Mining Longitudinal Web Queries: Trends and Patterns

Peiling Wang
School of Information Sciences, The University of Tennessee, Knoxville, Tennessee 37996-0341.
E-mail: peilingw@utk.edu

Michael W. Berry and Yiheng Yang
Department of Computer Science, The University of Tennessee, Knoxville, Tennessee 37996-3450.
E-mail: {berry, yyang} @cs.utk.edu

This project analyzed 541,920 user queries submitted to and executed in an academic Website during a four-year period (May 1997 to May 2001) using a relational database. The purpose of the study is three-fold: (1) to understand Web users' query behavior; (2) to identify problems encountered by these Web users; (3) to develop appropriate techniques for optimization of query analysis and mining. The linguistic analyses focus on query structures, lexicon, and word associations using statistical measures such as Zipf distribution and mutual information. A data model with finest granularity is used for data storage and iterative analyses. Patterns and trends of querying behavior are identified and compared with previous studies.

Introduction

The World Wide Web has brought information retrieval (IR) to end users who may have never searched other IR systems such as online library catalogs and CD-ROM databases. The amount of information on the Web is growing rapidly; the number of interactions by perpetual novice users grows as well. Several early studies of Web searching reveal that the system is neither simple nor easy to use. More than 30% of the queries resulted in zero-hit outcomes and users had difficulties in formulating queries (Pollock & Hockley, 1997; Shneiderman, Byrd & Croft, 1997; Wang & Pouchard, 1997). Many search engines now can avoid zero-hit results by partial matching or by processing multiword queries first with the Boolean AND operator then the Boolean OR, etc. As Cooper (1988) points out, on the other side of the coin, there is far too much output. The users are often at a loss and are unable to retrieve quickly needed entries when presented with a poorly ordered listing. To solve the fundamental problems of IR (such as presenting needs effectively and retrieving useful information efficiently), it is important for researchers and designers to understand what the users are searching for, how they search, and what problems they encounter. The number of published studies of Web users has grown rapidly in the last few years. Using tracking techniques to capture Web searches and interactions to observe users' behavior, several projects report on analyses of user queries logged by the search engines, such as the University of Tennessee, Knoxville's search engine (Wang & Pouchard, 1997), *Excite* (Jansen et al., 1998; 2000; Wolfram et al., 2001), and *AltaVista* (Silverstein et al., 1999). This project extends these studies in three ways: (1) analyzing truly longitudinal user queries across five calendar years (541,920 queries logged from May 1997 to May 2001); (2) exploring appropriate data model for query data organization and mining; (3) adopting mutual information statistics in combination of observed frequencies for measuring word association.

Related Studies

Web users' queries are a type of naturally occurring linguistic data that are quite different from textual document data that IR researchers have spent several decades investigating. It is only in the last few years that researchers have analyzed large numbers of user queries submitted to popular Web search engines: the *Excite* search queries were captured three times on selected days (Jansen et al., 1998; Spink et al., 2001; Ross & Wolfram, 2000; Wolfram et al., 2001); nearly one billion queries were logged by the *AltaVista* U.S. site during a 43-day period (Silverstein et al., 1999). Although differences exist among the log data, these projects all included exact query statements entered by the users and several comparable statistical analyses, such as frequencies of queries or query terms, number of unique queries, number of words in a query, and word associations. However, these studies have not reported data processing techniques in detail. Can new analyses be applied without

reprocessing data? Is there a potential extension of the analyses over a much longer period by adding new data?

There is clearly a need to model changes of Web users' searching behavior based on longitudinal analyses as new queries are added. Are there any improvements in query formation? Is there a better understanding of the system over time? Or, are there changes in the type of information submitted to a popular search engine? (Wolfram et al., 2001). The *Excite* researchers have collected queries again on selected days following their original 1997 study (Wolfram et al., 2001; Spink et al., 2002). By comparing three slices of data from *Excite* collected in September 1997, December 1999, and May 2001, they observed a shift of search topics but little change in users' search behaviors. These results from snapshot logs represent only a very small fraction of the actual search activities, which is a limitation noted by Ross and Wolfram (2000).

Jansen and Pooch (2001) present a framework for Web query analysis by examining reported studies focusing on Web queries. They urge researchers to promote consistency in statistical analysis and uniformity in terminology and definitions; they also call for detailed description on data collection and data analysis to facilitate comparisons.

Research Goal and Research Questions

This study identifies the patterns and trends of Web queries of an academic Website by mining longitudinal data covering a four-year period. An elaborate relational data model is used to store data and to conduct various statistical analyses on the queries, the vocabulary, the words and word associations, and problematic queries. The analysis of longitudinal data intends to complement the related studies of Web engines using data from snapshot logs. The ultimate goal is to contribute to our understanding of Web users' searching behaviors and to guide the design of useful functions for effective interactions between users and the system.

Specifically, our analyses address the following research questions:

(1) What are the characteristics and trends of querying activities on an academic site?
(2) What are the characteristics and trends of the user vocabulary (words and word associations)?
(3) What are the major problems of the user queries?
(4) How do these results compare to those from the studies of different Web search engines?

Methods

User queries entered to the Website of the University of Tennessee at Knoxville were captured unobtrusively from mid-May 1997 to mid-May 2001. The site had used a shareware program called SWISH (Simple Web Indexing and Searching for Humans, 2003) to build its search engine until January 2000 when the search engine was migrated to the software package from Inktomi with a period of both search engines available on the main page between January 2000 and May 2001. The queries analyzed in this project include only the ones entered in the old search engine by SWISH. The interface for the SWISH search engine was basically a simple slot accepting Boolean search statements; it defaulted multi-word queries to Boolean AND. Due to limited server space in the initial stage, only the homepages within three links from the main page were indexed by SWISH. The log file of the queries did not include IP addresses of individual users due to privacy concerns. Therefore, the search engine had limits in functions and coverage, and sessions of the individual users could not be identified from the log data. Nevertheless, we have valuable data on query level statistics that reveal users' search activities, as well as the actual queries that reveal both topics and linguistic structures. We report observations on behaviors of the user population as a whole.

Data Structure: the Model

The log file consists of the queries along with date stamps and numbers of hits in double-colon delimited ASCII format. Here are examples [typos are not corrected, e.g., *htpp* is a typo for *http*, *deparment* should be *department*]:

1997/12/11::htpp://www.unn.ac.uk/soc. . .c/about/misc/alcohol.htm::0
1998/08/14::graduate advisors deparment of germanic, slavic, and asian languages::0
1998/08/13::“ admissions + out-of-state”::0
1998/12/18:: 1-800 nomber to get grades::0
1999/01/05::Peyton Manning::7
2000/10/24:::0
2001/05/16:: “alpha kappa alpha”::12

These illustrative examples show that some searchers entered URLs as queries, natural language statements, Boolean statements in a format of other search engines, person's name using capitalized first letters (Peyton Manning was a quarterback of the football team), empty (no string), organization's name (sorority), punctuation marks, same word repetitively, and so on.

The first and foremost decision on data processing was to find a good data structure. The traditional implementation of a master file and inverted indexes for text data storage and processing proved neither efficient nor appropriate. A pilot analysis processed a portion of the queries using the traditional IR method to generate statistical measures, such as frequencies of queries and words, word-pair (word co-occurrences), etc. This method required time-consuming reruns to incorporate new data; it also lacked the flexibility to apply new analyses. For example, to change the breakdown of the file, it required substantial work to revise the programs as well as to reprocess the data. In recognition of the needs for longitudinal analysis of the queries and the

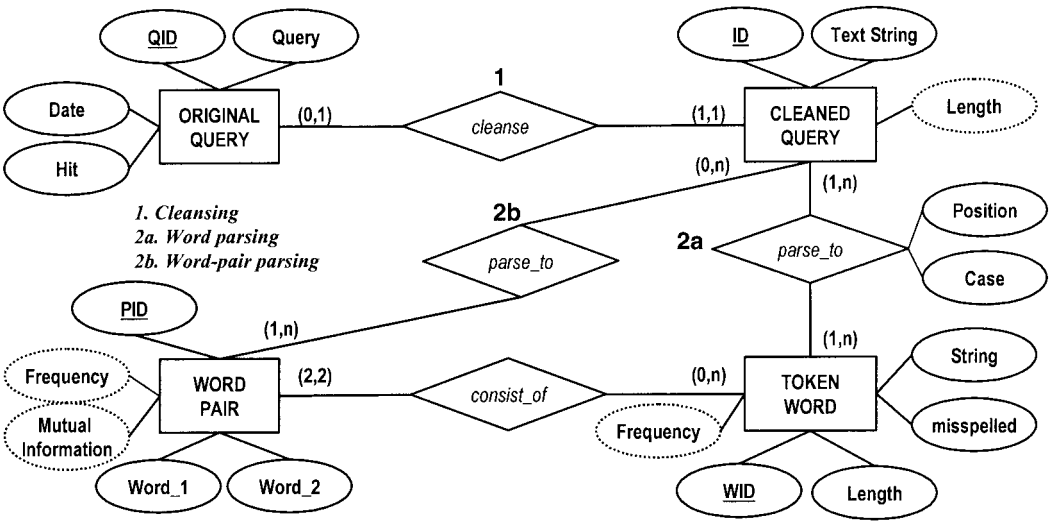


FIG. 1. Web query ER model for relational database.

iterative nature of data mining, the relational data model is developed at the finest granularity for data storage and analysis (Fig. 1). This approach takes advantage of the SQL standards, enables scalability to a large amount of data, and is flexible enough to incorporate new analyses as well as long-term mining needs.

Data Cleansing and Parsing

The pilot processing of queries found various unusual queries: blank queries (empty), strings with only punctuation marks, numeric-only (phone number, Social Security numbers, course numbers), Internet-related, computer or Internet-specific strings. The two projects mentioned above used different parsing strategies: (1) space parsing for the *Excite* queries and (2) space plus punctuation parsing for the *AltaVista* queries. We believe that certain queries should be understood as a whole to preserve the features of the Web queries. Therefore, the first step of our data processing, called cleansing, separated unusual queries from text queries; this processing also removed punctuation marks for further linguistic analysis.

The second processing step parsed the queries into single words and word-pairs (Fig. 1). The analysis of word association focuses on word-pairs that are the adjacent words or the words with one intervening word; word order is not differentiated. For example, the query "*petty cash reimbursement*" is parsed into three word-pairs: cash petty (first and second words in adjusted order), cash reimbursement (second and third words), and petty reimbursement (first and third words). This will preserve most significant word pairs in Boolean sense, given the fact that nearly 94% of the queries contain three words or less (Table 1 in the "Results" section), and the intervening word may be a grammatical word (such as the, of, and, in, etc.) or a Boolean operator (AND, OR, and NOT).

Misspelled words are included in the word table but are identified by the attribute *Misspelled* (using a spelling check plus human checking to allow personal names and acronyms). Caution was given to + and / signs that were often used to mean Boolean AND and OR, as evident in the following queries:

cheeting/plagerism [typo not corrected];
"Ratios"+"Student"+"Professor"

A URL address was not parsed; it was treated as an unusual query. The third decision on data processing was to include the common IR stop words (a, the, of, etc.) in order to preserve linguistic features of the user queries.

Data Analyses

Descriptive statistics (frequency and distribution) were generated to describe the characteristics of the queries. Linguistic analyses focused on establishing vocabulary, running statistics to test Zipf's laws using the following formula based on Baeza-Yates and Ribeiro-Neto (1999) and Korfhage (1997):

$$F_i \times R_i^{\theta} = C$$

where F_i is the frequency of the ith word with rank R_i in descending order and C a constant; $\theta > 0$.

Zipf law is one of the enduring interests in text analysis. The two *Excite* projects plot the trendlines of ranking by frequencies.

Another important analysis in text mining is word association. Callon et al. (1983) believe that the associations created between words in a text are more or less solid. Leydesdorff (1989) suggests that, as a linguistic approach, word-word associations are useful indicators of internal

structure of coherent document sets and possible tools to partition documents into clusters. There are several measures based on observed frequencies to identify word associations. Silverstein et al. (1999) use the Chi-square test and correlation coefficient in identifying word associations of the *AltaVista* queries; their analysis was limited to the 10,000 most common words and the queries containing these words. Ross and Wolfram (2000) focused on 1,054 of the most frequently occurring pairs out of the 673,105 pairs from 292,994 unique multi-word queries in *Excite*; they used Pearson's *r* distance measure to produce clusters.

Researchers in identifying co-occurrence preferences or strength of word associations have used mutual information statistics. The mutual information formula, as a measure of word association, is computationally simpler than the Chi-squared test or correlation coefficient; and importantly it does not assume mutual independence of the two words to calculate the expected frequencies. Church et al. (1991) demonstrate the use of this measure to identify semantic similarity and strengths between terms. Khoo, Dai, and Loh (2002) use mutual information to identify terms of Chinese texts.

This project observes all word-pairs, not just the most occurring word pairs in terms of strength—the low frequency pairs are not ignored. In fact, a low frequency word pair may be strongly associated if the two words always occur together. This can be indicated by the mutual information statistics. One of the commonly used formulas is as follows:

$$I(w_1, w_2) = \ln \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

where $P(w_1)$, $P(w_2)$ are probabilities estimated by relative frequencies of the two words and $P(w_1,w_2)$ *is the relative frequency of the word pair (order is not considered, therefore* $w_1 < w_2$). Relative frequencies are observed frequencies (F) normalized by the size of the queries:

$$P(w_1) = \frac{F_1}{Q};\ P(w_2) = \frac{F_2}{Q};\ P(w_1, w_2) = \frac{F_{12}}{Q'}$$

Although both the observed frequency of a word and the frequency of a word-pair are defined as the number of queries in which the word or the word pair occurs, the total size of the queries has different bases. The size of the queries for words is the actual size of the cleaned queries (Table 1), while the size of the queries for word pairs depends on the parsing algorithm. First, single-word queries do not produce pairs. Second, a two-word query produces one pair and is counted once, but a three-word query is parsed into three pairs and counted once for each of the three pairs. (That is, the query is counted a total of three times.) In the same way, a four-word query produces five pairs and is counted five times. (See "Data Cleansing and Parsing" for explanations of the rules). To cope with this phenomenon, the adjusted size of the multi-word queries should be calculated by the formula:

$$Q' = \sum_{n}^{m}(2n - 3)Q_n$$

where $Q_n$is the number of queries with n words ($n \geq 2$), m is the maximum number of words in queries ($m = 26$). Therefore, $Q = 526{,}360$ [the actual size of all cleaned queries including single word queries], and $Q' = 631{,}491$ [the adjusted size of all multi-word queries].

As noted in the previous section, the individual sessions could not be clearly identified. However, some consecutive queries with sequent log ID numbers show possible iterations by a same searcher. Our analysis only intended the speculations on possible users' mental models or misconceptions of the system for further observations using a different data collection method.

Analyses and Results

In this section, the results are presented along with comparison of other studies. The logged queries span five calendar years covering a four-year period (May 1997 to May 2001). Most of the analyses are based on calendar years unless noted otherwise.

Query Level

Table 1 indicates that a steady increase in search activities is seen from 1997 to 1999. The drastic decrease of the number for 2000 and 2001 was a result of the migration of the search engine. The log file in this study includes only the queries submitted to the old search engine, which covers a four-year period with complete data for two calendar years (1998 and 1999) and two academic years (1997 to 1999). The percentage of null outputs is consistently more than 30%. It is worth attention that the 30% zero-hit problem is a phenomenon that has not been improved over the years, which is also reported in the study by Shneiderman, Byrd, and Croft (1997). The search engine, as described in the "Methods" section, contributed partially to this problem: the default for multi-word queries was set to Boolean AND; only the homepages within three links from the main page were indexed; numerics were not indexed; and auto-correction on misspelled words was not supported. Nevertheless, the high percentage zero-hit queries reveal that most users did not comply with the system's rules that were readily available (as a help button and a link following the no match message for zero-hit queries). Misspelling as high as 26% was one of the major problems and searching for personal names also contributed to a high number of zero-hit because the directory was not indexed by the engine; rather, it was a separate search option.

TABLE 1. Query characteristics (original queries and cleaned queries).

Year[1]	Total Q cleaned Q[2]	Zero-hit (%)	Empty (%)	Longest[3] (average)	One-word[4] (%)	Two-word[4] (%)	Three-word[4] (%)	Four-word[4] (%)	Mean[4] (max)
1997[1]	73,834	24,826	424	101	29,400	29,544	8,813	4,088	2
	71,845	(33.6)	(0.6)	(13.1)	(40.9)	(41.1)	(12.3)	(5.7)	(16)
1998	172,492	55,229	1,079	131	66,204	69,584	21,611	10,332	2
	167,731	(32.1)	(0.6)	(13.3)	(39.7)	(39.5)	(12.9)	(6.2)	(26)
1999	233,442	75,533	1,200	121	86,770	95,384	30,892	14,878	2
	227,924	(32.4)	(0.5)	(13.6)	(38.1)	(41.8)	(13.6)	(6.5)	(17)
2000[1]	43,448	15,594	600	113	15,648	17,243	6,505	2,607	2
	42,003	(35.9)	(1.4)	(13.3)	(37.3)	(41.1)	(15.5)	(6.2)	(16)
2001[1]	18,704	7,472	1,568	104	6,201	6,797	2,908	951	2
	16,857	(39.9)	(8.4)	(12.2)	(36.8)	(40.3)	(17.3)	(5.6)	(16)
Grand Total	541,920	178,654	4,871	—	204,223	218,552	70,729	32,856	2
	526,360	(33.0)	(0.9)	—	(38.8)	(41.5)	(13.4)	(6.2)	(26)

[1] The 1997 queries include May to December; the 2001 queries include January to May; the number of queries for 2000 or 2001 drastically dropped due to the fact that the Website added a new search engine.

[2] Cleaned queries exclude empty and unusual queries such as numeric and Internet-related queries.

[3] Longest is the maximum number of positions in a query counting between the first character and the last character.

[4] From the cleaned queries: a word is the character string separated by white space or punctuation marks; calculations of means (average) exclude empty queries; max is the maximum number of words in queries.

Empty queries dropped 0.1 percent from 1998 to 1999, but increased afterwards from 1.4% in 2000 to 8.4% in 2001. The average for all queries was 0.9% (Table 1). Two previous studies report comparatively higher percentages of empty queries: 5% (*Excite* queries) and 20% (*AltaVista* queries).

Most queries are short with an average of two words or 13 character positions (the longest query is a 26-word natural language statement with 131 positions; see the example queries with the ID 103519 in the "Observation on Mental Models"). The one-word queries count for as high as 40.9% in 1997 (38.8% for all years); the two-word queries are around 41.1%. There is a steady increase in the three-word queries over the years (from 12.3% in 1997 to 17.3% in 2001; average 13.4%). Fewer queries (6.2%) contain four words or more (Table 1). These findings are consistent with the results from the two *Excite* studies and the *AltaVista* study.

The distributions of the queries across months for two academic years show similar patterns (Fig. 2). The peak month was January for both years; both June and August had comparatively less queries for both years.

The graphs of two selected weeks from Spring and Fall of 1999 show daily searching activities at the site (Fig. 3A). Mondays through Thursdays had substantially more submitted queries than Fridays and weekends. Similarly, there were fewer queries during the Christmas and New Year's holidays (Fig. 3B), when the university was closed. Two frequently searched queries, "*career services*" and "*football tickets*" (see Table 2), plotted by month show strong seasonal effects for both queries (Fig. 3C). These phenomena indicate that the searching activities correspond to the academic cycles and seasonal events. The majority of the UTK Web users are likely from the university community.

For the cleaned queries (526,360), there are 134,721 unique queries, of which, 97,990 queries (72.7%) occurred

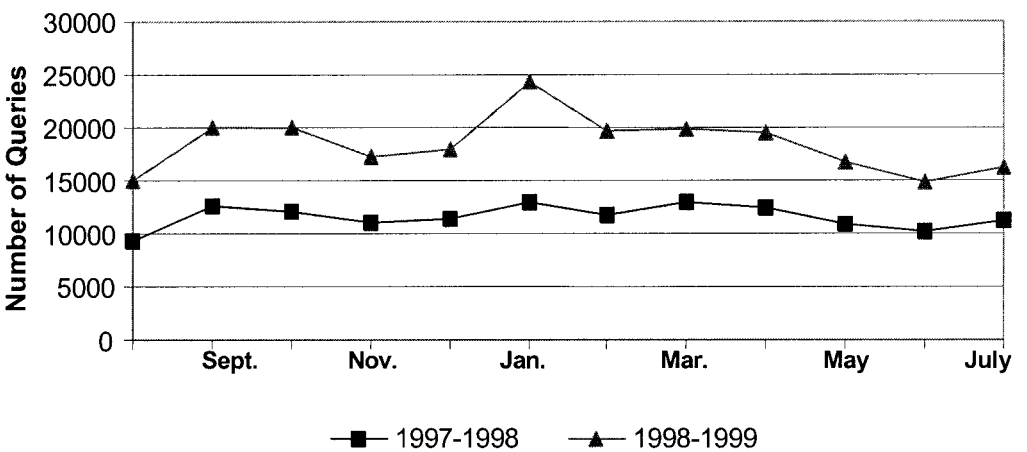


FIG. 2. Monthly query plots of two academic years.

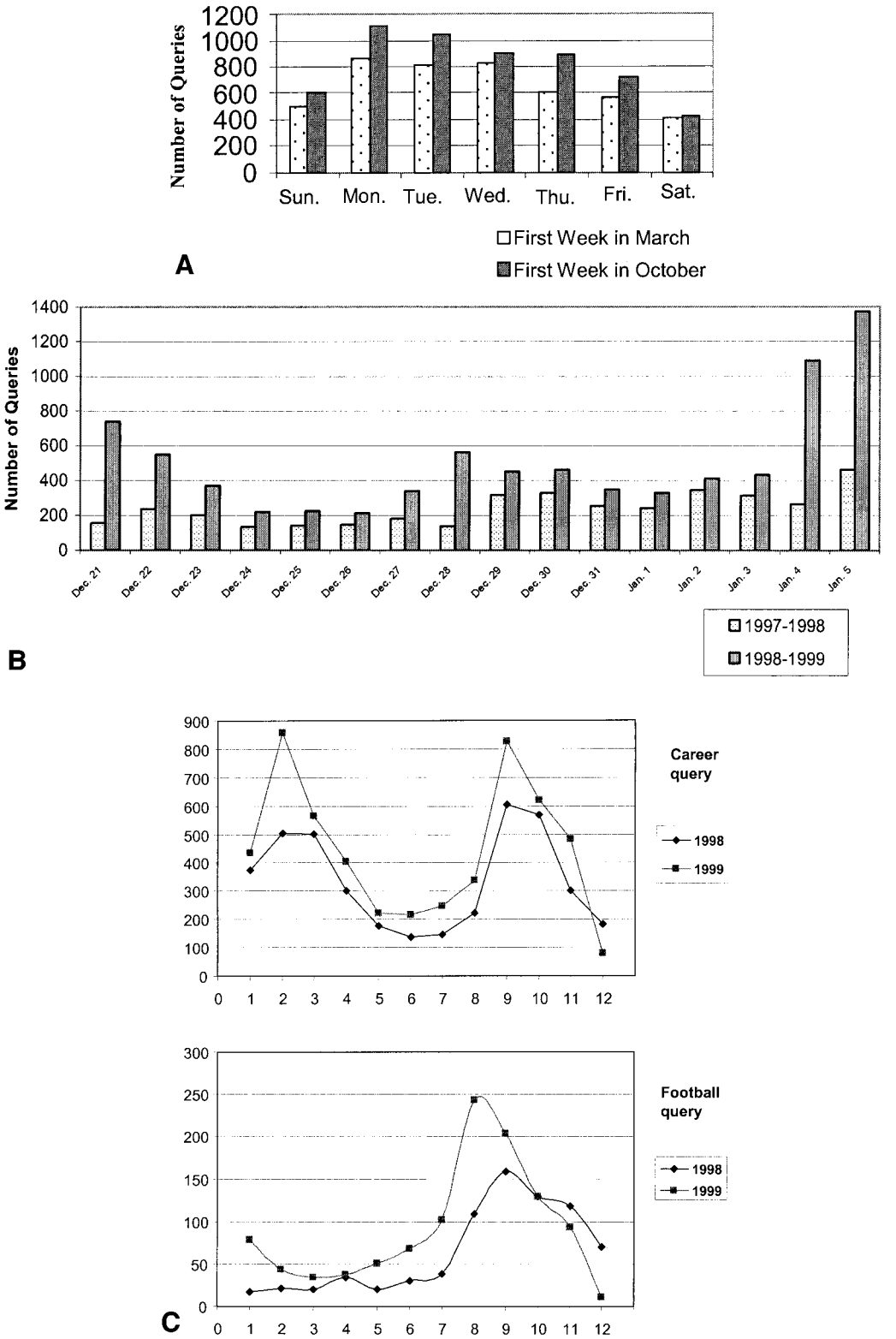


FIG. 3. A) Selected weeks in 1999. B) Number of queries during Christmas and New Year's holidays. C) Monthly queries for career vs. football.

only once and 16,378 queries (12.2%) twice. The most frequently occurring query is the two-word query "*career services,*" (Table 2) counting for 9,587 occurrences (4.4% of all two-word queries). Observations concerning query structures indicate that most queries are unique text strings and that only a small number of queries are frequently submitted. Our results are compatible with the findings from the *Excite* studies and the *AltaVista* study. Table 2 lists the top 28 queries (occurring more than 1,000 times) in comparison with *AltaVista*'s top 25. (*Excite* studies are not

TABLE 2. Comparison of top queries with *AltaVista.*

Rank	This study	Frequency	*AltaVista*	Frequency
1	career services	9587	sex	1551477
2	grades	5727	applet	1169031
3	tuition	4837	porno	712790
4	housing	4203	mp3	613902
5	timetable	4097	chat	406014
6	bookstore	3453	warez	398953
7	Rocky Top	2582	yahoo	377025
8	transcripts	2340	playboy	356556
9	Daily Beacon	2312	xxx	324923
10	employment	2156	hotmail	321267
11	cheerleading	1985	[non-ASCII]	263760
12	band	1914	pamela anderson	256559
13	registration	1683	p**** [vulgarity]	234037
14	scholarships	1537	sexo	226705
15	jobs	1488	porn	212161
16	football tickets	1465	nude	190641
17	career	1407	lolita	179629
18	marching band	1397	games	166781
19	cheerleaders	1377	spice girls	162272
20	resume	1375	beastiality	152143
21	financial aid	1331	animal sex	150786
22	webmail	1317	SEX	150699
23	tickets	1225	gay	142761
24	transcript	1211	titanic	140963
25	catalog	1187	bestiality	136578
26	Tennessee 101	1058		
27	football	1025		
28	biology	1000		

Notes: the *Daily Beacon* is a student newspaper. *Rocky Top* is a pep song played by the Pride of the Southland Band (marching band) during sports events and is one of the six state songs; *Rocky Top* is also the name of a campus bookstore. *Tennessee 101* lists aggregated student evaluations of courses.

included in this comparison because they report top words not top queries.) In contrast, the top queries in our study are academic related. These queries indicate that the frequently searched information needs typify students' information needs.

A further analysis, using pairwise comparisons of topical similarities or dissimilarities, of the top 28 queries (see Table 2) derived eight clusters. Specifically, if two queries were about the same aspect of university life, they were placed in one group; otherwise, they were set apart. The process was carried on until all queries were placed into one of the clusters. Then, each cluster was reexamined to see if all queries belonged to it. Two of the researchers conducted the analysis independently. The agreement of the results was 92.8%. A discussion resolved the discrepancies and reached eight categories. One problem we encountered was that some queries with the same wording might be searching for different topics. For example, "*Rocky Top*" is the name of the campus bookstore and the title of a pep song in sports. In this case, the query belongs to both clusters. The order of the list below is based on the number of queries with an adjustment for queries in multiple categories (the rank number in Table 2 is used to represent the actual query for brevity):

I. Academic related (18,303 queries): 2, 5, 8, 13, 24, 25, 26, 28
II. Sports & entertainment related (11,679 queries): 7, 11, 12, 16, 18, 19, 23, 27
III. Career related (14,936 queries): 1, 10, 15, 17, 20
IV. Finance related (8,783 queries): 3, 10, 14, 21
V. Housing (4,203 queries): 4
VI. Books & supplies (4,744 queries): 6, 7
VII. University publications (3,370 queries): 9, 26
VIII. Computing (1,317 queries): 22

Analysis of Vocabulary

It is found that the size of the vocabulary grows as the number of queries increases. The same is true with the word pairs. The increase in the number of words is much slower than the increase in the number of queries. Table 3 presents the statistical description of the user vocabulary across years. There are 44,455 unique words, of which only 2,966 words (6.7%) occur in all years, 2,584 words (5.8%) in four years, 3,768 words (8.5%) in three years, 6,653 words (15.0%) in two years, and 28,484 words (64.1%) in one of the years. The cumulative observed frequencies also put the query vocabulary into perspective: 175 words (0.4%) have a frequency greater than 1,000; 1,277 words (2.9%) occur

TABLE 3. Vocabulary characteristics.

Year	Cleaned queries	Unique words[2]	Words in pairs	Word-pairs
1997[1]	71,845	13,736	9,913	29,784
1998	167,731	21,872	15,962	58,037
1999	227,924	25,382	18,641	73,342
2000[1]	42,003	11,095	8,160	21,617
2001[1]	16,857	6,175	4,595	9,987
All	526,360	44,455	31,502	140,025[2]

[1] Data for 1997, 2000, and 2001 are incomplete; see Table 1, notes for explanation.

[2] See text for further exploration on overlapping across years.

100 times or more; 15,215 words (34.2%) occur three times and more; 21,767 words (49.0%) occur twice and more. In other words, 51.0% of the words occur only once.

As Korfhage (1997) points out, the most frequently occurring words in text documents are those included by grammatical necessity, such as *the*, *of*, *and*, and *a*; Zipf's law does not hold strictly. Both are true for our query data, which are from a large number of users and are very short in length. Tables 4a and 4b show that grammatical words *of*, *and*, and *for* rank high in occurrence, although *and* can also be interpreted as a Boolean operator.

For our query data, $\theta = 0.5$ or 0.6 produces best C values ($= F_i \times R^{i\theta}$) for higher rank words (Table 4a). Further exploration is done with plotting the trendlines of rank vs. frequency with respect to the natural logarithm (the *Excite* study used logarithm with base 10). Although the trendlines (Fig. 4) are not straight lines as Zipf's law would suggest, the lines for individual years show similar shapes. In fact, a small number of words occurred frequently and a large number of words occurred infrequently. This phenomenon is depicted as a second line for each year by ranking words based on unique frequencies. Therefore, words with the same frequencies cluster. The two lines overlap in the top part, but split near the point of $\ln(F) \approx 5$ and $\ln(R) \approx 6$, where the second line based on unique frequencies drops drastically. This drop corresponds to the fact that the number of words for low frequency increases as the frequency decreases (see also the last rows and notes in Table 4a).

Besides the grammatical words *of*, *and*, and *for* (*and* can mean a Boolean operator), the most frequently occurring words are consistent with top queries. From the list of the ranked words, another frequently searched category is the fraternities and sororities (*alpha* ranks 33rd with frequency 3,598; here are some of the queries including this word: "*Alpha Lambda Delta*," "*alpha kappa psi professional business fraternity*," "*Alpha Kappa Sorority*"). In addition, certain words occur seasonally, such as *football*, which is much higher during the fall months than spring months (see Fig. 3C).

Observed frequencies also reveal that certain words are queried more in one form than another such as plural vs. singular, for example: *services* (13,920) vs. *service* (1,542); *student* (10,668) vs. *students* (2,646); *transcripts* (2,632) vs. *transcript* (1,738); *cheerleading* (2,609) vs. *cheerleaders* (1,860) vs. *cheerleader* (547) vs. *cheerlead* (63).

In text data, a word is the basic unit. Using the word list in descending order of observed frequencies, the comparison of the top 20 words across different data corpora shows interesting linguistic regularities (Table 4b). Despite that the grammatical words such as *and*, *of*, and *for* are ranked high for all types of data corpora, more semantic words (such as *sex*, *nude*, *porn* in Excite; *services*, *career*, *student* in our academic site) than grammatical words are found as top words for queries compared to that for written or spoken English. The top 20s for the latter are mostly the pronouns such as *he*, *it*, *I*, and *you*. For the three query data corpora, the academic site has more semantic words than general search engines (Table 4b). This may reflect the fact that users with knowledge of conventional bibliographic databases would know that stop words should be excluded from search statements; Web users of academic sites are likely users of bibliographic databases.

To explore the highest ranking structural word *of* further, it is found that *of* occurs in queries with at least two words: there are 35 two-word queries containing *of*; 6,407 three-word queries containing *of*. Why would a short query with only two words need the structural word *of*? Some examples of such original queries show the problems and the possible needs: "*history of*," "*Department of*," "*head of*," "*location of*," "*best of*," "*#of students*," "*#of undergraduates*," "*of classes*," "*of theses*," and so on (some of these queries seem to be one of a series of reiterated queries; see the section "Observations on Mental Models" for more examples). Do users want to generate a list of headings similar to browsing a library catalog index?

In contrast some of the frequently occurring words in *Excite* or *AltaVista* queries did not rank high in our data. For example "sex" and "nude" ranked first and sixteenth in *Alta Vista*, and third and fourth in *Excite*. In our data, however, "sex" ranked 358th in 1997, 446th in 1998, 786th in 1999, and 686th in 2000; "nude" ranked 921st in 1997, 1245th in 1998, 1484th in 1999, and 1740th in 2000. For all the queries, a total of 639 queries (0.1%) contain sex-related words and 97 queries ($\approx$ 0%) contain *nude*(*s*). The sex-related queries, however, seem to look for very different types of information, although a few of them were vulgar (not quoted here): *sex*, *transsexuals*, *sex jokes*, *sexual abuse*, *sexual harassment*, "*sexually transmitted diseases*," "*critiques on Sex Without Love by Sharon Olds*," "*sexual behavior or conduct and faculty and students*," etc.

Word Associations

Single word queries are likely to produce higher numbers of hits with lower precision. As Table 1 indicates, more than 41% of the queries contained two words and nearly 20% of the queries contained three or more words. The words occurring in the same queries represent various associations: the words form a phrase (Rocky Top), one defines a context for the other (financial aid), one is a synonym of the

TABLE 4a. Top ten words and Zipf fitting for selected ranks across four years.

	1997			1998			1999			2000		
Rank	Word	Freq.	F*R^0.5	Word	Freq.	F*R^0.5	Word	Freq.	F*R^0.6	Word	Freq.	F*R^0.5
1	*of*	2096	2096	*of*	5076	5076	*of*	7350	7350	*student*	1642	1642
2	*and*	1457	2061	*services*	4826	6825	*services*	6618	10031	*alpha*	1091	1543
3	*Services*	1450	2511	*grades*	4532	7850	*career*	5901	11408	*of*	1082	1874
4	*grades*	1409	2818	*career*	4487	8974	*student*	4555	10465	*and*	842	1684
5	*career*	1181	2641	*and*	3292	7361	*and*	4522	11877	*services*	770	1722
6	*football*	1090	2670	*student*	2918	7148	*school*	4275	12526	*delta*	700	1715
7	*school*	1060	2804	*housing*	2652	7017	*timetable*	4145	13322	*kappa*	669	1770
8	*student*	1055	2984	*school*	2649	7493	*tuition*	4115	14329	*center*	612	1731
9	*office*	964	2892	*schedule*	2297	6891	*football*	3295	12314	*career*	589	1767
10	*tuition*	872	2758	*football*	2261	7150	*housing*	3279	13054	*phi*	579	1831
74	*programs*	305	2624	*international*	687	5910	*transfer*	946	12515	***daily***	124	1067
113	*arena*	189	2009	*mba*	500	5315	*learning*	705	12024	***clubs***	75	797
147	***marching***	132	1600	***cost***	360	4365	*final*	492	9826	***care***	41	497
Last	[*]	1	16	[&]	1	20	[^]	1	41	[#]	1	14
	Distinct F values: 268			Distinct F values: 407			Distinct F values: 481			Distinct F values: 187		

Bold cells have several words of the same frequency:
* 7,022 words occur once; & 11,006 words occur once; ^12,850 words occur once; # 5,719 words occur once.

other (career employment), one serves syntactic function (of southland), etc. Some word associations can be used for query expansion at the interface level. For example, when a user types in a one-word query, highly associated pairs containing this word can be displayed for possible expansion. The sources of these pairs can be from the document corpus as well as the user query corpus.

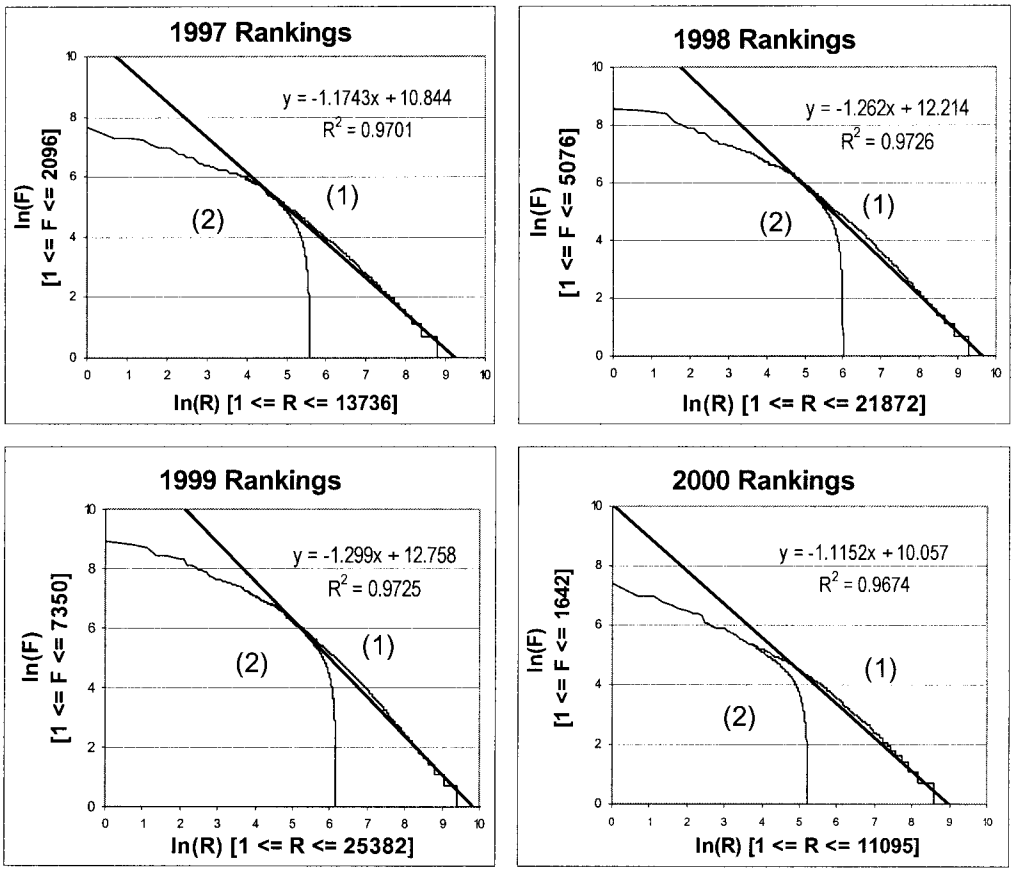
A common way of understanding associations is by means of matrices or two-way contingency tables. A close look at the data reveals that such matrix is sparse even if infrequent pairs are excluded, because neither all words participate in pairing nor all participating words pair with each other (see the "Methods" section for the operational definition of word pairs). Of the lexicon (44,455 words), there are 31,502 words resulting in 140,025 word-pairs, which exclude the 154 pairs of the same words (such as *alpha alpha*, *delta delta*, *and and*, *of of*, *pom pom*, etc.). There are 751 word-pairs (5.4%) with observed frequencies

TABLE 4b. Comparison of lexical top 20 words across data corpora.

Rank	Written English[1]	Spoken English[1]	*Excite*[2] (51K)	*Excite*[2] (1M)	This study
1	the	the	and	and	of
2	of	and	of	of	services
3	to	I	the	sex	career
4	in	to	sex	free	student
5	and	of	nude	the	and
6	a	a	free	nude	grades
7	for	you	in	pictures	school
8	was	that	pictures	in	tuition
9	is	k	for	university	housing
10	that	it	new	pics	football
11	on	is	+	chat	timetable
12	at	yes	university	for	schedule
13	he	was	women	adult	center
14	with	this	chat	women	office
15	by	but	on	new	band
16	be	on	gay	xxx	for
17	it	well	xxx	girls	department
18	an	he	to	music	UT
19	as	have	or	porn	Tennessee
20	his	for	music	to	graduate

[1] Data on Written English and Spoken English are from *The Cambridge Encyclopedia of Language* by Crystal (1987, p.86); original data were from the published studies of newspaper writing by Alexejew et al. (1968) and the London-Lund Corpus of Spoken English.

[2] The *Excite* studies rank the top words based on unique queries, while this study ranks the top words based on total cleaned queries. The *AltaVista* study does not report top words.



Note: Natural logarithms of frequencies and ranks are plotted for each year as two lines: (1) ranking all words; (2) ranking words with unique frequencies (i.e., words with same frequency are clustered). The two lines show overlapping (the upper portions) and divergences (the lower portions); the lower portion of line (1) has a close linear fit while the upper portion needs quadratic polynomial fitting. A drastic drop for line (2) occurs at the middle indicating lower frequencies cluster more words.

FIG. 4. Frequency versus rank plots and linear trendlines.

100 or greater and 50,989 pairs (36.4%) occur twice or more. In other words, 63.6% of the pairs occur only once. The word-pairs overlap across years: only 2,271 pairs (1.6%) occur in all years; 3,708 pairs (2.6%) in four years; 7,461 pairs (5.3%) in three years; 17,612 pairs (12.6%) in two years; 108,973 pairs (77.8%) in only one of the years. Typos, as high as 26.0%, might have contributed partially to such phenomenon. Selected illustrative pairs with high frequencies or high mutual information values are listed in Table 5; the ranking is based on observed pair frequency (F_{12}) then mutual information value (I). Frequencies provide intuitive word associations, which must be put into context. The mutual information statistic (I) measures the strength of the association between words. The I values for this data corpus range between -4.94 and 12.99 (Table 5). In a large well-structured document corpus, high I values should mean significant associations. However, for the user query data, the top I values often signal typos or personal names (the selected three examples with $I = 12.99$ in Table 5). Therefore, the two measures should be combined to understand word associations found in user queries. The pairs sharing the same frequency may have different I values (see examples with ranks between 184 and 187 in Table 5). Pairs with low F_{12} and high F_1 and F_2 are weakly associated with negative I values.

To explore further how high frequency pairs with relative low I values associate with other words, we selected the word pair *career* and *services*. This pair is number 1 in the ranking based on observed frequency 9,778, but only scored a moderate I value of 3.22 (Table 5). We retrieved the set of 705 pairs containing either words or both. Table 6 presents the illustrative examples of the set. *Career* is paired with 234 words and *services* with 471 words; corresponding I values range from -4.72 to 3.57. There are 97 pairs with higher I values than 3.22 (the I for *career services*). Most of the higher I values are from the pairs with usually one misspelled word pairing with either *career* or *services*. These cases require a careful look at mutual information statistics. When $f(w_1\ w_2) \approx f(w_i)$ $(i = 1\ or\ 2)$, the pair scores a high mutual information value if one of the words has a high frequency, which may indicate an anomaly such as misspelling (see the first, fourth and fifth rows in Table 6). If misspelled words are excluded, a high I value indicates that a rare word (low frequency) is strongly associated

TABLE 5. Selected word pairs with frequencies and mutual information.

Rank	$word_1$	$word_2$	F_{12}	F_1	F_2	I
1	career	services	9778	12328	13920	3.22
2	rocky	top	3384	3450	3533	4.80
3	beacon	daily	2727	3817	2808	4.72
4	of	the	2187	16063	4165	2.66
5	band	marching	1981	4955	2053	4.45
6	football	tickets	1959	7076	4291	3.34
7	college	of	1952	3293	16063	2.78
8	of	university	1840	16063	4347	2.45
9	e	mail	1827	2229	2734	4.88
10	aid	financial	1781	2027	2145	5.19
22	distance	learning	1117	1979	1521	5.09
23	physical	therapy	1111	1585	1582	5.27
25	football	schedule	1084	7076	6567	2.33
26	mini	term	1078	1129	1333	5.75
27	of	southland	1074	16063	1281	3.13
28	health	student	1073	3170	10668	2.63
29	arena	thompson	1060	1317	1751	5.31
30	boling	thompson	1049	1120	1751	5.46
32	get	grades	1033	1184	9000	3.75
34	center	medical	999	6143	2667	3.29
184	numbers	phone	**275**	**507**	**1389**	**5.14**
185	courses	online	**275**	**3660**	**2413**	**2.61**
186	the	university	**275**	**4165**	**4347**	**1.90**
187	classes	schedule	**275**	**2773**	**6567**	**1.89**
50990	shally	singler	1	1	1	12.99
50998	Caorl	Professr	1	1	1	12.99
51109	egnlish	instituion	1	1	1	12.99
126743	international	leadership	1	2061	114	0.62
140015	daily	of	1	2808	16063	−4.63
140025	beacon	of	1	3817	16063	−4.94

Note: F_{12} is based on adjusted size of queries Q' (see text for details). Typos are not corrected.

with a high frequency word, such as *picking* and *carnival* with *career* (Table 6). However, we cannot say that a high frequency word is strongly associated with a low frequency word solely based on the *I* value. Therefore, *career* is not a good predictor of the word pairs *picking career* or *carnival career.* We cannot predict the pair from the high frequency word even through the *I* is high (Table 6). For very low negative *I* values, there is no doubt that the pairs are weakly associated as compared with other pairs in the set. In combination with frequencies, *I* values are good indicators of strengths of associations interpreted within the context. In comparison, the association between *career* and *services* is higher than that of *dining* and *services*, or *parking* and *services*; *dining services* is a stronger pair than *food services.* As mentioned earlier one particular linguistic form seems to occur more often than another, i.e., *services* (plural) has a much higher frequency than *service*. Similarly, certain word pairs seem to be stronger than their linguistic variations. For example, the pair *career services* ($I = 3.22$) is stronger than *career service* ($I = 1.60$), but *food services* ($I = 1.83$) is weaker than *food service* ($I = 3.31$). Using both measures, a semantic network can be built to visualize this set including only the valid pairs by correcting misspelled words. Further, the set can also be linked to other sets. In order to derive semantic networks based on pair analysis, we need to set thresholds and develop formulas.

To compare the pair *career services* in connection with queries (Table 2), it is found that this pair occurs as a top-ranked two-word query. Longer queries containing these words also are related to career seeking. For example,

92356::1998/02/13::funnelweb career services::3
208345::1998/10/27::where do you go for an interview that you have signed up for using career services::0
396831::1999/08/26::career services - salaries::23
466612::1999/11/19::job listings through career services::0
485035::2000/02/20::career services website::20
531476::2001/02/15::ut career services resume writing::0

These queries form a cluster representing students' uppermost need for career-seeking support. Because this is a frequently searched category, the most frequently occurring pair *career services* can well be a good category heading for an academic Web directory that collects relevant pages on job listings, salaries, resume writing, interviews, schedules, and consulting. As searching is concerned, an effective matching algorithm would be to respond to keyword searching containing any of the career related words (see also Table 2) with a directory organized by related facets of career seeking instead of a list of ranked individual pages.

TABLE 6. Selected pairs in set of career services (N = 705).

$word_1$	$word_2$	F_{12}	F_1	F_2	I
career	cervices	7	12328	7	3.57
career	picking	1	12328	1	3.57
career	carnival	9	12328	10	3.47
carrear	services	6	6	13920	3.45
counselingdeafnesshuman	services	1	1	13920	3.45
carear	services	16	18	13920	3.33
career	*services*	*9778*	*12328*	*13920*	*3.22*
disbility	services	1	1	13920	3.45
dining	services	488	904	13920	2.83
parking	services	444	1767	13920	2.07
network	services	130	521	13920	2.06
printing	services	17	73	13920	1.99
food	services	119	600	13920	1.83
retirement	services	52	276	13920	1.78
career	service	214	12328	1542	1.60
data	services	38	282	13920	1.45
enrollement	services	2	15	13920	1.44
health	services	420	3170	13920	1.43
creative	services	14	111	13920	1.38
career	path	1	12328	9	1.37
career	homepage	11	12328	462	−0.17
career	professor	2	12328	189	−0.98
career	find	1	12328	95	−0.98
career	class	1	12328	2730	−4.34
career	employment	1	12328	3285	−4.53
football	services	2	7076	13920	−4.72

Note: negative values signal weak associations when F_{12} is much less than $F_1 * F_2$.

The trendline of the frequency versus rank for word pairs has similar features as those of single words plotted in Figure 4. Similar to single words, our high frequency word pairs, as expected, are quite different from those found in *Excite* or *AltaVista* results.

Misspelling

The percentage of misspelled words is as high as 26.0%. This means that the actual size of the lexicon would be 74.0% of the reported number of unique words if the typos were corrected. The types of spelling errors include missing or extra characters, missing spaces, and inverting or substituting of characters due to keyboard proximity or graphemes of the same phoneme (e.g., *cervices* in Table 6). *Career*, well-qualified as a frequently misspelled word, was found in at least eleven different misspellings seen in the following queries: "*carear services*," "*caree service*," "*careear services*," "*careeer service*," "*careert services*," "*carees services*," "*careet services*," "*carer sevices*," "*carere services*," "*carerr services*," and "*casreer services*." Athletic had the misspelled formats in queries: "*athlic training*," "*athlltic scholarships*," and "*athlwtic department*," etc.

Observations on Mental Models

Mental models are users' conceptualizations of the systems in terms of how a system works (Norman, 1983). Many users of Web search engines seem to learn searching through trial and error experience without a real understanding of the systems, which may never lead to a correct conceptual model of how various search engines work. We made some observations on searching behaviors. Queries searched on just about everything such as computing and Internet-specific symbols, numeric (phone numbers, Social Security Numbers, course numbers, semester, time, date, IP address, and forms). Here are a few examples: "*974-pool*" (the university phone numbers start with 974), "*423-525-8438*," "*1-800number*," "*377-78-xxxx*," "*246-41-xxxx*" (the last four digits are hidden for confidentiality), "*zoology230*," "*125 basic calculus section 60028*," "*102grades*," "*spring1999schedule*," "*11:10-12:25*," "*11-15-97*," "*Feb.22,2001*," "*206.172.235.48*," "*T-18*" [travel authorization form], address "*123 manor way apt 1*." Users also expect the system to understand natural language statements. Most long queries are natural language statements that include punctuation marks, for example:

> 63151::1997/12/02::Drop out rate for Freshmen at the University of Tennessee in the 1996-1997 school year::0
> 103519::1998/03/11::I have had this problem ever since the beginning of this week. I have not been able to dial into UTK network even with one success::0
> 460145::1999/11/11:: the relationship of loneliness, social isolation, and physical health to dietary adequacy of independently living elderly::0
> 507033::2000/09/05:: "how much garbage does the univer-

TABLE 7. A block of consecutive queries with reiterations likely by the same searcher.

ID	Month	Day	Year	Original query	Words	Hits
490766	4	9	2000	Teacher and student relationships	4	9
490767	4	9	2000	Teacher and student relationships	4	9
490768	4	9	2000	tenure and policy	3	24
490769	4	9	2000	tenure and policy and student-teacher	5	0
490770	4	9	2000	tenure and policy and student and teacher relationship	8	1
490771	4	9	2000	Sex	1	50
490772	4	9	2000	Sex and students and teachers	5	1
490773	4	9	2000	Sex and faculty and students	5	9
490774	4	9	2000	faculty and sex and students	5	9
490775	4	9	2000	Sex and tenure	3	6
490776	4	9	2000	sexual behavior or conduct	4	79
490777	4	9	2000	sexual behavior or conduct and faculty and students	8	39
490778	4	9	2000	sexual behavior or sexual conduct and faculty	7	4

sity of Tennessee at Knoxville dispose of in one month?" ::0
532922::2001/02/26:: "price comparison of three leading grocery stores with regard to the 35 most frequently purchased items"::0

Although individual users were not identified in these logged queries, a comparison of similar consecutive queries within a short span—executed sequentially as documented by the log IDs—showed features of mental models of Web searching (Table 7). That is, a later query shows a change of terms, adding, or dropping terms. The following three sets of queries may be the results from the assumption that adding word(s) could resolve the zero-hit problem:

55389::1997/11/11::the beacon::0
55390::1997/11/11::the daily beacon::0

32593::1997/09/16::TENNESSEE VS. ALABAMA FOOTBALL GAME::0
32594::1997/09/16::10/18/97 TENNESSEE VS. ALABAMA FOOTBALL GAME AT BIRMINGHAM::0
32595::1997/09/16::ALUMNI HEADQUARTERS FOR 10/18/97 TENNESSEE VS. ALABAMA FOOTBALL GAME AT BIRMINGHAM::0

499590::2000/07/10::of theses::0
499591::2000/07/10::preparation of theses::0
499592::2000/07/10::Guide to the Preparation of theses::0

Search algorithms and Boolean operators are difficult for many users. The following consecutive examples of original queries, if from the same searcher, reflect an effort by adding words or using Boolean operator AND to deal with zero-hits unsuccessfully (see also Table 7):

535084::2001/03/13::ewan unix::0
535085::2001/03/13::ewan unix domain::0
535086::2001/03/13::ewan AND unix AND domain::0
535087::2001/03/13::ewan AND unix AND domain AND help AND setup::0

Many users also use Boolean syntax from other systems, such as

cheeting/plagerism
Athletics+Media+Relations+Director
"Ratios"+"Student"+"Professor"
Building/Room+Number

Our observations of users' problems with Boolean operators are consistent with the *Excite* studies.

As mentioned earlier, many queries are natural language statements with "the," "of," etc. Similarly, the following revision of the original query with exact wording seems to test whether case matters in searching:

55796::1997/11/12:: Little red riding hood::0
55797::1997/11/12:: Little Red Riding Hood::0

However, there is also evidence indicating possible systematic learning during the following three interactions likely by the same user:

249598::1999/01/05::pride of the southland::0
249600::1999/01/05::pride of the southland marching band::0
249601::1999/01/05::pride southland marching band::1
[Query 249599::1999/01/05::Anthropology. . . was not included because it was likely from a different user]

The modifications of the original query were systematic: (1) added the words "marching band" when receiving a zero-hit [perhaps intended to be more specific about the query]; (2) dropped "of the" to modify a zero-hit [this might be the case that the searcher was testing a hypothesis that "of the" were not searchable]. The following examples show several reiterations with little success (null output or too much output) there seemed little understanding of the rules]:

56772::1997/11/14::history of utk::0
56773::1997/11/14::history of ut::0
56774::1997/11/14::history of::0

56775::1997/11/14::history::100
56776::1997/11/14::history of the university::0
56777::1997/11/14::history::100

Discussion

The results of this study are certainly limited by the nature of the unobtrusively captured data, which lack session identifiers and information on users' evaluations of the search output. Thus, users' real information needs cannot be fully understood and the results should be interpreted with caution. Nevertheless, the analyses of this collection of substantial data provide useful structural descriptions and linguistic characterization of Web queries of an academic site by its user population as a whole across a four-year period. This study analyzed truly longitudinal Web query data rather than slices of queries on selected days by the *Excite* studies. In the field of IR, substantial attention has been placed on the analyses of text document collections. It is only in recent years that researchers realize that the language of Web users and their behaviors deserve research effort because an IR system cannot achieve its goals without truly knowledgeable users.

The longitudinal data show similar patterns across years indicating the perpetual behavior of Web users, especially the continued zero-hit problems (Pollock & Hockley, 1997; Shneiderman, Byrd, & Croft, 1997; Wang & Pouchard, 1997). The logs are from a Boolean Web search engine built with much stricter algorithms than the majority of today's general Web search engines. Although by deviating from the pure Boolean design a search engine can reduce the chances for null output. The other side of the coin is too much output (Cooper, 1988). The fundamental problem of users' lack of knowledge of IR basic concepts and individual systems remains unsolved. Another significant problem is that the high percentage of typos cannot be ignored in linguistic analyses. Further mining should test if certain results will change if misspelled words are corrected or excluded.

Statistical analyses of queries reveal the existence of linguistic regularity. The vocabulary of Web users is comparatively small with a large number of misspelled words and personal names. Both words and word-pairs have similar trends for observed frequencies, ranked orders in descending frequencies, and the amount of overlapping across years. Zipf's distribution expects linear fit. Our data show that the high frequencies zone in Figure 4 needs a quadratic polynomial. As the frequency decreases, more words share the frequency. The two plotted lines based on all words and unique frequencies split at approximately the middle point where frequencies are shared by more and more words. The most frequently occurring words in queries are quite different from those in written or spoken English.

The word associations measured by frequencies and mutual information of word-pairs can identify most frequently occurring words with strong associations. The word-word matrix is sparse; a set of pairs can be identified based on the associations of the words. Each set has a core of pairs identifiable by frequencies of the individual words and the pairs in combination with the mutual information values. Most pairs may well be irregular pairs, such as misspelled words. This phenomenon is similar to gradience of word classes, in which typical words center the category. Our analysis of the pair *career* and *services* shows potential uses of these measures. One use will be to cluster pairs by developing formulas and defining threshold values. Another is to build semantic networks surrounding most frequently occurring pairs, in which mutual information values specify distances between words.

User queries between this academic Website and the general popular Websites *Excite* and *AltaVista* show some similarities. The mean length of queries is short (2 words per query). The majority of the queries are unique in linguistic format (occurring once or twice). The queries become slightly longer over the years. Content-wise, the queries in the academic Website are quite different from the two popular general Web search engines according to vocabulary analysis. The majority of the queries in the academic Website involve topics related to career, academics, finances, sports, and student life on campus that typify university students' information needs.

In addition, we also observed users' mental models of Web searching and learning of the system through modifications of queries as shown in consecutive queries. One may criticize that our interpretations hazard guesses; we note that additional data are needed to support our interpretations of these iterations. We present our interpretations as hypotheses to be further tested using different data collection methods that can track individual users and their thoughts along the search. Further studies of Web users also need to explore how the system can embed instruction and learning in iterations.

Some implications for system design are worth noting. First, users need help with correcting typos, syntax errors, as well as contextual explanations or feedback to build correct conceptual models of Web searching. Second, user vocabulary from the queries of a specific Website can provide a solid base for building the effective end-user thesauri proposed more than fifteen years ago by Bates (1986). Bates advocates that such end-user thesauri should provide entry vocabulary geared to users' propensities (p. 369). This study provides the methods and techniques to identify such user vocabulary based on real data. Third, knowledge of lexicon and word associations of Web queries can provide guides for how a site should represent and organize its contents, especially the frequently queried categories. Fourth, measures of word associations have the potential for clustering queries into categories and building semantic networks, which can improve search effectiveness and help users to map information needs and reiterate queries.

Regarding research on Web user queries, analyses should move beyond lexical units (words and word-pairs) and toward a conceptual level. We intend to continue our analysis to explore query clustering at the conceptual level using

the WordNet synset file to transform words into concept classes. WordNet is an electronic lexical database developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller (Fellbaum, 1998).

Our data model parses queries at the finest granularity, while previous Web query studies have not shared their data models as pointed out by one of the referees of this paper. Jansen and Pooch (2001) calls for consistency in levels of analysis and common metrics to facilitate valid comparisons of results. To reach this goal, it is necessary to look closely at and debate on the various parsing techniques and statistical analyses implemented by individual studies to derive standard methods and techniques so that the results from these studies are more comparable and can be better integrated. Mining query behaviors should become an integral part of search engine design. Although relational database technologies provide a solid basis for data storage and analysis, tools must be developed to automate data cleansing and loading, and to build interfaces for interactive analysis and visualization.

Conclusions

The Web queries entered in the academic site show seasonal changes in terms of the number of queries and the topics; some of the trends are predictable based on the academic cycle and local events. There was little change in either topics or searching behaviors over the four years. The analysis of the vocabulary reveals that rank and frequency plots in natural logarithm have a good linear fit for the lower frequencies, where more and more words share the same frequency. Typos have been a sustained problem for failed search and the major part of the low frequency words. Word associations can provide the basis for building semantic network of words as a part of end-user thesauri. The mutual information statistics along with frequency are useful in identifying significant word pairs in a set made of pairs that share a word with the high frequency pair. The difficulty is to find the threshold because mutual information value varies from set to set.

The results of this study are applicable to university Websites and sites with similar users. To improve Website design, the content of a site should contain topics identified from the queries. The organization of the topics should take into account the ranking by frequency and make the frequently searched topics more accessible. These topics should be clustered into broader categories as well as broken down into facets or narrower topics. To promote accessibility, WebPages indexes should assign higher weights to terms in user vocabulary. Word associations can be used for automatic query expansions or helping users to refine queries. Query analysis should be an integral part of the Web search engine. Appropriate data mining tools are needed in addition to a good data model.

Acknowledgments

This project was partially supported by a grant from the Computational and Information Sciences Interdisciplinary Council, The University of Tennessee at Knoxville. Special thanks go to Jennifer L. Bownas for research assistance and Rick D. Thursby for pilot data processing using the inverted file structure. Thanks also due to Livy I. Simpson and Sarah D. Sewell for editorial help. The valuable comments from three anonymous referees are highly appreciated.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: Addison-Wesley.

Bates, M.J. (1986). Subject access in online catalogs: a design model. Journal of the American Society for Information Science, 37(6), 357–376.

Callon, M., Courtial, J., Turner, W.A., & Bauin, S. (1983). From translations to problematic networks: an introduction to co-word analysis. Social Science Information, 22(2), 191–235.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In Uri Zernik, (Ed). Lexical acquisitions: Exploiting on-line resources to build a lexicon. Hillsdale, NJ: Lawrence Erlbaum Associates, p. 115–163.

Cooper, W.S. (1988). Getting beyond Boole. Information Processing and Management, 24(3), 243–248.

Crystal, D. (1987). The Cambridge encyclopedia of language. Cambridge: Cambridge University Press.

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. Cambridge, MS: The MIT Press.

Jansen, B.J., & Pooch, U. A review of Web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology, 52(3), 235–246.

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.

Khoo, C.S.G., Dai, Y., & Loh, T.E. (2002). Using statistical and contextual information to identify two- and three-character words in Chinese text. Journal of the American Society for Information Science and Technology, 53(5), 365–377.

Korfhage, R.R. (1997). Information storage and retrieval. New York: Wiley.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. Research Policy, 18(4), 209–223.

Norman, D. (1983). Some observations on mental models. In Dedre Gentner and Albert L. Stevens, (Eds.). Mental Models. London: Lawrence Erlbaum Associates Publishers. p. 7–15.

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.

Pollock, A., & Hockley, A. (1997). What's wrong with Internet searching. D-Lib Magazine, 3, 1-5. Internet. http://www.dlib.org/march97/bt/03pollock.html. Accessed on June 14, 1997.

Ross, N.C.M. & Wolfram, D. (2000). End user searching on the Internet: an analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information Science and Technology, 51(10):949–958.

Shneiderman, B., Byrd, D., & Croft, W.B. (1997). Clarifying search: a user-interface framework for text searches. D-Lib Magazine, (1), 1–18. Internet. http://www.dlib.org/dlib/january97/ retrieval/01shneiderman. html. Accessed on June 14, 1997.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR Forum, 33(1), 6–12.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. IEEE Computer Magazine, 35(3), 107–109.

Spink, A., Wolfram, D., Jansen, 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.

Sproat, R., & Shih, C.L. (1990). A statistical method for finding word boundaries Chinese text. Computer Processing of Chinese & Oriental Languages, 4(4), 336–351.

SWISH. (2003). Originally developed by Kevin Hughes is currently maintained and enhanced by a team at http://www.swish-e.org/ (Accessed on February 21, 2003).

Wang, P., & Pouchard, L. (1997). End-user searching of Web resources: problems and implications. In Proceedings of the 8th ASIS SIG/CR workshop, Washington, DC (pp. 73–85). Medford, NJ: Information Today.

Wolfram, D., Spink, A., Jansen, B.J., & Saracevic, T. (2001). Vox populi: the public searching of the Web. Journal of the American Society for Information Science and Technology, 52(12), 1073–1074.