A Dual-approach to Web Query Mining: Towards Conceptual Representations of Information Needs

Peiling Wang
School of Information Sciences
College of Communication and Information
University of Tennessee, Knoxville

Final Report
2005 OCLC/ALISE Library and Information Science Research Grant Project

March 31, 2006

Suggested citation:
Wang, Peiling. 2006. "A Dual-approach to Web Query Mining: Towards Conceptual Representations of Information Needs." 2005 OCLC/ALISE research grant report published electronically by OCLC Online Computer Library Center, Inc. Available online at: http://www.oclc.org/research/grants/reports/2005/wang-p.pdf

ABSTRACT

Using a data mining approach, this project analyzed two Web query corpora collected from a university Website. One corpus includes 0.5 million queries (1997-2001); the other includes 5 million queries (2002-2004). The objectives of the project are (1) to identify appropriate methods for conceptual representations of information needs based on Web query corpora, and (2) to develop tools for effective interactions between the searcher and the IR system. The project team experimented with a dual approach, which combines quantitative and qualitative techniques, for mining longitudinal Web query corpora. This report presents (1) a description of the query corpus in relation to searching behaviors and comparison of the two corpora, (2) algorithms and techniques for identifying consecutive queries likely submitted from the same user, and query clusters representing users' topical needs, (3) methods for visualizing knowledge structures using concept maps. The results of this project have implications for advancing Web query analysis and developing new methods to enhance searchers' interaction with the Web.

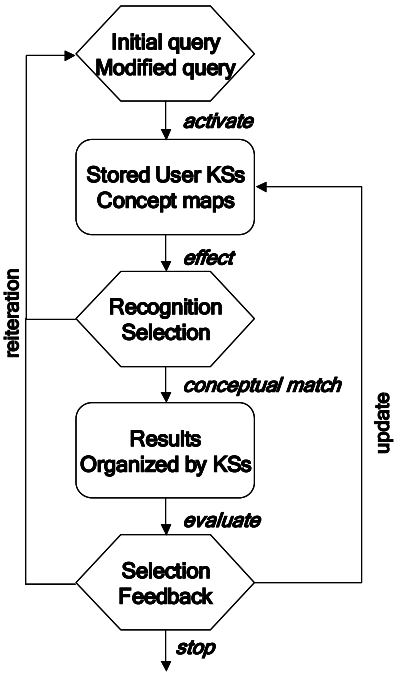

Fig. 1. User-Web Interaction Model

I. The project and completed tasks

The project is designed based on the User-Web Interaction Model (Figure 1). It proposed to carry out the following tasks and all are completed:

1. Comparison of the two corpora from two periods of time: 0.5M (1997-2001) and 5M (2002-2005) to discover patterns and changes over time (data mining) -- Completed October 2005.

2. Identification of queries looking for similar information (query clustering) – Completed December 2005.

3. Construction of concept maps to represent information needs in clustered user queries – Completed March 2005.

II. Research Findings

The Web queries were logged from October 14, 2002 to January 31, 2005. Except for missing data for the days when the Web server was migrated to a new machine or software was updated causing log file corruption, we collected a corpus of 4,597,478 queries (hereafter the 5M). In the 5M corpus 3,681,711 (80.1%) queries were duplicate (repetitive) queries of the 915,767 (19.9%) unique (distinctive) queries. A small number of 1,259 (<0.1%) queries were empty or not English characters. The unique queries provide a basis for linguistic analysis and for modeling information needs. The length of these queries ranges between one and 36 words with a mean of 2.03 words per query. (The descriptive statistics for 0.5M are in Wang, Berry, & Yang, 2003).

The following three subsections present our major findings: (1) comparison of the two corpora; (2) strategies for clustering queries looking for similar information; (3) construction of concept maps to visualize information needs.

(1) Comparison of the Two Corpora

The two corpora show both similarities and differences in searching behaviors. Queries are slightly longer than 0.5M corpus (Table 1): one-word queries slightly reduced from 39% (0.5M) to 36% (5M); two-word queries also slightly dropped from 42% to 40%; queries containing three words or more increased 2%. The majority of the queries now consist of 2 and 3 words (Table 1). The mean number of words is 2 for both corpora. The longest query in the 0.5M corpus is 131 characters and in the 5M corpus is 36 characters.

Table 1. Query Distribution by Number of Words, in Percentage

Number of words in query	0.5M (1997-2001)	5M (2002-2005)
Mean	**1.88**	**2.06**
Empty, non English, or Default	0.9	1.0
One-word	38.8	35.6
Two-word	41.5	39.9
Three-word	13.4	15.1
Four or more words	6.2	8.3
Total	***100.8***	***98.9***

Notes

- The default text in search box is "Enter search terms" (See Figure 2 and Table 2)
- The total does not add to 100 due to rounding.

The high frequency queries in the two corpora show differences and similarities. Table 2a lists the top 30 queries with descending ranking by frequency. Twelve of them (in bold font) occurred in both corpora with slightly different rank positions. Between the word variations, both the singular form *transcript* and its plural form *transcripts* made the cut, but the latter was entered twice as much as the former. The number two query in the 5M corpus flags a design problem of the Web engine. Designed as a reminder to the searcher in the textbox, *Enter search terms* was executed daily as a result of searchers simply hitting the Go button to go to the search Engine page (Figure 2).

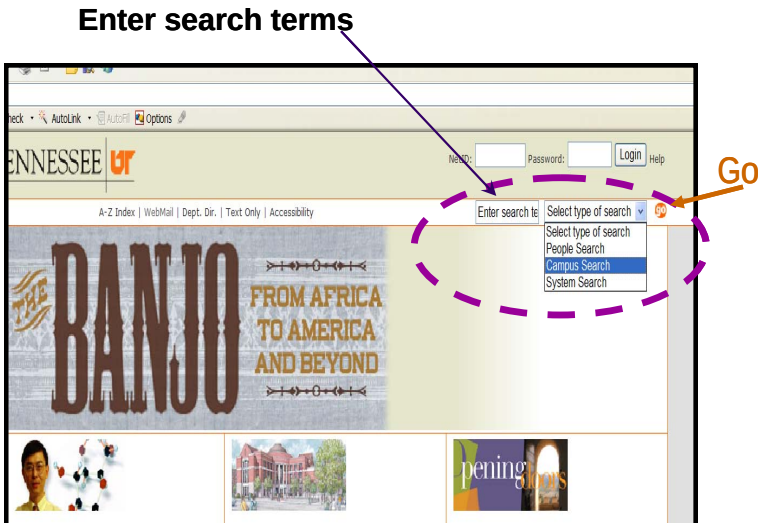

Figure 2. "Enter search terms"the reminder is the top query

Table 2a. Top Queries Compared

Rank	0.5M Corpus		5M Corpus	
	Query	Frequency	Query	Frequency
1	**career services**	9587	blackboard	74316
2	grades	5727	*Enter search terms*	45564
3	**tuition**	4837	**housing**	45270
4	**housing**	4203	circle park	32552
5	**timetable**	4097	registrar	28284
6	**bookstore**	3453	**tuition**	26312
7	rocky top	2581	**career services**	21327
8	**transcripts**	2340	**bookstore**	20507
9	**daily beacon**	2312	**timetable**	20207
10	**employment**	2156	**transcripts**	19436
11	cheerleading	1985	circle park online	17023
12	band	1914	calendar	15178
13	registration	1683	map	15159
14	**scholarships**	1537	campus map	14992
15	jobs	1488	**financial aid**	13994
16	football tickets	1465	online	13514
17	career	1407	**daily beacon**	12126
18	marching band	1397	**football**	12015
19	cheerleaders	1377	cpo	11446
20	resume	1375	**transcript**	11260
21	**financial aid**	1331	**employment**	11088
22	webmail	1317	academic calendar	10827
23	tickets	1225	parking services	9250
24	**transcript**	1211	**scholarships**	9162
25	**catalog**	1187	library	7982
26	Tennessee 101	1058	computer store	7933
27	football	1025	admissions	7439
28	biology	1000	oit	7416
29	sororities	983	philosophy	7208
30	anthropology	970	**catalog**	7149

Notes:

- All queries are aligned to lower case with the exception of "Enter search terms" (see text for explanation)
- Rank 19 for 5M is cpo (circle park online); rank 28 is oit or OIT (Office of Information Technology)
- The queries that occurred in both corpora are in bold font

Table 2b. Top Queries that Occurred in Both Corpora

Query	Rank in 0.5M	Rank in 5M
career services	1	7
tuition	3	6
housing	4	3
timetable	5	9
bookstore	6	8
transcripts	8	10
daily beacon	9	17
employment	10	21
scholarships	14	24
financial aid	21	15
transcript	24	20
catalog	25	30

Notably, several high frequency queries in the 5M corpus seem to reflect the change in registration method at the university. Since 2000, the registration system gradually moved from multi-mode (paper, phone, walk-in) to single-mode registration (online only). Online registration is called Circle Park Online or CPO, thus the frequently searched queries (Table 2a).

Both corpora show seasonal patterns. We identified queries following the registration cycle, as the example in Figure 3 shows. The big peak for August 2004 reflects the freshman enrollment increase in fall 2004.

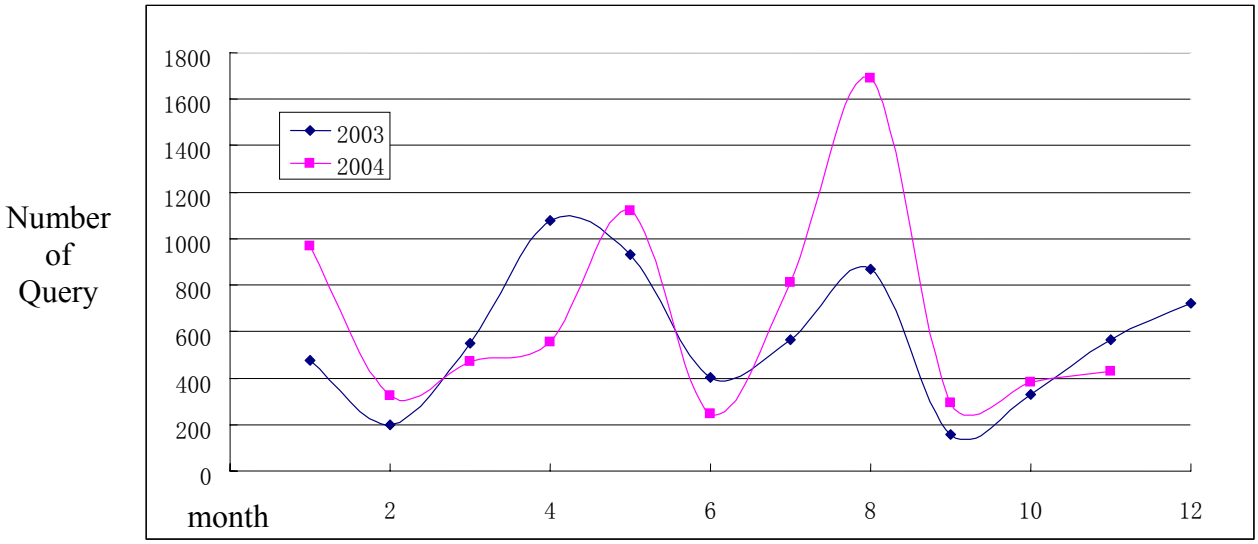

Figure 3. distribution of query *circle park online(17023)*

A comparison of the top token words indicates that *of* remains the number one word among unique queries. There are seven tokens (35%) on the top-20 lists of both corpora (Table 3).

Table 3. Top 20 Tokens Compared

Rank	Token in 0.5M	Token in 5M
1	**of**	**of**
2	services	**and**
3	career	**for**
4	**student**	the
5	**and**	in
6	grades	**student**
7	school	**ut**
8	tuition	2003
9	housing	2004
10	football	to
11	timetable	university
12	schedule	course
13	center	school
14	office	**graduate**
15	band	**tennessee**
16	**for**	a
17	department	center
18	**ut**	education
19	**tennessee**	summer
20	**graduate**	program

Notes:

- All tokens are aligned to lower case
- Tokens in bold fonts are found in both corpora
- For single word queries (Tables 2a, 2b), the word is the token

The unique queries were parsed into adjacent pairs. If a query has two words, they form one pair. Thus, three-word queries form two pairs: w_1w_2 and w_2w_3; four-word queries form three pairs: w_1w_2, w_2w_3, and w_3w_4, and so on.

Table 4 lists the top 20 pairs from both corpora. There are 10 pairs (50%) that occurred in both corpora. If the order (position) of the pair is differentiated, 5 pairs (25%) occurred in both corpora.

Table 4 Top Pairs Comparison

Rank	0.5M corpus (1997-2001)		5M corpus (2002-2005)	
1	*university*	*of*	**of**	**the**
2	**of**	**tennessee**	*of*	*university*
3	**college**	**of**	**of**	**tennessee**
4	**of**	**the**	tennessee	university
5	**department**	**of**	**college**	**of**
6	e	mail	and	of
7	*office*	*of*	edu	utk
8	graduate	school	**department**	**of**
9	*school*	*of*	2003	fall
10	phone	number	2004	fall
11	center	for	*of*	*office*
12	*on*	*campus*	is	what
13	high	school	**how**	**to**
14	the	university	*of*	*school*
15	**how**	**to**	2003	spring
16	class	schedule	*aid*	*financial*
17	Web	Page	*campus*	*on*
18	rocky	top	title	title
19	*financial*	*aid*	in	the
20	Neyland	Stadium	2004	spring

Note:

- All pairs are aligned to lower case
- The pairs in **bold** fonts are found in both corpora
- The pairs in *italic* fonts are in both corpora in reversed order

(2) Two Strategies for Web Query Clustering

The most significant contributions this project made to current research on Web user queries are the strategies derived to identify Web queries that are likely looking for the same information. We derived two major strategies: (A) identification of consecutive queries in the same session by the same user; (B) mutual information based on word-pair statistics.

(A) *Consecutive queries*. Table 5 shows a series of five queries submitted by the same searcher. The searcher reiterated the original query 4 times by changing search terms, dropping terms and adding terms, or adding previously dropped terms. The information need is represented with four unique concepts in this series of queries.

Table 5. A Series of Queries in a Session

08/11/2003	**20:38:10**	**apartments**	
	20:42:53	**off campus housing**	**283**
	20:43:24	**off campus housing**	**31**
	20:43:55	**real estate, +off +campus**	**31**
	20:44:25	**real estate, +off +campus, apartments**	**30**

(B) *Word pair and Mutual Information*. When queries consist of two words, they are parsed to form word pairs. The co-occurrence of words provides information on their dependence. Mutual information (MI) is defined as a quantity that measures the mutual dependence of the two words. The measure is quantified by the formula:

$$MI(w_1, w_2) = \ln (P(w_1, w_2)/ (P(w_1)*P(w_2))$$

where $P(w_i)$ is the probability of word w_i

$$P(w_1) = F_1/Q,$$

$$P(w_2) = F_2/Q,$$

$$P(w_1, w_2) = F_{12}/Q'$$

where F_i is the frequency of word w_i

$$Q' = \sum_{i=2}^{n} Q_i(i-1))$$

where Q_i is the total number of queries containing *i* words. *n* is the number of words in the longest query. The formula for Q' is dependent on the parsing algorithm. In the previous analysis of the 0.5M corpus (Wang et al., 2003), the pairs were formed by adjacent words and words with one intervening word. For this project, the pairs in the 5M corpus are formed by only adjacent words (See also Page 8). The rationale for adopting this algorithm is to preserve the original word pair patterns as they occurred in the queries. According to this principle, grammatical structure words (of, the, on ...) are not removed in parsing pairs.

We first focus on high frequency pairs by selecting one pair that has the potential to bring out a good set of conceptual words. For example, *course description* is selected and a cluster is derived by expanding the pairs to include other words that are paired with *course*. We found that MI values can be used to cluster the strength of the dependency (Table 6). That is, the low-frequency word can predict the high-frequency word. In Table 6, the first 4 pairs have comparatively lower frequencies than that of the pair *course description*. The corresponding MIs are much higher than the high-frequency pair. Figure 5 depicts the network with selected pairs derived from *course* (including *courses*). In addition, we also found that MI value ranges vary across clusters. Therefore, a normalization algorithm is needed to further this line of analysis. The continuation of this analysis is supported by a National Leadership Grant from the IMLS (http://web.utk.edu/~peilingw/IMLS).

(3) Construction of Concept Maps

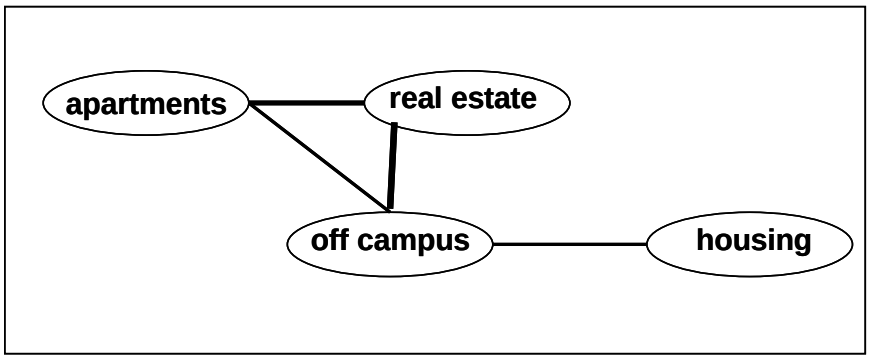

Figure 4. Semantic Network of A Single User

Algorithm A – Single User Session

The identification of queries in a single session enables us to build a semantic network for a topic. The four unique queries in Table 5 include four concepts representing needs by a user in a search session. *Algorithm A* states that if the two concepts occurred in the same query, there is a link between the two, and the strength of the link is based on the co-occurrence of concepts. Figure 4 depicts the network in which three concepts, *apartment*, *real estate*, and *off campus* are interconnected, but *housing* links only with one concept, *off campus*.

Algorithm B – Mutual Information of Word Pairs Derived from Unique Queries

Unique queries provide linguistic features of user queries. Each unique query may have a different level of popularity which the query frequency can measure. The top queries are selected using such measures. However, regardless how popular a query may be, it represents one statement for an information need. Therefore, the linguistic structures can be observed closely to identify patterns. One such observation in this project is to use two measures to identify semantic associations of words: mutual information and word-pair frequency. These two measurements allow words to cluster and be linked in a web. *Algorithm B* chooses a word pair from the output of high frequency F_{12} word pairs (such as top 10, 50, 100, etc); the pair is then expanded to include additional pairs that contain either word. Table 6 lists the pairs derived from (or including) *course(s)*. Three clusters are identified based on mutual information. These pairs are depicted in Figure 5 based on the strength of association between each pair. The first cluster includes four pairs with MI values greater than 2. The second cluster includes five pairs with MI values between 1 and 2. In the second cluster, one additional pair is introduced to illustrate the potential propagation of the network. For example a strong association is found between *degree* and *offer/ed/ing*. This recursive process can continue so that the cluster will grow bigger. To generate a meaningful and manageable size of the network, a strategy is needed to make an appropriate cut point; that is, to set an appropriate threshold for exclusion.

Table 6. Three clusters of pairs derived from *course*

Pair	F_{12}	F_1	F_2	MI
course outline	87	16720	229	2.31
course offer/ed/ing	432	16720	1277	2.19
repeat course	33	103	16721	2.14
correspondence course	74	257	16720	2.03
required course	219	1250	16757	1.53
course equivalencies	123	16720	838	1.36
course description	2610	16720	4057	1.33
course evaluation	188	16720	1531	1.18
course available	55	16720	533	1.00
online course	599	6335	16812	0.91
course catalog	730	16720	3678	0.82
course listing	430	16720	5107	0.80
course schedule/timetable	908	16721	11164	0.77
drop / withdraw	143	16720	2225	0.53
course syllabus	371	16756	6139	0.47
blackboard course	37	906	16720	0.13
add course	40	1284	16720	0.01
degree offer/ed/ing	*106*	*5857*	*956*	*2.13*

Note:
Morphemic variations are aggregated. Clusters are based on MI. The last pair *degree offer/ed/ing* here is derived from *offer/ed/ing,* not from *course*.

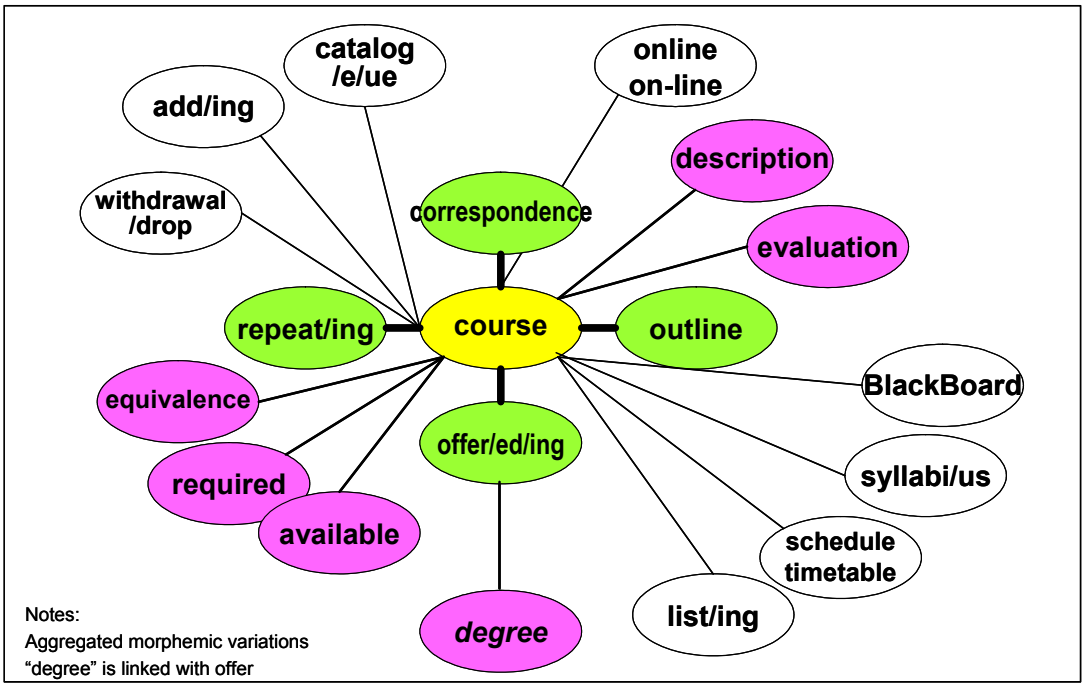

Figure 5. Semantic Network based on Mutual Information

III. Future Research

Two main areas need further research following this project:

(1) A threshold is needed for identifying session boundaries..

(2) A threshold is needed for identifying clusters. Is there a normalization strategy that provides a basis for calculating a cut point across different clusters?

IV. Research Outcomes

This research has resulted in a successful research proposal funded by the IMLS National Leadership Grant in the amount of $199,995 for two years, 2006-2007. The co-PIs are Dietmar Wolfram and Jin Zhang at the University of Wisconsin, Milwaukee. The IMLS project will expand the analysis and collect new data to compare user queries submitted to three different search engines: the academic, general, and medical Websites. The goal is to build a new interaction model that is less complex than the traditional retrieval systems and more effective than the simplistic Web search engines. This new model is grounded on users' needs and query behaviors and incorporates algorithms for conceptual matching.

The preliminary results have been presented at two conferences:

1. Wang, Peiling and Wu, Lei (2006) *Presenter*. OCLC/ALISE 2005 LISRGP Award Papers. Association for Library and Information Science Education 2006 Annual Conferences (January 19, San Antonio, TX)

2. Wang, Peiling (2005) *Panelist*. Internet usage transaction log studies: The next generation. American Society for Information Science and Technology 2005 Annual Meeting (October 31, Charlotte, NC).

Currently, a paper is in preparation to be submitted to an appropriate journal or conference. This grant will be acknowledged.

V. Budget Report

Submitted in a separate file.

References

Wang, Peiling; Berry, Michael, and Yang, Yiheng. Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*. 2003, 54(8): 743-758.

Wang, Peiling; Bownas, Jennifer, and Berry, Michael W. Trend and behavior detection from Web queries. In Berry, Michael W., ed. *Survey of text mining: clustering, classification, and retrieval*. New York: Springer; 2004. 173-183.