Term Co-occurrence in Internet Search Engine Queries: An Analysis of the Excite Data Set

Dietmar Wolfram

School of Library and Information Science
University of Wisconsin--Milwaukee
P.O. Box 413
Milwaukee, WI 53201
Internet: dwolfram@uwm.edu

ABSTRACT

Unique queries submitted to the Excite search engine were analyzed for empirical regularities in the co-occurrence of search terms. The distribution of frequency of term pair occurrences was fitted to three models used in informetric studies to determine whether the pattern of term usage followed a Zipfian distribution. Relatively poor fits were obtained for two of the models tested, leading the author to conclude that the distribution is not Zipfian. Based on empirical distributions for term co-occurrences and terms submitted per query, two simulation models were developed to determine if binary dependence for specific query terms and combinations of term sizes were evident. A strong binary dependence relationship was observed for specific co-occurring terms. An analysis of co-occurrence based on term sizes revealed that the simulation model underestimated the co-occurrences of less frequently used terms and highly used terms. The implications for web-based IR system searching and design are discussed.

INTRODUCTION

Search engines provide Internet users with the ability to quickly locate sites containing information of interest. Every day the larger search engines process millions of submitted queries from users around the world. Despite the ubiquity of these technologies, very little research exists on how searchers actually make use of these services. This is largely due to the fact that few Internet search companies make available data from their engines for study. One of the few exceptions has been Excite, Inc., which has provided access to a large set of data from the Excite search engine. This data file consists of more than one million queries submitted to Excite on a single day. This wealth of data offers many areas of opportunity for analysis. The purpose of the present study is to perform an analysis of one facet of the Excite data--specifically, patterns of binary term co-occurrence within queries.

Several research studies have investigated co-occurrence of index terms within databases. Nelson (1983) found that simulation models that incorporated binary dependence of index terms within databases better modeled the distribution of term co-occurrences than a model that assumed binary independence. Peat & Willett (1991) examined the limitations of using term co-occurrence data for the automatic query expansion and concluded that similarity coefficients based on term co-occurrence used for automatic query expansion should only be used for the identification of alternative query terms that occur infrequently in the database. Wolfram (1996) used descriptor term co-occurrence to develop a simulation model for representing inter-record linkage structure in a hypertext bibliographic retrieval system where common occurrences of descriptor terms of records were used as the basis for inter-record linkages. The author found that a complex pattern existed for the observed system that could not be adequately represented by three different models. These studies have looked at term co-occurrence from the database side. At present, little research exists in this area from the query perspective, where patterns of search term co-occurrence may be useful in identifying the nature of searches and end-user search behaviours.

This study contributes to the small but growing area of Internet-based informetric research that examines regularities that exist in the numbers of Internet sites, overall growth of the Internet, and Internet user search behaviour (Huberman et al., 1998; Ingwersen, 1998; Lawrence & Giles, 1998). Most notable has been the recent research undertaken by Jansen, Spink & Saracevic (in press) which examined searcher behaviours using the Excite search engine based on a 50,000 query data set collected by Excite. The authors found that there were empirical regularities in the usage of query terms and the query size. They also found that fewer than 10% of queries submitted used Boolean operators.

The present study focuses on search patterns of end-users through the investigation of query term co-occurrence on a larger set of queries submitted to the Excite search engine. Query term co-occurrence is defined as the common occurrence of two search terms within a query.

Specific questions investigated include:

1. Is the distribution of query term co-occurrences, like many informetric phenomena, Zipfian?
2. Does binary dependence for query term co-occurrence hold true as it does for database terms?
3. What conclusions about user search habits can be drawn from co-occurrences of search terms?

A misconception of the search characteristics of Internet users is that the majority of users, assumed to be novice searchers, submit queries consisting of only a single term, thereby making study of query term co-occurrence data impractical. In fact, queries with two or more terms constitute roughly two-thirds of all submissions (Jansen, Spink & Saracevic, in press), making the analysis of query term co-occurrence quite feasible.

METHOD

Raw data from the Excite data set were imported into an MS Access database. The data included numeric identifiers for searchers/machines, time information on when each query was submitted, and the full queries entered by users. The 1,000,000+ queries represent a subset of queries submitted to the Excite search engine on a single day and only include submissions where an identifier, stored as a cookie on a user machine, was available. Queries submitted from machines on which the browser cookie facility was disabled were not included.

Queries were initially parsed for individual terms based on alphanumeric starting characters. Terms were delimited using spaces and other non-alphanumeric characters. Non-alphanumeric characters that were part of a URL or email address ('/', '.', '@', ':') were not treated as delimiters. Occurrences of the words 'and', 'or', and 'not' were treated as terms since the context of their use as Boolean operators or as parts of phrases was difficult to determine. Other modifiers allowed by Excite for inclusion or exclusion of terms (+, -) were ignored.

For this study, only unique queries were examined. The raw query data includes many examples of identical queries being submitted by the same searcher/machine identifier in succession. These submissions most likely represent the request to view the next page of hits for the query. Similarly, different searchers may have submitted the same query at different times. Again, multiple occurrences of the same query were not included, only unique queries were examined in this study to determine relationships between the co-occurrences of query terms. Of the original 1,025,910 queries in the raw data set, 363,282 were found to be unique. Based on the generated tables of queries and terms, the following data were collected: terms used per query, term frequency distribution, co-occurrence frequencies, queries submitted per identifier, and pages visited per query.

Term co-occurrences may only be studied using queries containing at least two terms. Jansen, Spink & Saracevic (in press), in a study of a 50,000 query Excite data set, reported that the average number of terms per query was 2.21terms, with the majority of queries containing 2 or more terms. Although the Excite search engine will only process the first 10 terms submitted in a query, a small percentage of the queries submitted in the present data set contain more than 10 terms. These larger queries were not truncated since they represent a more realistic account of actual searcher behaviour, regardless of how the search engine processes the queries.

The observed size-frequency distribution data for term co-occurrences were fitted to models used in informetrics (Zipf, Mandelbrot Zipf, Shifted Generalized Waring) to determine if the pattern of term co-occurrences follows a Zipfian form. Parameter values for each model were estimated using minimum chi-square estimation. An iterative routine relying on a quasi-Newton algorithm was used to estimate the parameter values that minimized the chi-square goodness-of-fit between the observed and the modeled data. The routines were re-run using different starting points to reduce the likelihood that outcomes represented local minima.

Observed query term co-occurrences were also compared to generated values from two simulation models to determine whether binary dependence holds true for specific term pairs and for terms of different frequencies of use. Binary dependence refers to the observed co-occurrence of two terms in a query that is greater than the product of the individual probabilities that they would co-occur. Each model focused on a different type of co-occurrence.

Model A - Term selection based on specific terms

Hypothetical queries were generated using the distribution of terms per query and the probability of a given term occurring within a query given its frequency of use. The model, by default, assumes binary independence of term co-occurrence. A computer simulation program was developed to generate hypothetical queries using the observed distribution of terms per query and the probability of occurrence of specific terms within a query based on their overall frequency of use. The number of hypothetical queries generated was equivalent to the total number of observed multi-term queries. The resulting distribution of simulated unique term pairs and their frequencies was compared with the observed distribution.

Model B - Term selection based on frequency of use

This model focused on the co-occurrence of terms based on their 'popularity' of use (i.e. how often terms of a given frequency of use co-occur). Are there patterns of co-occurrence based on the popularity or rarity of terms that result in co-occurrences that are more or less frequent than is probable? If so, where do the differences lie?

Another simulation program was developed to generate the same number of hypothetical queries based on the observed distribution of co-occurring frequencies of use or term sizes. The simulated co-occurrence distribution was compared to the observed equivalent using chi-square values for each cell of the observed and simulated term size co-occurrences.

FINDINGS

Figure 1 summarizes the distribution of terms per query for the unique query set[1]. The mode is two terms per query, with roughly two-thirds of all queries containing two or more terms, representing 292,994 unique multi-term queries. The unimodal distribution is highly skewed with 99.5 percent of the queries containing 10 or fewer terms.

A chart of the observed size-frequency distribution of term co-occurrences

[1] Zero terms in a query is also possible. These queries represent searchers who clicked on the search button without entering any terms, who selected the 'More Like This' option, or who entered one or more non-alphanumeric characters.

appears in the Figure 2. The 25 most frequently occurring term pairs appear in the Appendix. The list of term pairs shows that many of the most frequently occurring terms do co-occur with one another. Among the most frequently co-occurring pairs are Boolean operators and other non-semantic pairs of terms. Note that although 'not' only occurs 324 times, it co-occurs with 'and' more frequently than this number. This is due to the multiple occurrences of 'and' observed in some queries.

The distribution of term co-occurrences follows the classic reverse 'J' shape found in many informetric phenomena. However, results of the model fitting indicate that the observed data do not fit the models tested. With a large number of cells for the chi-square comparison, large chi-square values are not uncommon. The results are, nevertheless, highly significant.

TABLE 1
Model Fitting for Distribution of Term Co-occurrences

Model	Distribution & Parameters	X^2 value	d. f.
Zipf	$f(x) = \frac{a}{x^b}$ $x = 1,2, ..., Xmax$ $a = 0.699 \quad b = 2.285$	10,326 *	89
Mandelbrot Zipf	$f(x) = \frac{a}{(x+c)^b}$ $x = 1,2, ..., Xmax$ $a = 2.649 \quad b = 2.876 \quad c=0.613$	699 *	88
Shifted Generalized Waring	$f(x) = \frac{\Gamma(\nu+\alpha)\,\Gamma(x+\nu-1)\,\Gamma(x+\beta-1)}{B(\alpha,\beta)\,\Gamma(\nu)\,\Gamma(x+\nu+\alpha+\beta-1)\,(x-1)!}$ $x = 1,2, ... Xmax$ $\Gamma(*)$ and $B(*,*)$ represent the Gamma and Beta functions respectively $\alpha = 1.969 \quad \beta = 0.991 \quad \nu = 0.982$	1087 *	88

* indicates a significant outcome at .01

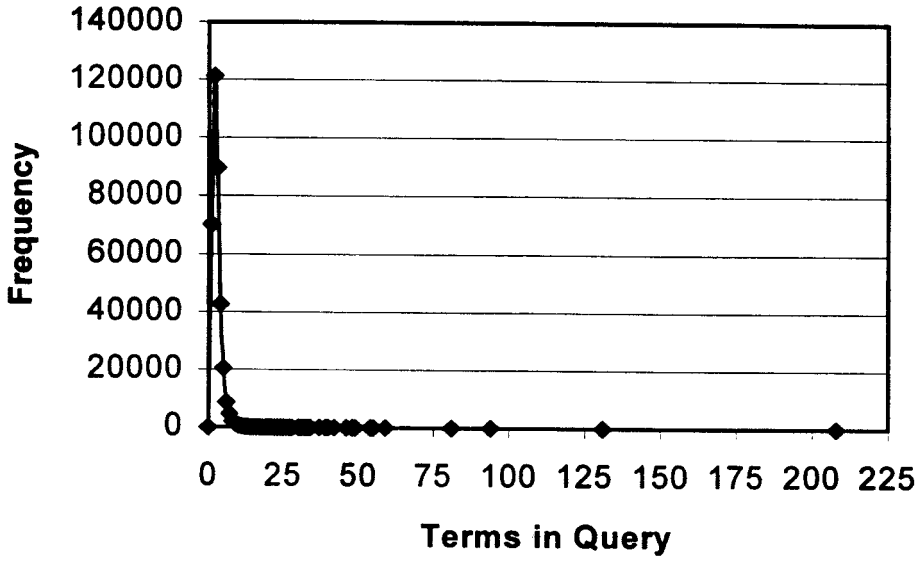

FIGURE 1 - Distribution of Terms Per Query

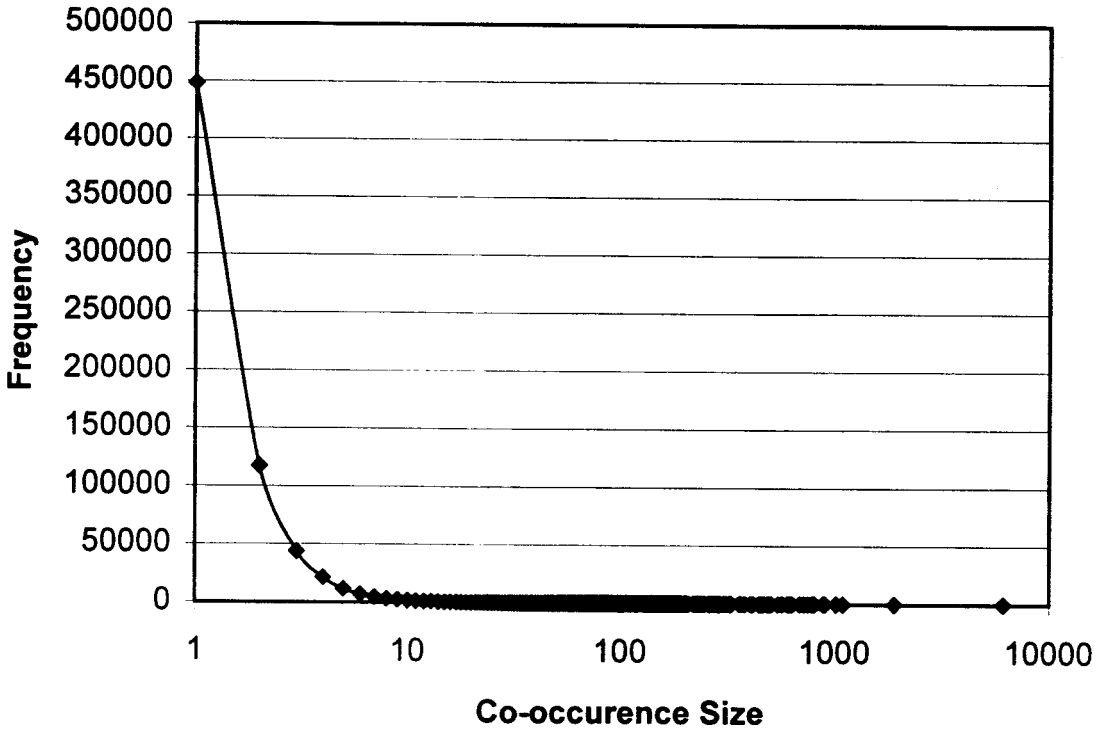

FIGURE 2 - Distribution of Term Co-occurrences

The simulation models reveal that binary dependence plays an important role in queries submitted.

Model A Results

A comparison of the overall characteristics of the observed and simulated queries appears below:

	Observed	Simulated
Number of queries	292,994	292,994
Number of unique terms	96,004	96,004
Number of non-unique term pairs	1,387,767	1,481,139
Number of term pair types	673,105	1,247,863

The simulation model generates 85% more term pair types, or unique term pairs, than is observed. This indicates that binary dependence of term co-occurrence plays an important role in how queries are formulated. Figure 3 displays the observed and simulated frequencies of term co-occurrence. The differences are so large that they do not merit chi-square goodness-of-fit testing. The simulation model, which assumes independence of term co-occurrence, underestimates the frequency of occurrence of the most frequently occurring term pairs, producing a maximum co-occurrence size of 99 (compared to 6116 for the observed data). It also grossly overestimates the number of singly co-occurring term pairs. Nelson (1983) found a similar pattern when estimating term co-occurrences within several databases. The differences are even more pronounced in the present study, leading the author to suspect that binary dependence plays an even stronger role in the development of queries than in the co-occurrence of terms within databases.

Model B Results

A comparison of the overall characteristics of the observed and simulated queries is summarized below:

	Observed	Simulated
Number of queries	292,994	292,994
Number of unique term frequencies	567	567
Number of non-unique size pairs	1,387,767	1,449,942
Number of size pair types	110,492	139,514

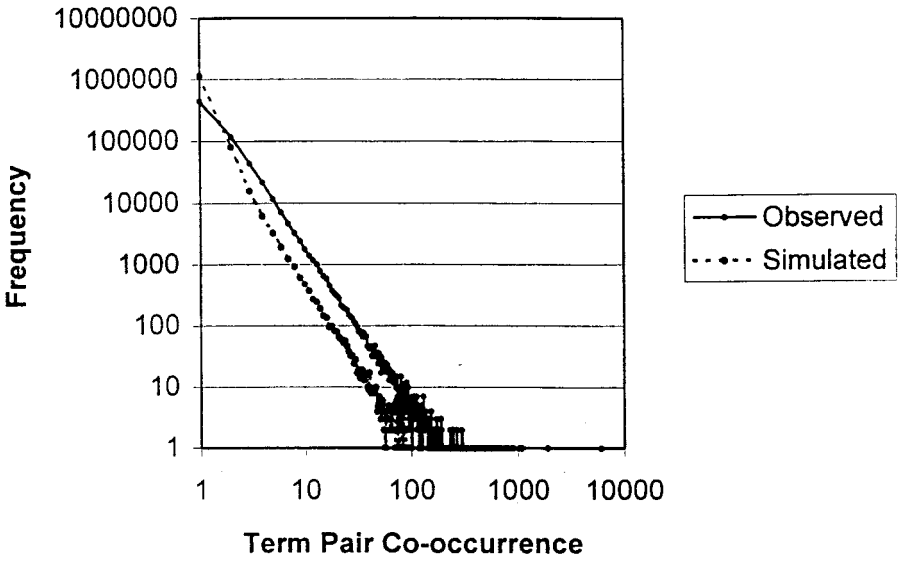

FIGURE 3 - Distribution of Observed and Simulated Term Pair Co-occurrences - Model A

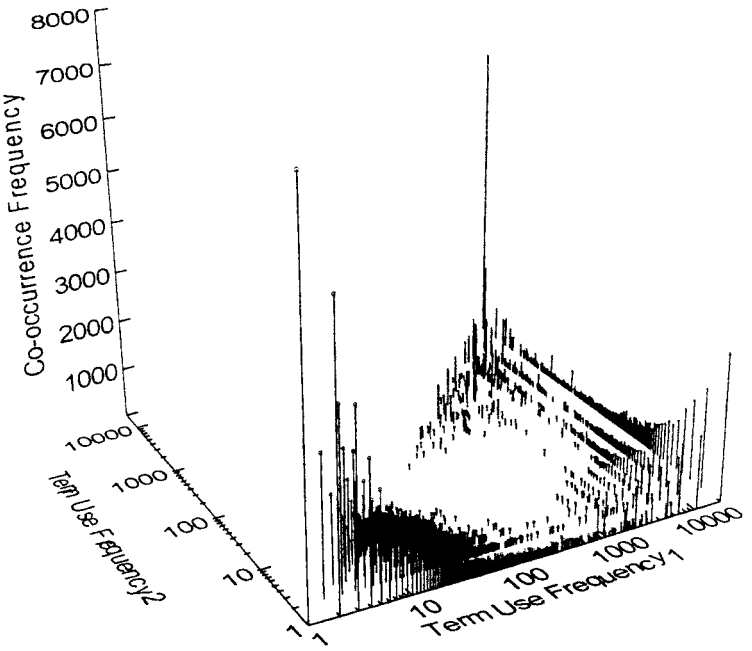

FIGURE 4 - Most Frequently Co-occurring Term Sizes - Observed Queries

A scatter plot of the most frequently observed co-occurring term sizes appears in Figure 4. Due to the large number of size combinations, only co-occurrence frequencies greater than 100 are represented. Since the term use frequencies are symmetric, only half of the plot contains data points. Infrequently used terms most often co-occurred with other infrequently used terms. Similarly, the most frequently used terms also co-occurred with one another, with co-occurrences of frequently and infrequently used terms also co-occurring in large numbers. This is not surprising since one would expect that with a large number of infrequently occurring terms and terms with high frequency, such terms would co-occur frequently by chance alone.

The simulation model overestimates the total number of term size pairs by about 25%. To determine where these differences lie, a plot showing the chi-square cell value differences between the observed data and simulated model was constructed. Combining the data for the distinct pair sizes results in a plot that contains approximately 149,000 points. Since this large number is impractical for a small plot, a subset of the chi-square cell values was selected. An analysis of the differences between the observed and simulated size pairs, reveals that more than 75% of the cells generated chi-square values of less than 5. Therefore, to see where the most significant differences lie, a plot showing only those cells where chi-square values are greater than 100 is included here. This results in a plot of more than 1,600 points, representing the cells with the most significant departures between the observed and simulated data (Figure 5). A top-down view, showing where the significant differences lie without the magnitude of the differences appears in Figure 6.

The graphs reveal that most of the differences occur between the very low term usage frequencies and, particularly, with the very high frequencies. There is a distinct absence of high cell values towards the center of the plot, indicating a relatively good fit between the observed and simulated co-occurrences. In general, the simulation model underestimated the numbers of co-occurrences for the less frequently used terms and the most frequently used terms, indicating that term dependence is also important for these pairs. The model also tended to over estimate the numbers of co-occurrences for terms of mid-range frequency, but to a much lesser degree.

The results indicate that many queries are using combinations of terms that are unique, or conceivably more specific, and do not incorporate 'popular' search terms. If we ignore the most frequently used terms ("and", "the", "of"), which provide little semantic contribution, the peaks at the high end of the observed data plot in Figure 4 disappear, as do a number of the highest cell chi-square values comparing the observed and simulated results in Figure 5.

DISCUSSION AND CONCLUSIONS

The distribution of terms used per query demonstrates a wide range of query complexity, from the simple to the absurdly long in a few cases. The lengthiest queries were a result of cutting and pasting of textual passages into the query box, or were due to a lack of understanding ofthe limitations of the search engine in handling lengthy queries.

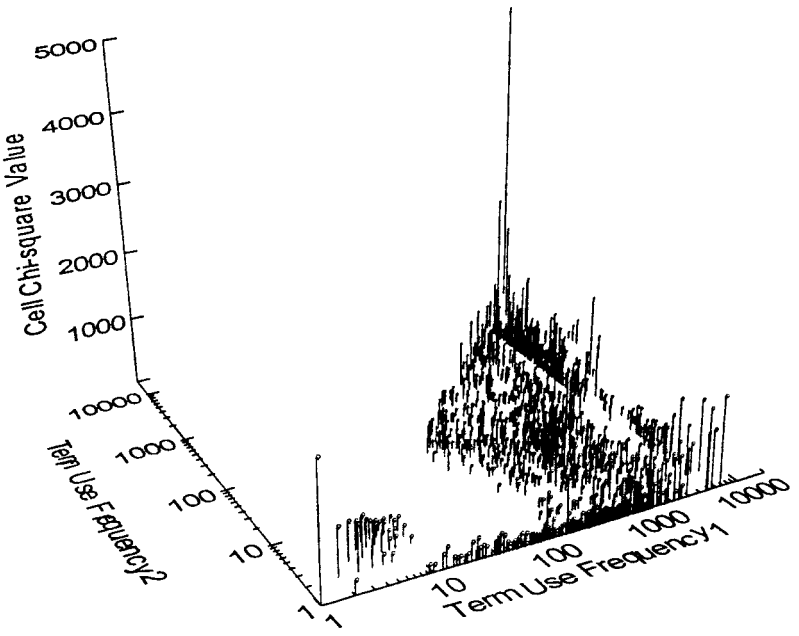

FIGURE 5 - Large Cell Chi-Square Value Differences Between Observed and Simulated Queries – Model B

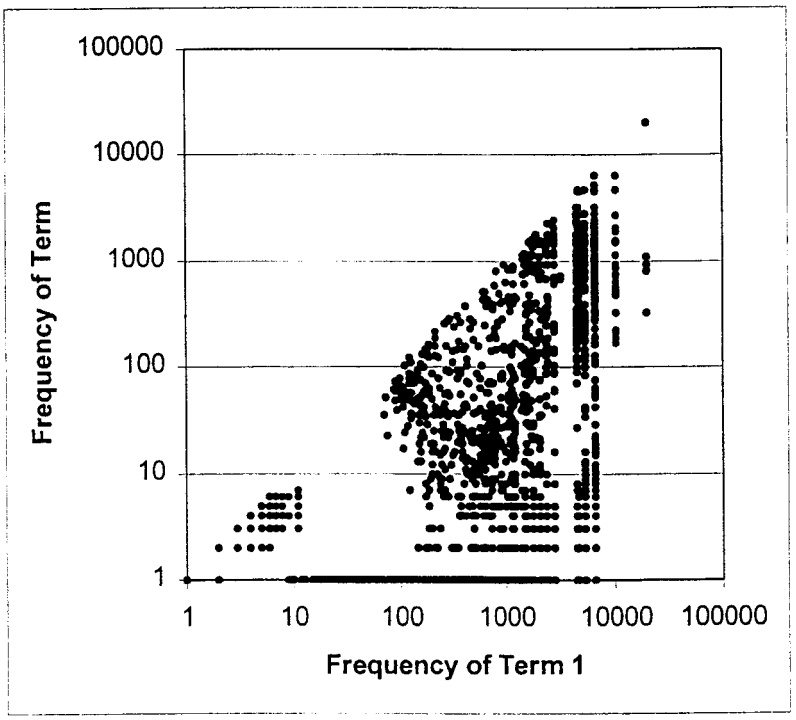

FIGURE 6 - Large Cell Chi-Square Value Differences Between Observed and Simulated Queries (No Magnitude) – Model B

Many of the longest queries appeared to incorporate large numbers of terms or synonyms without the use of Boolean operators in an attempt to be as inclusive as possible. The twenty-five most frequently occurring term pairs reveal a range of broad subject areas including education, media, and adult-oriented material. It is not possible to draw any meaningful conclusions from such a small representation of the data. The author is currently investigating classifications of search topics through term co-occurrences as part of another study.

The poor fits for the distribution of term co-occurrence frequencies using the three models leads the author to conclude that the data do not follow a classic Zipfian form. In informetric studies, the largest differences between observed data and fitted models usually occur at the tail end of the distribution with the highest size values for size-frequency models. For the present study, the largest differences were observed at the low end of the distributions, where the observed values descended more sharply than the models would have predicted. Other classes of distributions may provide better fits.

Binary dependence does indeed hold true on some level for the co-occurrence of query terms. One easily suspects that the selection of query terms by searchers does result in a binary dependence between specific search terms, since searchers do not randomly select search terms. But it is the plots comparing the observed and simulated (binary independent) data that reveal this clearly. Trinary co-occurrence (i.e. the co-occurrence of triples of specific terms) may provide additional insights; however, it increases the complexity of the analysis and reduces the number of queries which can be studied to those queries with three or more terms, representing a smaller percentage of the total queries submitted.

It is through simulation model B and the frequency of usage of query terms that a better understanding is gained of how co-occurrences take place. Instead of relying on specific pairs of search terms, the model looks at combinations of frequency of use. Do frequently used or popular terms co-occur with other popular terms, or do they co-occur with more unique terms? Large numbers of searches relied on pairs of infrequently used terms[2]. So, searchers do not simply rely on combinations of popular terms in their queries, which would be an indication that they are primarily interested in a smaller number of search topics. The co-occurrence of frequent and infrequently used terms was also observed. The pairing of infrequent terms with popular terms is, perhaps, due to a desire to provide a more specific, limiting term in addition to a more frequently used, and assumedly broader, term. Co-occurrences of mid-frequency usage terms appeared less frequently than the simulation model predicted, whereas the most popular and the more unique terms appeared more frequently than predicted, indicating that searchers relied more heavily on the more popular and unique terms in their queries. This could indicate that searchers were unaware of other relevant search terms that could have been incorporated into their searches.

Based on searcher query development behaviours, can one make recommendations for more efficient search engine design or for improved usage? One recommendation for search engine developers, considering the finding that fewer mid-range frequency search terms were used than predicted, would be the inclusion of an

[2] A portion of the 49,376 terms that occur only once, undoubtedly, represent spelling errors.

online thesaurus which searchers could consult for additional search terms of relevance to their topic of interest.

Even with today's faster computers, the need for efficient methods to access electronic contents is vital. Would the pre-coordinated storage of postings for frequently co-occurring terms provide more efficient processing of queries? Probably not. The most frequently occurring semantically meaningful search term pair ("free pics") occurs in approximately 0.37% of all the unique queries. When taking into account the frequency of occurrence of this term pair in non-unique queries, it represents an upper limit to the frequency of co-occurrences of term pairs over all queries. Since this most frequently occurring pair of terms represents such a small percentage of total term pairs, and these numbers quickly dwindle for subsequent pairs, providing access to pre-determined postings lists is unlikely to improve retrieval efficiency noticeably.

Should the most frequently used individual search terms be made more readily accessible in memory than less frequently used terms? Most definitely. The 10 most frequently occurring terms represent 0.01% of the 96,004 unique search terms. However, they constitute approximately 5% of all search terms used in the unique, multi-term set of queries studied. In this case, it makes sense from an efficiency perspective to provide faster access to the most frequently used terms.

In conclusion, the Excite query set presents a wealth of data for additional investigation. The author is currently examining relationships between different aspects of the data, including searcher and query characteristics and the interactions of these characteristics. Term co-occurrence data are also being used to determine the subject content of queries submitted by categorizing the thousand most frequently used term pairs and performing a cluster analysis of the resulting categories to reveal relationships among the subjects searched.

Acknowledgment:

The author would like to thank Excite, Inc. for making the query data accessible for study and Nancy Ross for research assistance.

REFERENCES:

Huberman, B A. et al. (1998). Strong regularities in World Wide Web surfing. *Science*, 280 (April 3), 95-97.

Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2), 236-243.

Jansen, B. J., Spink, A. & Saracevic, T. (In Press). Real life, real users, and real needs: A study of user queries on the web. *Information Processing and Management.*

Lawrence, S. & Giles, L. (1998). Searching the World Wide Web. *Science*, 280 (April 3), 98-100.

Nelson, M. J. (1983). The use of term co-occurrence information in information retrieval. *The Canadian Journal of Information Science*. 8, 67-73.

Nelson, M. J. (1989). Stochastic models for the distribution of index terms. *Journal of Documentation*, 45(3), 227-237.

Peat, H. J. & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*. 42(5), 378-383.

Wolfram, D. (1996). Inter-record linkage structure in a hypertext bibliographic retrieval system. *Journal of the American Society for Information Science*. 46(10), 765-774.

APPENDIX

Twenty-Five Most Frequently Occurring Term Pairs

Term	Term Frequency	Term	Term Frequency	Co-occurrence Frequency
and	19974	and	19974	6116
of	10433	the	6512	1901
pics	2817	free	6469	1098
university	2732	of	10433	1018
new	2454	york	933	903
sex	5270	free	6469	886
the	6512	in	4768	809
real	1056	estate	825	787
home	2062	page	1782	752
free	6469	nude	4555	720
of	10433	and	19974	690
pictures	4691	of	10433	637
how	791	to	2094	625
and	19974	the	6512	614
free	6469	pictures	4691	605
high	940	school	2062	571
xxx	1782	free	6469	569
and	19974	free	6469	545
adult	2062	sex	5270	508
and	19974	or	794	505
or	794	or	794	501
sex	5270	pictures	4691	496
nude	4555	pictures	4691	486
for	3179	sale	683	467
and	19974	not	332	456