the intent behind query transformations and, hence, to derive a taxonomy of semantic query transformations.

Preliminary results

Occurrence of query transformation types

Figure 2 shows the occurrence frequencies of query transformation characters (listed in Table 1 of the Appendix). Each query transformation to the left of the vertical line consists of a main transformation type, which may be modified by one or more of the supplementary characters B, b, Q, q and _.

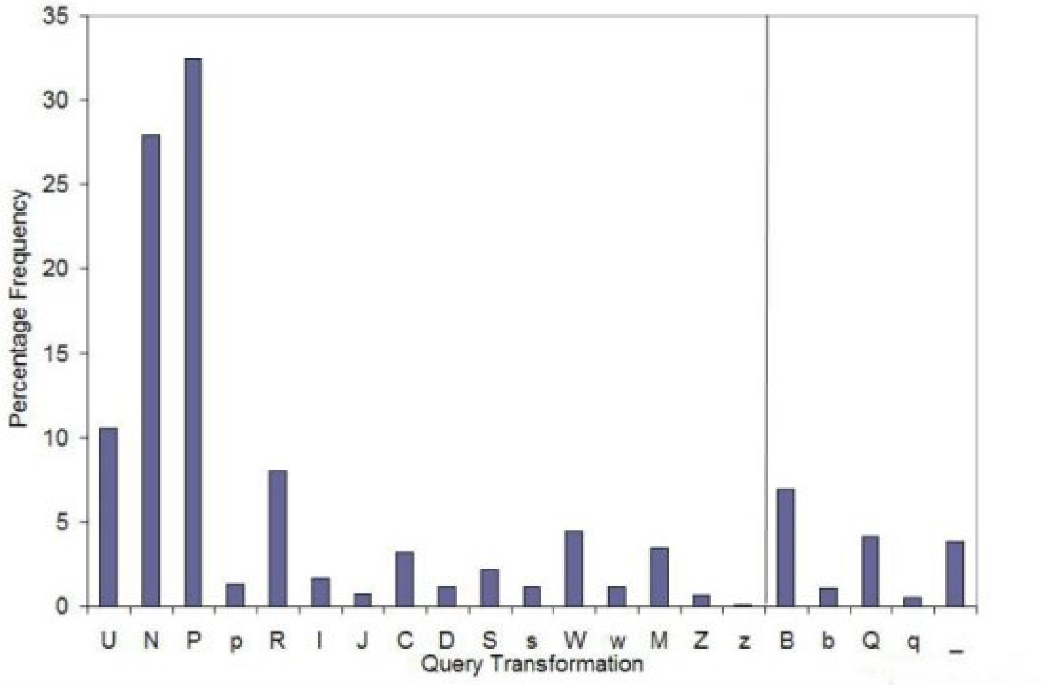

Figure 2. Query transformation percentage frequencies for QMT = 0.3 and WST = 0.4.
Note: The transformation codes are described in Table 1 of the Appendix. Classes U-z sum to 100% and results shown to the right of the vertical line refer to the supplementary characters for Boolean, quotation marks and delays. The occurrences of these are distributed amongst the main characters. A total of 1,025,838 queries was analysed.

The main features of Figure 2 can be summarised as follows. Forward page viewing, 'P', which He *et al.* (2002) have called Browsing, is the most popular activity and occurs more frequently than the submission of new queries, 'N'. Unique queries, 'U', form over 10% of entries and these are followed in popularity by repetition, 'R', which has also been interpreted as relevance feedback requests (Spink *et al.* 2000). Of the transformations of more significant interest, 'W' indicating a single common word, 'C' indicating the lengthening of a phrase and 'M' an unresolved textual similarity are the most common.

The graphs that we have examined in detail suggest that there is considerable variety in the patterns that may be obtained by these methods. The hope is that further analysis of these will aid the development of search models. At this stage we have examined the overall features of the graphs through their connectivity. Disregarding direction, query 22 in Figure 1 is connected to four other queries, queries 5 and 6 are connected to three, while other nodes are connected to just one or two neighbours. Amongst the 368,513 sub-sessions we found 165,578 that contained two queries or more and constituted viable graphs. By counting the k connections to each node for each sub-session length and accumulating the results in a histogram, we can obtain an average frequency distribution $f(k)$. Figure 3 shows a semi-logarithmic plot of this distribution of connectivity with the preferred parameters QMT = 0.3 and WST = 0.4. Many naturally occurring networks exhibit connectivity with a scale-free, power-law distribution characterised by relatively few highly connected nodes amongst a majority of low connectivity (Barabási and Albert 1999). However, the fitted lines in Figure 3 reveal that exponential scaling is obtained over much of the range following $f(k) \sim Ae^{\gamma k}$ with exponent $\gamma = 1$ and where A is a constant. For clarity, we show results for graphs of just two sub-session lengths in the range of interest, but other values have been sampled and show similar behaviour. The implication is that there is no preferential attachment of new queries to nodes that are already well subscribed (Barabási and Albert 1999). A possible influence on this result is the observed preponderance of simply connected page-viewing and repetitive queries. However, as shown in Figure 3, ignoring these trains in the connectivity count makes a significant difference only to values for *k* = 2 other points being virtually unchanged when viewed on the logarithmic scale.

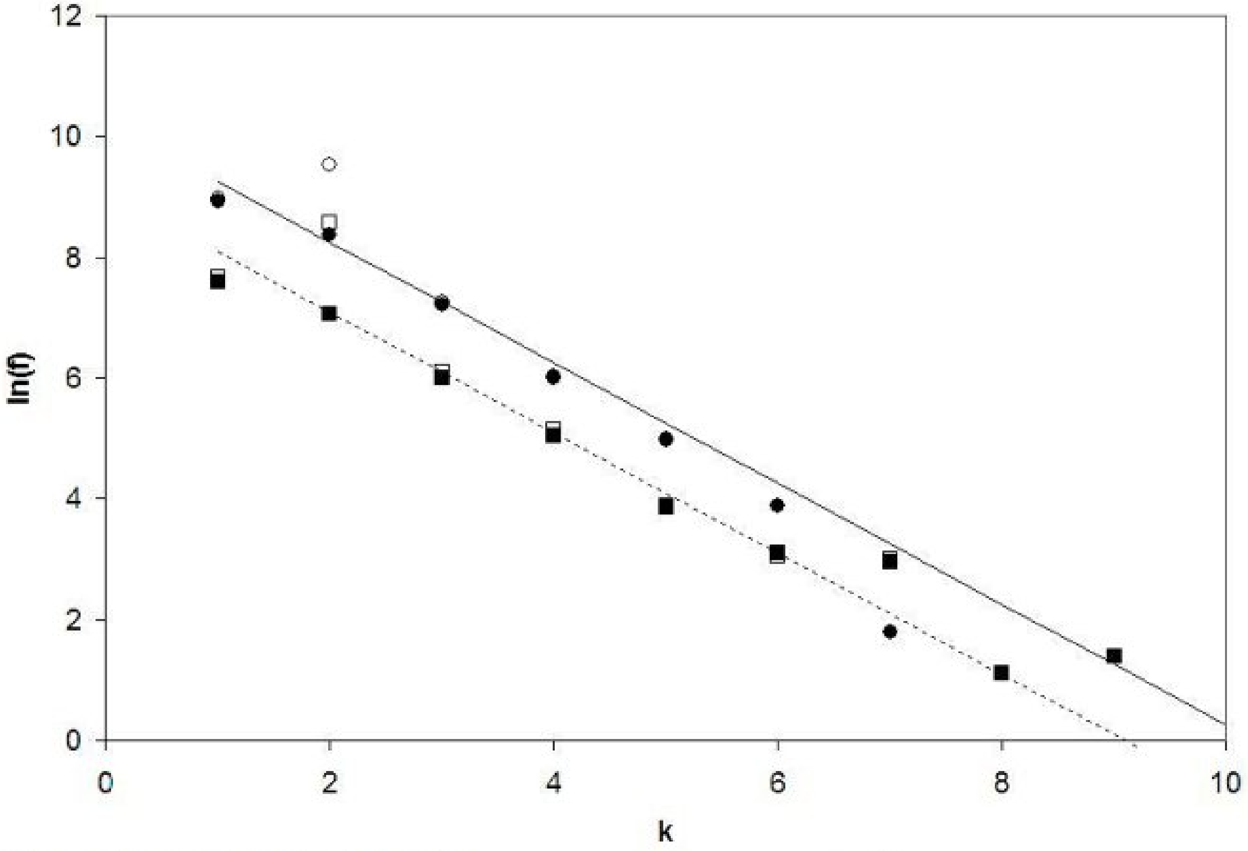

Figure 3. This natural logarithmic plot shows the frequency, *f*, of nodes with *k*-connections found in our graphical representations of query trains obtained with WST = 0.3 and QMT = 0.4.
The plot shows results for sub-sessions of length 10 (filled circles); and of length 20 (filled squares). Also shown, using open symbols are results obtained by excluding page-rank and relevance feedback trains: for lengths 10 (open circles); and length 20 (open squares). The lines have a slope of -1 and are fitted through the points corresponding to k = 3 in each case.

Correlation studies

As we have described, each query transformation code may contain up to four characters describing the transformation. An analysis of the correlations between these characters reveals information about the relationships between the supplementary codes and the primary codes relating to the construction of query transforms; these are the Intra-QT correlations. Following a standard correlation analysis (Hájek & Dupac 1967) we write the occurrence frequencies for these as $f(A_j , A_i)$ for pairs in which a character A_j is coincident (i.e. part of the same query transform) with character A_i. From the occurrence frequencies we evaluate expectation frequencies and an estimated standard deviation for the term correlations under the assumption that the codes appear randomly. We then compare the differences between observed and expected values expressed in multiples of the estimated standard deviation using the quantities D_f

$$D_f(A_j, A_i) = \frac{f(A_j, A_i) - E\{f(A_j, A_i)\}}{\sqrt{V\{f(A_j, A_i)\}}} \qquad (1)$$

Where *V* is the variance. Compared in this way, large positive differences in these suggest a strong association while negative values indicate that the codes concerned appear together less than expected. On the basis of values returned, values of greater than 50 suggest a strong association whilst values between 5 and 50 are significantly associated. Values between -5 and +5 were considered insignificantly associated and those less than -5 negatively associated. Results are given in Table 2

Type	B	b	Q	q	—
U	**20.6**	-	1.32	-	-
N	-1.48	-	**23.26**	-	**78.27**
P	-	-	-	-	-66.16
p	-	-	-	-	-9.63
R	-	-	-	-	**10.53**
I	-	-	-	-	4.45
J	**61.85**	**47.37**	**136.42**	**78.73**	-5.47
C	**46.02**	*-42.81*	-15.14	-19.22	-4.70
D	*-34.07*	**62.20**	-15.09	**13.45**	-4.79
S	*-24.52*	-11.14	*-20.69*	*-7.63*	*-5.65*
s	*-2.62*	**9.93**	*-7.05*	3.65	*-8.04*
W	*-35.00*	*-10.35*	*-32.99*	*-6.81*	-6.05
w	*-2.63*	**9.14**	*-11.51*	-0.98	*-8.18*
M	*-21.05*	*-12.98*	*-37.31*	*-13.28*	-1.97
Z	-2.26	**14.11**	*-10.06*	2.23	-0.90
z	1.78	2.82	0.55	1.45	0.95
B	*0.00*	-	1.16	**76.78**	*-15.01*
b	-	*0.00*	**74.95**	**10.05**	*-11.07*
Q	1.16	**74.95**	0.00	-	-0.28
q	**76.78**	**10.05**	-	0.00	*-7.77*

_	*-15.01*	*-11.07*	-0.28	*-7.77*	0.00

Table 2: Intra-QT deviations from expectation obtained for the Excite logs. In this case the '—' symbol refers to a 1hour delay. The results are printed for 5 ≤ D_f ≤ 50, and <u>bold underlined</u> for D_f > 50and in *italics* for D_f < 50.

Discussion

Table 2 records the values of D_f for correlations between codes within the same transformation. Using this measure, many of the results are significantly different from the random expectation.

The primary character, 'J' is, by definition, normally in association with one of the supplementary terms (the erroneous use of a single quotation mark may trigger it without an auxiliary character) and so the high values in row J of Table 2 are expected. The main conclusions that can be drawn from Table 2 can be summarised as follows:

- A Boolean term is relatively likely to be used with a unique query.
- Quotation marks are relatively likely to be used with a new query.
- There is a strong correlation (D_f = 46.02) between the use of Boolean terms and the character code "C", indicating that the inclusion of a Boolean term often involves a new word as well (the Boolean terms AND OR NOT are not counted as words). Similarly, there is a strong correlation between 'b' and 'D' - the removal of a Boolean term often includes the removal of word.
- Conversely, there is a negative correlation between 'B' and 'D', indicating that the inclusion of a Boolean term is rarely combined with the removal of a word. There is also a negative correlation between 'b' and 'C'.
- There is a somewhat more surprising correlation between 'B' and 'q' indicating that the simultaneous inclusion of a Boolean term and the removal of quotation marks is relatively common. Similarly the removal of a Boolean term 'b' and the addition of quotation marks 'Q' are also strongly linked.
- However, it seems that the inclusion of Boolean terms and the addition of quotation marks, 'B' and 'Q', appear together roughly as expected according to the random model.
- The delay code '—' is strongly associated with new queries (D_f = 78.27). Within a session several different enquires (sub-sessions) may be represented and some of these may be answered by a single query. It seems that there is then frequently a delay before the next enquiry is embarked on.
- The delay code is positively associated with relevance feedback, suggesting that this behaviour is being used to look again with search terms that have been successful.
- The delay code is negatively correlated with most other transformations particularly the inclusion of Boolean terms and page viewing. This suggests that these are part of active searching behaviour.

These observations include some that may be expected and some that are not. A similar analysis can be performed on the inter-QT correlations and this will be reported elsewhere.

Conclusions

In this paper we report on the initial phase of research into Web searching behaviour by the general public. By linking queries to the most similar preceding query in a session we have devised a system that will process a search engine log and convert individual sessions into a string of query transformations. Each transformation type is coded by up to four symbols and includes a connection number to the most similar preceding query. These can be converted into a graphical representation of the search process that could potentially reflect an operator's thoughts more accurately than could a simple relationship to the immediately preceding query. We have used this system to analyse the transformations obtained from an Excite search engine log in terms of their relative frequency and co-occurrence. A number of expected correlations have been identified and some that are unexpected. The connectivity of the graphs obtained has been investigated and found to follow an exponential scaling law, suggesting that their evolution does not involve preferential attachment to existing nodes.

The original motivation for developing this analysis was to apply it to our own logs derived from experimental records of searches by volunteers. This is now in progress. The codes can either be applied, as here, to link queries to the most similar previously occurring query, or to the immediately occurring query. Two sets of codes add an extra dimension to the richness of the data. These codes, representing the syntactic changes between queries, are being embedded in a database that includes volunteer background, scores from a cognitive style analysis, scores representing the subjects appraisal of a search, the queries themselves, Webpage titles visited as well as the keystroke record and activity timings. Volunteers are encouraged to talk through their searches and the temporal database will represent a flexible repository of data that can be used in conjunction with audio records of the volunteers' comments in a qualitative analysis.

The database approach has the aim of developing an understanding of the problem domain by elaborating a model that captures its semantics. Further, the inductive nature of this process means that this model will evolve, particularly given the progression from quantitative to qualitative analysis and the consequential progressive increasing in understanding of the semantics of the problem domain. Therefore, the applications for analysing snapshots of the data will benefit from views of the data that are persistent and independent of subsequent changes to the data model, i.e., data independence. Finally, we note that the phenomena that we wish to model, i.e., Web searches, are characterised by time. There is potential, therefore, to utilise the large body of research into temporal databases. In particular, we identify parallels between temporal data modelling requirements encountered in previous research into museum databases (Eaglestone *et al.* 1996)) and those of the current study, which is also an area of ongoing research into support for inductive research methods.

Acknowledgements

We thank Mark Sanderson and Amanda Spink for making the Excite logs available to us and the Arts and Humanities Research Council (formerly Board) for funding this project.

References

- Barabási, A-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**(5439), 509-512.
- Broder, A., (2002). A taxonomy of Web search. *ACM SIGIR Forum*, **36**(2). Retrieved 10 September, 2006 from http://www.acm.org/sigir/forum/F2002/broder.pdf.
- Chau, M., Fang, X. & Sheng, O. R. L, (2005). Analysis of the query logs of a Web site search engine. *Journal of the American Society for Information Science and Technology*, **56**(13), 1363-1376.
- Eaglestone, B., Holton, R. & Rold, L. (1996). GENREG: a historical data model based on event graphs. In *Proceedings of the 7th International Conference on Database and Expert Systems Applications*, (pp. 254-263). London: Springer-Verlag. (Lecture Notes in Computer Science, 1134) .

- Hájek, J. & Dupac, V. (1967). *Probability in science and engineering*. Prague: Academia.
- He, D., Göker, A. & Harper, D.J., (2002). Combining evidence for automatic Web session identification. *Information Processing and Management*, **38**(5), 727-742.
- Hölscher, C. & Strube, G. (2000). Web search behaviour of Internet experts and newbies. In *9th International World Wide Web Conference, Amsterdam, May 15-19, 2000. Conference proceedings*. Retrieved 10 September, 2006 from http://www9.org/w9cdrom/81/81.html
- Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E. & Lukose, R.M. (1998). Strong regularities in World Wide Web surfing. *Science*, **280**(5360), 94-97
- Jansen, B.J., Spink, A., Bateman, J. & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the Web. *SIGIR Forum*, **32**(1), 5-17.
- Jansen, B.J., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. *Information Processing and Management*, **36**(2), 207-227.
- Jansen, B.J. (2000). The effect of query complexity on Web searching results. *Information research*, **6**(1). Retrieved 1 February, 2006 from http://InformationR.net/ir/6-1/paper87.html.
- Jansen, B.J. & McNeese, M. D. (2005). Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *Journal of the American Society for Information Science and Technology*, **56**(14), 1480-1503.
- Jansen, B.J. & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, **42**(1) 248-263.
- Jansen, B.J., Spink, A. & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, **56**(6), 559-570.
- Jansen, B.J. & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, **42**(1) 248-263.
- Lau, T. & Horvitz, E. (1999). Patterns of search: analyzing and modelling Web query dynamics. In J. Kay (Ed.) *UM99, Proceedings of the Seventh International Conference User Modelling*, (pp. 119-128). New York, NY & Vienna, Austria: Springer-Verlag.
- Nicholas, D., Huntington, P. & Dobrowolski, T. (2003). Re-appraising information seeking behaviour in a digital environment. *Journal of Documentation*, **60**(1), 24-43.
- Ozmutlu, H. C. & Cavdur, F. (2005). Application of automatic topic identification on Excite Web search engine data logs. *Information Processing & Management*, **41**(5), 1243-1262.
- Ozmutlu, S., Spink, A. & Ozmutlu, H. C. (2004). A day in the life of Web searching: an exploratory study. *Information Processing & Management*. **40**(2), 319-345.
- Ozmutlu, S., (2006). Automatic new topic identification using muliple linear regression. *Information Processing & Management*, **42**(4), 934-950.
- Siochi, A.C. & Ehrich, R.W. (1991). Computer analysis of user interfaces based on repetition in transcripts of user sessions. *ACM Transactions on Information Systems*, **9**(4), 309-335.
- Spink, A., Bateman, J. & Jansen, B.J. (1999). Searching the Web: a survey of EXCITE users. *Internet Research: Electronic Networking Applications and Policies*, **9**(2), 117-128.
- Spink, A., Jansen, B. J. & Ozmultu, H.C. (2000). Query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policies*, **10**(4), 317-328.
- Spink, A., Wolfram, D. Jansen, M. B. J. & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*, **52**(3), 226-234.

- Spink, A. & Jansen, B. J. (2004). *Web search: public searching of the Web*. Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Wildemuth, B. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, **55**(3) 246-258.
- Willett, P., Barnard, J.M. & Downs, G.M. (1998). Chemical similarity searching. *Journal of Chemical and Information Computing Science*, **38**(6), 983-996.
- Zukerman, I., Albrecht, D.W. & Nicholson, A.E. (1999). Predicting users' requests on the WWW. In J. Kay (Ed.) *UM99, Proceedings of the Seventh International Conference User Modelling*, (pp. 275-284). New York,NY & Vienna, Austria: Springer-Verlag.

How to cite this paper

Whittle, M., Eaglestone, B., Ford, N., Gillet, V.J. and Madden, A. (2006). "Query transformations and their role in Web searching by the general public" *Information Research*, 12(1) paper 276 [Available at http://InformationR.net/ir/12-1/paper276.html]

Find other papers on this subject

Scholar Search Google Search Windows Academic

Appendix

Code	Query Transformation
U	A unique query. Only used for a single query session.
N	A new query (recognised by being at the start of a session or having low textual similarity to preceding queries) appearing during a session of at least two queries.
R	A repeated query with the same page rank - probably seeking relevance feedback.
P	A repeated query with increased page rank - further investigation of results from the current query.
p	A repeated query with reduced page rank - further investigation of results from the current query by returning to earlier pages.
I(*k*)	Indicates an identical query (including quotation marks and Boolean operators) to the one in the *k*'th position. This excludes identical queries in the immediately preceding position, which are covered by codes 'R', 'P' and 'p' in Table 1.
J(*k*)	Indicates an identical query apart from quotation marks, and/or Boolean + marks.
C(*k*)	A conjoint modification, which extends query *k* and retains it as a sub-phrase.
D(*k*)	A disjoint modification, which reduces a query *k* in which it is contained as a sub-phrase.

S(*k*)	A modified query that has a sub-phrase in common with query *k*; this sub-phrase is shorter than either query in question. The sub-phrase forms only part of the prior query.
s(*k*)	As S(*k*) above but in this case the number of words in common is also greater than the word-length of the common sub-phrase: an indication of some re-ordering of words, word insertion or removal.
W(*k*)	A modified query with a single word in common with query *k*. The single word forms only part of the prior query
w(*k*)	A modified query that has more than one word in common with query *k* but these are separated in one or both queries. Usually indicates an insertion or word replacement between common words.
M(*k*)	A modified query recognised on the basis of some textual similarity with a previous query above the threshold level. It cannot be further categorised as one of the above and probably contains changed word endings.
Z(*k*)	Queries not recognised as similar but found to have a single word in common with query *k*. This word is probably short, in comparison with the query length; frequently 'and' 'of' or 'in'.
z(*k*)	Queries not recognised as similar but found to have more than one word (as above, usually short) in common with query *k*.

Table A1. The following primary codes are used to describe the main features of a given modification (Where a code is followed by a value *k* in parenthesis, *k* records the sequential position of the most similar previous query. Where it is missing the code always refers to a transformation from the immediately preceding query or none at all)

Code	Query Transformation
B	Indicates the inclusion of a Boolean operator (+, AND, OR, NOT).
b	Indicates the removal of a Boolean operator.
Q	Indicates the inclusion of quote marks.
q	Indicates the removal of quotation marks
_	Delay term indicating prolonged inactivity prior to a subsequent query.

Table A2. The supplementary codes shown here may appear before (B,b,Q,q) or after (-) one of the primary transformation codes listed in Table A1.

Coding

The following pseudo code represents the algorithm used to generate the query transformation codes. It is essentially a double loop over the queries that retrieves the most similar pairs and then applies a series of tests on the pairs to establish a code. Details of the similarity methods used will be given elsewhere. These are represented here as T_i^B, T_i^Q for the Boolean, and quotation mark modifications, T_i^P, the primary code and, T_i^D, the delay code. We assume that n queries have previously been identified in a session. The symbol Q_i represents the ith query and Q_i^* the same query stripped of quotation marks and Boolean symbols. The corresponding page rank value, read from the logs, is represented by P_i. Coding symbols are shown italicised and the codes themselves are shown plain. Other terms are identified in the code or are self-explanatory.

Code A1 : Algorithm for code identification

Input: A set of *n* queries from a single user session.

Output: A set of four query transformation codes T_i^B, T_i^Q, T_i^P ,T_i^D for each query *i*.

Process:

```
for $i$= 1 to $n$ // Label first query transformation
$Q_i$ = a query
$T_i^B = f$ ;  $T_i^Q = f$ ; $T_i^P = f$ ;  $T_i^D = f$
end for
// Count and strip out quotation marks and Boolean terms from $Q_1$
$Q_1 => Q_1^*$
if($N_Q(1) > 1$) then $T_1^Q$ = Q
if($N_B(1) > 0$) then $T_1^B$ = B
if($n > 1$) then $T_1^P$ = N
if($n == 1$) then $T_1^P$ = U
for $i = 2$ to $n$ //  Loop over remaining queries
 //   First compare $Q_i$ with $Q_{i-1}$
if (timestamp($i$) - timestamp($i$-1) = delay limit) then $T_i^D$ = _
if ($Q_i$==  $Q_{i-1}$) then
```

```
if($P_i$ < $P_{i-1}$) then $T_i^P$ = p   //  Decreased page rank

if($P_i$ > $P_{i-1}$) then $T_i^P$ = P   //  Increased page rank

if($P_i$ == $P_{i-1}$) then $T_i^P$ = R //  Repeat query

else search for most similar previous query

//  Count and strip out quotation marks and Boolean terms from $Q_i$

$Q_i$ => $Q_i^*$

$N_B(i)$ = number of Boolean terms in query $i$

$N_Q(i)$ = number of quotation marks in query $i$

for $j$ = 1 to $i$-1

    Compare $Q_i^*$ with $Q_k^*$  and find similarity

    Record the most similar $Q_k^*$  set $k = j$

end for

 Compare  $Q_i^*$ with $Q_k^*$:

Obtain smallest word count = $W_{min}$

Count common words = $N_{com}$

Find largest common sub-phrase; word count = $MCS$

// Comparisons above similarity threshold QST

if (similarity($Q_i^*$ , $Q_k^*$)  >=  $QST$) then
```

Assign Boolean and quotation mark codes T_i^B , T_i^Q (Code A2)

```
if($Q_i == Q_k$)  then   $T_i^P$ = I              // Identical queries
elseif( $MCS = W_i$ ) then// Common phrase
    if( $W_j > W_i$ ) then  $T_i^P$ = C // Conjoint modification
    if( $Wj < W_i$ ) then  $T_i^P$ = D // Disjoint modification
    if($Wj = W_i$ ) then  $T_i^P$ = J// Identical apart from Boolean or quotation marks
endif
if( ($MCS = 0$)  AND  ($N_{com} > 0$))  $T_i^P$ = X // Error - should not happen
if($MCS = 1$) then
    If ($N_{com} > MCS$) then $T_i^P$ = w    // Single words in common
    If($N_{com} = MCS$)  then  $T_i^P$ = W         // Single common word
endif
if($MCS$ != 1) then
    if ( ($MCS$ !=> 0) AND ($MCS < W_{min}$)) then    // MCS shorter than shortest query
        if ($N_{com} > MCS$) then $T_i^P$ = s  // Common sub-phrase with re-ordering
        if ($N_{com} = MCS$) then  $T_i^P$ = S   // Common sub-phrase
    endif
endif
```

```
if ( ($MCS$ = 0) AND ($N_{com}$ = 0))  then  $T_i^P$ = M  // Similar queries but no common words or phrase

endif

 // Comparisons below similarity threshold QST

if( (similarity($Q_i$, $Q_k$) < QST) AND ($MCS$ > 0 ) )  then

        Assign Boolean and quotation mark codes $T_i^B$ , $T_i^Q$ (Code A2)

if ($N_{com}$ = 1) then $T_i^P$ = Z            //  One similar word

if ($N_{com}$ >1) then  $T_i^P$ = z                          //  More than one similar word

endif

end for                            // End loop over i
```

Code A2. Process assigning quotation mark and Boolean terms for a transformation between query i and an earlier query k.

Input: Quote mark and Boolean term count for queries *i* and *k*.

Output: Query transformation codes T_i^B ,T_i^Q for query*i*.

Process:

```
if($N_Q(i)$ - $N_Q(k)$  1) then $T_i^Q$ = Q        // Assign quotation mark codes

if($N_Q(i)$ - $N_Q(k)$ < -1) then $T_i^Q$ = q

if($N_B(i)$ - $N_B(k)$ > 0) then $T_i^B$ = B       // Assign Boolean codes

if($N_B(i)$ - $N_B(k)$ < 0) then $T_i^B$ = b
```

Contents | Author index | Subject index | Search | Home