Patterns and Transitions of Query Reformulation during Web Searching

Bernard J. Jansen
Email: jjansen@ist.psu.edu
Fax: +1 814 865 6426
College of Information Sciences and Technology
The Pennsylvania State University
329F Information Sciences and Technology Building
University Park, PA 16801, USA

Mimi Zhang
Email: mzhang@ist.psu.edu
Fax: +1 814 865 7492
College of Information Sciences and Technology
The Pennsylvania State University
321D Information Sciences and Technology Building
University Park, PA 16801, USA

Amanda Spink
Email: ah.spink@qut.edu.au
Fax: + 61 7 3864 2703
Faculty of Information Technology
Queensland University of Technology
Gardens Point Campus, 2 George Street, GPO Box 2434
Brisbane QLD 4001 Australia

***Abstract*—**

Purpose: To investigate and identify the patterns of interaction between searchers and search engine during Web searching.

Design: We examined 2,465,145 interactions from 534,507 users of Dogpile.com submitted on May 6, 2005. We compared query reformulation patterns. We investigated the type of query modifications and query modification transitions within sessions.

Findings: We identified three strong query reformulation transition patterns: between (1) specialization and generalization, between (2) video and audio, and between (3) content change and system assistance. In addition, our findings show that Web and images content were the most popular media collections.

Value: This research sheds light on the more complex aspects of Web searching involving query modifications.

***Index Terms* — Query modification, query reformulation, search engine, Web search, pattern**

I. Introduction

The interaction between user and search engine is a learning process (Jansen, Smith, & Booth 2007), an exploratory process (Stojanovic 2005), and an iterative interactive process (Rieh & Xie 2006, p.752). All of these characteristics are due to the nature of Web search engine. Since there is only partial overlap between the knowledge of the user and that of search engine (Amitay et al 2005), users have to constantly adjust their queries as they approach the information they need. Therefore, query reformulation is a most popular search tactics. Query reformulation is a crucial step in fulfilling users' information needs during Web searching.

The purpose of our study here is to advance our understanding of query reformulation, to expand our knowledge in predicting the future actions of searchers on searching systems, and to improve the searching system capability of providing searching assistance at the right time. Specifically, we aim to determine the query modification patterns through which users search for information on Web searching systems. We refer to each query modification event during a session as a search state. We will investigate how these patterns transit between each other and if there is any identifiable transition pattern.

Our research is important because if a Web searching system can predict the future state of searchers, the system can provide targeted searching assistance to aid searchers in their information seeking task. If we can determine an appropriate order of the search process (i.e., number of predictive states), this indicates an upper bound for prediction, which will provide us the most predictive power with the least computational complexity.

In the following sections, we first review relevant prior work in the field and present our research questions. We then address our research design and data analysis. Following this, we discuss our research results. We end with implications and future research.

II. Relevant Studies and Research Question

Different aspects of query reformulation have been studied by various researchers. Bates (1979a, 1979b, 1990), Fidel (1985, 1991), Rieh and Xie (2006), and Chen and Dhar (1990) focused on theorizing query reformulation, defining it, and identifying the patterns. Jansen, Spink, Saracevic, and Zhang (Jansen, Spink, & Saracevic 2000; Spink & Jansen 2004; Zhang, Jansen, & Spink 2006) worked on analyzing a Web transaction log to discover the features of query reformulation and its patterns on the Web. Gauch and Smith (1991) and Amitay et al (2005) designed information search systems that automated the query reformulation process. Church et al (2007) studied query reformulation in mobile search domain.

A. Web transaction log analysis

In the early age of Web log analysis, researchers focused on the statistics of simple query reformulation classification, like how many queries have been modified, and how often the users made modifications. Jansen Spink, and Saracevic (2000) conducted the Excite log analysis and reported 67 percent users submitted only one query and 19 percent users made only one modification (i.e. submitted two queries). Thirty-five percent of the queries were the initial queries, and 22 percent were the modified queries. Among those modified queries, 34.76 percent queries were of the same length as the previous ones, but 19.03 percent added one more term to the previous queries. There were 16.33 percent that subtracted a term from the previous ones.

Zhang, Jansen and Spink (2006) tried to advance the understanding of query reformulation. They reported their examination on patterns and features of query reformulation. They developed a list of query reformulation patterns based on linguistic, information retrieval behavior, information seeking behavior knowledge, and the observation of data. Query reformulation patterns are mainly identified from the linguistic aspect, for example, subtracting noun or adding verb after term. The stratified sample sessions were extracted from an AltaVista transaction log and manually coded with query reformulation pattern list. They found that changing the query topic was the primary means to modify queries, and most of the time the users were inclined to modify nouns or subtract some types of words when changes were made. The searchers appear to know how to increase and decrease the coverage (i.e., number of results retrieved) of queries.

B. Query reformulation pattern

Researchers have tried to theorize on the query reformulation process using traditional library search system. Bates defined four levels of search activities, which are move, tactic, stratagem, strategy from bottom to top according to Bates (1990). Move is “an identifiable thought or action that is a part of information searching” (Bates 1990, p.578). Tactic is “one or a handful of moves made to further a search” (Bates 1990, p.578). Stratagem is defined as “a larger, more complex set of thoughts and/or actions than the tactic; a stratagem consists of multiple tactics and/or moves, all designed to exploit the file structure of a particular search domain thought to contain desired information" (Bates 1990, p.578). Strategy is "a plan, which may contain moves, tactics, and/or stratagems, for an entire information search" (Bates 1990, p.578).

For search tactics, they could be monitoring tactics, file structure tactics, search formulation tactics, terms tactics or idea tactics. (Bates 1979a, 1979b, 1990) Monitoring tactics refer to means to guarantee that the search on the right path and effective. Files structure tactics are employed to get "desired file, source, or information" (Bates 1979b, p.207) via "the file structure of the information facility" (Bates 1979b, p.207). Search formulation tactics are techniques assisting people in reformulating queries. Terms tactics refer to ways of choosing and modifying terms during search formulation process, which include SUPER, SUB, RELATE, REARRANGE, CONTRARY, RESPELL, and RESPACE. Idea tactics assist in developing new ideas or solutions to problems during information searching process. (Bates 1979a, 1979b, 1990)

Fidel (1985, 1991) sorted query reformulation into two categories: operational moves and conceptual moves. Operational moves refers to query modifications that employ the same meaning of query components. In contrast, conceptual moves change the meaning of query components. People who prefer one type of move are called operationalists or conceptualists correspondingly. The moves can also be classified into three categories: moves to reduce the size of a set, moves to enlarge the size of a set, and moves to simultaneously increase both precision and recall.

Fidel (1991) conducted a case study on 47 professional searchers working on their job-related searches. From the observation and analysis of verbal and search protocols, she found that when people employed operational and conceptual moves to improve recall, they were more likely to used operational moves to improve precision or to reduce recall. People employed conceptual ones to improve both precision and recall. Operationalist was inclined to employ textwords, avoid consulting the thesaurus and make fewer recall moves (i.e., moves to increase recall) compared with the conceptualist.

Chen and Dhar (1990) studied the query refinement process of an online catalog system. In their study, they employed the semantic network to represent the semantic contents of queries and a Problem Behavior Graph to describe the flow of search process. In this representation system, there is a concept of "semantic operator" similar to our concept of query modification. They referred it as "moves or actions that change the content of the query" (Chen & Dhar 1990, p.121) and identified five operators: synonymous term operator, broader term operator, narrower term operator, adjacent term operator, and disjointed term operator.

However, most of these studies were on traditional library systems rather than commercial search engines. Although there are similarities, these two sets of system are also different. For the library system, it is easy to identify individual user via his/her login information. The information content of the system is limited and more systematical. On the Web, there is almost no boundary of Web information source. The single user is hard to identify. We cannot simply map results from library research directly to the Web sphere.

However, one common aspect they have is the discussions on the specialization and generalization concepts. For example, SUPER and SUB in Bates' papers (1979a, 1979b, 1990), moves to enlarge the size of a set and moves to reduce the size of a set, also noted by Fidel (1985, 1991). There are the broader term operator and narrower term operators in Chen and Dhar's (1990) research.

Our reformulation pattern classification will include these two patterns as well. Ours will include more contemporary means of modifying queries. We develop the classification method for queries based on prior research in Web search (Jansen, Spink, & Saracevic 2000).

Our study here follows this direction and adopts an automated method to study the reformulation pattern. Our research question is: *How people modify their queries during the interaction with the search engine*? We investigated the manner of query modification during sessions. We provide aggregate results for each query classification category, and then extend this first level classification by analyzing intra-session query transactions (i.e., movement from one type of query to the next).

III. Research Design

A. Web Data

Dogpile.com (www.Dogpile.com) is a meta-search engine, owned by InfoSpace, Inc. When a searcher submits a query, Dogpile.com simultaneously submits the query to multiple other Web search engines, collecting the results from each, removing duplicates results, and aggregating the remaining results into a combined ranked listing using a proprietary algorithm. Dogpile.com integrates the results of the four leading Web search indices (i.e., Ask, Google, MSN, and Yahoo!) along with other search engines into its search results listing. Meta-search engines provide a unique service by presenting the alternate results provided by the various search engines, which have a low rate of overlap (Spink et al 2006).

B. Data Collection

We collected the records of searcher – system interactions in a transaction log that represents a portion of the searches executed on Dogpile.com on May 6, 2005. The original general transaction log contained 4,056,374 records, each containing seven fields:

- *User Identification*: a code to identify a particular computer.
- *Cookie*: an anonymous cookie automatically assigned by the Dogpile.com server to identify unique users on a particular computer.
- *Time of Day*: measured in hours, minutes, and seconds as recorded by the Dogpile.com server on the date of the interaction.
- *Query Terms*: the terms exactly as entered by the given user.
- *Location*: a code representing the geographic location of the user's computer as denoted by the computer's Internet Protocol (IP) address.
- *Source*: the content collection that the user selects to search (e.g., *Web, Images, Audio, News,* or *Video*), with *Web* being the default.
- *Feedback*: a binary code denoting whether or not the query was generated by the *Are You Looking for?* query reformulation assistance provided by Dogpile.com.

We imported the original flat ASCII transaction log file of 4,056,374 records into a relational database. We generated a unique identifier for each record. We used four fields (*Time of Day*, *User Identification*, *Cookie*, and *Query*) to locate the initial query and then recreate the sequential series of actions from a particular user, determined by *User Identification* and *Cookie*. An analysis of the dataset shows that the interactions of Dogpile.com searchers was generally similar to Web searching on other Web search engines (Jansen, Spink, & Koshman, 2007).

C. Data Preparation

The terminology that we use in this research is similar to that used in other Web transaction log studies (Jansen & Pooch 2001; Park, Bae, & Lee 2005). For this research, we are interested in queries submitted by humans, and the transaction log contained queries from both human users and agents. There is no acknowledged methodology for precisely identifying human from non-human submissions in a transaction log. Therefore, researchers normally use a temporal or interaction cut-off (Silverstein et al 1999).

We selected the interaction cut-off approach by removing all sessions with 100 or more queries. This cut-off is substantially greater than the reported mean number of queries (Jansen, Spink, & Saracevic 2000) for human Web searchers. This cut-off most likely introduced some agent sessions; however, we were reasonable certain that we had included most of the queries submitted primarily by human searchers.

D. Data Analysis

We used an algorithm to classify content changes within sessions utilizing the fields: *IP*, *Cookie*, and *Content Change*. We used a contextual method to identify sessions. We used the searcher's IP address and the browser cookie to determine the initial query and subsequent queries. Instead of using a temporal cut-off, we used changes in the content of the user queries.

For this method, we assigned each query into a mutually exclusive group based on an IP address, cookie, query content, use of the feedback feature, and query length. The classifications are:

- *Assistance:* the current query was generated by the searcher's selection of an *Are You Looking For?* query (see www.dogpile.com).
- *Content Change*: the current query is identical but executed on another content collection.
- *Generalization:* the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more general information.
- *New:* the query is on a new topic.

- *Reformulation:* the current query is on the same topic as the searcher's previous query and both queries contain common terms.
- *Specialization*: the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more specific information.

The *initial query* (Q_i) from a unique IP address and cookie always identified a new session. In addition, if a *subsequent query* (Q_{i+1}) by a searcher contained no terms in common with the previous query (Q_i), we also deemed this the start of a new session. Naturally, from an information need perspective, these sessions may be related at some level of abstraction.

However, with no terms in common, one can also make the case that the information state of the user changed, either based on the results from the Web search engine or from other sources (Belkin, Odd, & Brooks 1982). Also, from a system perspective, two queries with no terms in common represent different executions to the inverted file index and content collection. We classified each query using an application that evaluated each record in the database. We built our algorithm from the concept presented in the paper from He, Göker, and Harper (2002).

IV. Results

A. Query reformulation pattern

Table I presents the classification of query patterns. 36.66 percent of queries are modified queries. Compared with previous finding of 22 percent modified queries (Jansen, Spink, & Saracevic 2000), more queries were modified in this dataset. Among these reformulated queries, 42.46 percent were modified by adding or deleting terms from the previous query (i.e., Reformulation: 22.73 percent; Specialization with reformulation: 9.95 percent; Generalization with reformulation: 9.78 percent). The percentage of reformulating queries is 23.28 percent (Specialization: 16.28 percent; Generalization: 7.20 percent). This results shows that people were inclined to make some changes on the previous queries by adding or subtracting terms.

There were 26.23 percent of the queries that show users' effort to narrow down the range of the results (Specialization: 16.28 percent; Specialization with reformulation: 9.95 percent), while 16.98 percent display users' attempts to get more results, 9.25 percent and about one third less than the specialization pattern (Generalization: 7.20 percent; Generalization with reformulation: 9.78 percent). This result illustrates users' slight preference to specialization. There were 22.25 percent queries generated by using system's assistance, which was the second major means of modifying queries. The percentage of query shifting between different media collections is 11.81 percent. People seemed not to have much interest in changing content once the searching process begins. Specialization with reformulation and Generalization with reformulation are combined Specialization, Generalization, and Reformulation as defined above.

We did explored users' content changes, though. Table II includes the results of analysis. Web and images were the major media collections to users. The shifts between Web and images were the two most popular content changes (Web to images: 37.21 percent; Images to Web: 21.20 percent). Switches to Web and images were the top two options in almost all the content change categories. From Web repository, 73.70 percent queries were shifted to image collection. From image collection, 70.71 percent queries were switched to Web. From audio depot, 42.92 percent queries were transited to Web and 16.22 percent were transited to images. From video collection, 47.41 percent queries were channeled toward Web. From news repository, 64.12 percent queries were altered to the Web and 21.32 percent were to images. Video and audio seemed to have a connection. The transitions between audio and video were ranked in the second place in both media collections (Audio to Video: 39.37 percent; Video to Audio: 46.04 percent). Interestingly, there was no transition from video to images. The transition to news collection was in the last place in each media collection, which depicts the unpopularity of using search engine solely on news collection.

B. Transition between query reformulation patterns

Table III shows how different types of query shift between each category. Again, *New* is the first query of a user session. From Table I, there are 964,780 new queries in our dataset. 221,278 of them were modified and 48.60 percent of queries were the unmodified new queries among the whole dataset. Among those modified, 39.45 percent queries were reformulated to narrow down the range of the results (New to specialization: 28.20 percent, New to specialization with reformulation: 11.25 percent); 17.29 percent were modified to get more results, about 1/2 of the specialization (New to generalization with reformulation: 8.93 percent, New to generalization: 8.36 percent). The proportion of specialization and generalization here is slightly higher than the proportion of these two patterns in the whole dataset. Therefore, after the initial query, users were more likely to narrow down the result list.

TABLE I

QUERY REFORMULATION PATTERNS

Search Pattern	**Occurrence**	**Percentage**	**Occurrence (excluding *New*)**	**Percentage (excluding *New*)**
New	964,780	63.34%		
Reformulation	126,901	8.33%	126,901	22.73%
Assistance	124,195	8.15%	124,195	22.25%
Specialization	90,893	5.97%	90,893	16.28%
Content change	65,949	4.33%	65,949	11.81%
Specialization with reformulation	55,531	3.65%	55,531	9.95%
Generalization with reformulation	54,637	3.59%	54,637	9.78%
Generalization	40,186	2.64%	40,186	7.20%
TOTAL	**1,523,072**	**100.00%**	**558,292**	**100.00%**

TABLE II

TRANSITION AMONG CONTENT

Content Transition	**Occurrence**	**Percentage**	**Percentage (within sub-category)**
Web to Images	12,080	37.21%	73.70%
Web to Audio	2,411	7.43%	14.71%
Web to Video	1,298	4.00%	7.92%
Web to News	602	1.85%	3.67%
Images to Web	6,882	21.20%	70.71%
Images to Video	2,096	6.46%	21.53%
Images to Audio	553	1.70%	5.68%
Images to News	202	0.62%	2.08%
Audio to Web	1,537	4.73%	42.92%
Audio to Video	1,410	4.34%	39.37%
Audio to Images	581	1.79%	16.22%
Audio to News	53	0.16%	1.48%
Video to Web	1,036	3.19%	47.41%
Video to Audio	1,006	3.10%	46.04%
Video to News	143	0.44%	6.54%
News to Web	370	1.14%	64.12%
News to Images	123	0.38%	21.32%
News to Video	59	0.18%	10.23%
News to Audio	25	0.08%	4.33%
TOTAL	**32,467**	**100.00%**	

TABLE III

TRANSITION OF QUERY MODIFICATION WITHIN WEB SESSION

Query Pattern Shift	**Occurrence**	**Percentage**	**Percentage**
		(within sub-category)	**(within entire dataset)**
New to specialization	62,405	28.20%	14.50%
New to assistance	58,471	26.42%	13.58%
New to content change	37,242	16.83%	8.65%
New to specialization with reformulation	24,895	11.25%	5.78%
New to generalization with reformulation	19,767	8.93%	4.59%
New to generalization	18,498	8.36%	4.30%
SUBTOTAL	**221,278**	**100.00%**	**51.40%**
Specialization to reformulation	13,049	32.02%	3.03%
Specialization to generalization with reformulation	9,090	22.31%	2.11%
Specialization to generalization	6,714	16.48%	1.56%
Specialization to specialization with reformulation	4,745	11.64%	1.10%
Specialization to content change	3,585	8.80%	0.83%
Specialization to assistance	3,567	8.75%	0.83%
SUBTOTAL	**40,750**	**100.00%**	**9.47%**
Reformulation to specialization with reformulation	8,826	22.70%	2.05%
Reformulation to specialization	7,450	19.16%	1.73%
Reformulation to generalization with reformulation	6,979	17.95%	1.62%
Reformulation to assistance	5,951	15.30%	1.38%
Reformulation to generalization	5,523	14.20%	1.28%
Reformulation to content change	4,160	10.70%	0.97%
SUBTOTAL	**38,889**	**100.00%**	**9.03%**
Assistance to content change	18,474	57.84%	4.29%
Assistance to reformulation	3,754	11.75%	0.87%
Assistance to generalization with reformulation	3,378	10.58%	0.78%
Assistance to specialization with reformulation	2,626	8.22%	0.61%
Assistance to generalization	2,147	6.72%	0.50%
Assistance to specialization	1,563	4.89%	0.36%
SUBTOTAL	**31,942**	**100.00%**	**7.42%**
Specialization with reformulation to generalization with reformulation	9,719	36.35%	2.26%
Specialization with reformulation to reformulation	7,423	27.76%	1.72%
Specialization with reformulation to generalization	3,788	14.17%	0.88%
Specialization with reformulation to specialization	2,251	8.42%	0.52%
Specialization with reformulation to assistance	1,899	7.10%	0.44%
Specialization with reformulation to content change	1,658	6.20%	0.39%
SUBTOTAL	**26,738**	**100.00%**	**6.21%**

Query Pattern Shift	**Occurrence**	**Percentage (within sub-category)**	**Percentage (within entire dataset)**
Generalization with reformulation to reformulation	8,455	31.91%	1.96%
Generalization with reformulation to specialization with reformulation	7,114	26.85%	1.65%
Generalization with reformulation to specialization	4,789	18.08%	1.11%
Generalization with reformulation to assistance	2,945	11.12%	0.68%
Generalization with reformulation to generalization	1,659	6.26%	0.39%
Generalization with reformulation to content change	1,531	5.78%	0.36%
SUBTOTAL	**26,493**	**100.00%**	**6.15%**
Content change to assistance	10,688	40.91%	2.48%
Content change to reformulation	4,527	17.33%	1.05%
Content change to specialization	4,096	15.68%	0.95%
Content change to generalization	2,907	11.13%	0.68%
Content change to generalization with reformulation	2,050	7.85%	0.48%
Content change to specialization with reformulation	1,859	7.12%	0.43%
SUBTOTAL	**26,127**	**100.00%**	**6.07%**
Generalization to specialization	6,790	37.17%	1.58%
Generalization to reformulation	3,278	17.95%	0.76%
Generalization to assistance	3,248	17.78%	0.75%
Generalization to specialization with reformulation	2,137	11.70%	0.50%
Generalization to content change	1,838	10.06%	0.43%
Generalization to generalization with reformulation	975	5.34%	0.23%
SUBTOTAL	**18,266**	**100.00%**	**4.24%**
OVERALL TOTAL	**430,483**	**100.00%**	**100.00%**

Generally speaking, users were more likely to expand the result lists after narrowing the queries. From specialization, 38.79 percent queries were modified to be more general (Specialization to generalization with reformulation: 22.31 percent, Specialization to generalization: 16.48 percent). Only 11.64 percent were modified to narrow the range of results. From specialization with reformulation, 50.52 percent queries were transited to get more results (Specialization with reformulation to generalization with reformulation: 36.35 percent; Specialization with reformulation to generalization: 14.17 percent); only 8.42 percent continued to be narrow down.

On the other hand, after narrowing the queries, users were inclined to expand the result lists. From generalization with reformulation, 44.93 percent queries were reformulated to reduce the range of the result (Generalization with reformulation to specialization with reformulation: 26.85 percent; Generalization with reformulation to specialization: 18.08 percent). A small percentage, (6.26 percent) of the queries, was changed to get fewer results. From generalization, 48.87 percent queries were modified to reduce the range of queries (Generalization to specialization: 37.17 percent; Generalization to specialization with reformulation: 11.70 percent). There was 5.34 percent of the dataset changing from generalization to generalization with reformulation.

Moreover, we can see from the results in Table III, that there appears to be a connection between the searcher shifting content collections and the use of system assistance with a majority (57.84 percent) of assistance usage occurring just before a content change or just after (40.91 percent) a content change. These shifts accounted for 25 percent of all assistance usage.

C. Accuracy of classification

We conducted a verification of our classification algorithm by manually classifying 2,000 queries. We arrived at five categories of errors, developed a priori:

1) *Misspelling:* a word was misspelled or a previously misspelled word causing a change resulting in a misclassification (causes a false *New* or *Reformulation*).
2) *Cookie: either cookie not defined or change in cookie but not a change in user (causes a false New).*
3) *Special character change: the original query contained special characters (causes a false New or Reformulation).*
4) *Time gap: time gap between queries was too large to be considered a session, but Q_i and Q_{i-1} were still related (causes a false New).*
5) *Other: a miscellaneous collection of other reasons (causes a false New).*

Of the 2,000 queries manually classified, there were 400 deemed to be mistakes. We see from Table IV that most of the errors were due to misspellings (i.e., the algorithm counted the word as a new term when in reality the searcher had misspelled a term in the original query and corrected the term in the subsequent query. Most misspellings occurred due to missing spaces in words. However, the sum total of all misclassifications was 4.45 percent, resulting in a 95.55 percent accuracy rate for the algorithm.

TABLE IV

MISCLASSIFICATION OF QUERY FROM 2,000-QUERY SAMPLE

Type of Misclassification	**Occurrence**	**Percentage**
Misspelling	52	47.27%
Cookie	23	20.91%
Time gap	21	19.09%
Special character change	5	4.55%
Other	9	8.18%
TOTAL	**110**	**100.00%**

V. DISCUSSION

Our study shows that the query reformulation process is like the focus adjustment in picture shooting. You have to adjust further or move closer to your model and find the best position to shoot the picture. The search engine users specialize and generalize their queries a couple of times to locate their desired information. For example, 26.23 percent of the queries were specialized and 16.98 percent were generalized. More queries were modified to generalize than to specialized. The same case happened to generalization. From generalization, queries were more likely to be specialized than to continue to be generalized. These two query reformulation patterns, generalization and specialization, are the basic patterns. Both have been identified by researchers before in domains besides other than the Web. (Bates 1979a, 1979b, 1990; Fidel 1985, 1991; Chen & Dhar 1990)

Web and images were the most popular media collections. As far as we could determine, this is one of the first analyses of content transitions in Web searching. The shifts between Web and images were the two most popular content change options. Switches to Web and images were the top two tactics in almost all the media collections. These two media collections have a longer history of usage than the others. Moreover, Web and images are the most commonly used resources. Comparatively, audio and video are more about entertainment and recreation. In addition, Web and images are the first two options on the content selection list. All of these conditions probably make them the most used media collections.

However there was no transition from video to images. Is it because video is so-called “motion picture” and has provided some visual information? The information has provided enough details for users to make judgment or fulfill their information needs. Thus, they do not need switch to images for the visual information. Therefore, when the system generates recommendations, it might not want to suggest this transition: from video to images.

Another interesting finding in transition patterns between different contents is the transition between audio and video. Each transition was ranked in the second place in both media collection and the percentages are 39.37 percent (Audio to Video) and 46.04 percent (Video to Audio). It seems there is a connection between them. They are next to each other on the content option list. These two media are roughly equally distributed. Li et al (2005) employed Web a crawler to track 17 million Web pages from different parts of the world and extract around 30,000 streaming audio and video clips. 43

percent of the media clips are audio only and 57 percent are video clips.

We believe as search technology on video is improved, more queries will be transited to video since it could embrace more diverse information compared with audio. For example, people could watch ballet over the Web. If this ballet show is broadcasted via audio, people could only hear the music. Therefore, as the bandwidth becomes wider and the search technology on video makes progress, people could prefer video to audio.

There appears to be a connection between the searcher shifting content collections and the use of system assistance with the majority of assistance usage occurring just before a content change or just after a content change. This implies that most content changes happened due to system recommendations and after switched to a new media collection. It also indicates that this is when people are most open to such assistance.

People modified queries again according to the system's assistance. Thus, the designer could utilize this finding to channel queries to certain media collection, such as news. News was found to be the least popular media collection. In another design scenario, say a search engine company wants to promote its new technology on video search. Its Web systems may want to make more suggestion on shift to search results on video collection and get a chance to show people how well their system performs. The finding on shift between different media collections due to system recommendation will certainly benefit the design and the industry.

VI. Conclusion and Future Study

To sum up, our algorithm has done a neat job of query modification classification. The accuracy rate is high at 95.55 percent. We found that more queries were modified compared with findings in prior work (e.g., Jansen, Spink, Saracevic 2000). When modifying queries, users were more likely to add or remove term(s) from the previous queries. Utilizing system assistance was one of the major modification options. Content change was not very popular among users. Shift between different contents had several obvious patterns, which are Web and images were the most popular media collections; video and audio had a connection and were often shifted between each other; there was no shift from video to images; news was the least popular media collection. After the initial queries, about half of the queries were modified, and most queries were reformulated to be more specific. After specialization, users were more likely to generalize their queries. Generally speaking, specialization and generalization alternated in the query reformulation process. There appears to be another connection between the searcher shifting content collections and the use of system assistance with the majority of assistance usage occurring just before a content change or just after a content change

For future research, the classification algorithm here may be used as a model to facilitate cross-system investigations. An attempt to standardize query reformation detection would also enhance comparative transaction log analyses. Certainly, the use of semantic classification of queries and terms is an area of future research. We are currently conducting qualitative analysis of Dogpile users' query reformulation that we will compare with the results reported in this paper.

Acknowledgement

We thank InfoSpace, Inc. for providing the Web search engine transaction log data without which we could not have conducted this research. We encourage other search engine companies to find ways to collaborate with the academic research community. We also thank Ms. Danielle Booth for coding the algorithm used in this research.

Reference

Amitay, E, Darlow, A, Konopnicki, D, and Weiss, U. (2005), "Queries as Anchors: Selection by Association", Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia, ACM Press, Salzburg, Australia, pp.193-201.

Bates, J. M. (1979a), "Idea tactics", Journal of the American Society for Information Science, Vol 30 No 5, pp.280-9.

Bates, J. M. (1979b), "Information Search Tactics", Journal of the American Society for Information Science, Vol 30 No 4, pp.205-14.

Bates, J. M. (1990), "Where Should the Person Stop and the Information Search Interface Start?", Information Processing & Management, Vol 26 No 5, pp. 575-91.

Belkin, N, Oddy, R, and Brooks, H. (1982a), "ASK for Information Retrieval, Parts 1", Journal of Documentation, Vol 38 No 2, pp.61-71.

Belkin, N, Oddy, R, and Brooks, H. (1982b), “ASK for Information Retrieval, Parts 2”, Journal of Documentation, Vol 38 No 3, pp.145-64.

Chen, H. and Dhar, V. (1990), “Online query refinement on information retrieval systems: A process model of searcher/system interactions”, Proceedings of the Thirteenth Annual International ACM SIGIR Conference, ACM Press, Brussels, Belgium, pp.115-32.

Church, K, Smyth, B, Cotter, P, and Bradley, K. (2007), “Mobile Information Access: A Study of Emerging Search Behaviour on the Mobile Internet”, ACM Transactions on the Web, Vol 1 No 1 Article 4. Retrieved September 8, 2007, from ACM Digital Library.

Fidel, R. (1985), “Moves in online searching”, Online Review, Vol 9 No 1, pp.61-74.

Fidel, R. (1991), “Searchers’ Selection of Search Keys: III. Searching Styles”, Journal of the American Society for Information Science, Vol 42 No 7, pp.517-27.

Gauch, S, and Smith, J. B. (1991), “Search improvement via automatic query reformulation”, ACM Transactions on Information Systems, Vol 9 No 3, pp.249-80.

He, D, Göker, A, and Harper, D. J. (2002), “Combining Evidence for Automatic Web Session Identification”, Information Processing & Management, Vol 38 No 5, pp.727-42.

Jansen, B. J, Smith, B, and Booth, D. (2007), “Understanding Web Search via a Learning Paradigm”, paper presented at the Sixteenth International World Wide Web Conference (WWW2007), May 8-12, Banff, Canada.

Jansen, B. J, and Pooch, U. (2001), “Web User Studies: A Review and Framework for Future Work”, Journal of the American Society of Information Science and Technology, Vol 52 No 3, pp.235-46.

Jansen, B. J, Spink, A, and Koshman, S. (2007), “Web Searcher Interaction with the Dogpile.com Meta-Search Engine”, Journal of the American Society for Information Science and Technology, Vol 57 No 5, pp. 744-55.

Jansen, B. J, Spink, A, and Saracevic, T. (2000), “Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web”, Information Processing & Management, Vol 36 No 2, pp.207-27.

Li, M, Claypool, M, Kinicki, R, and Nichols, J. (2005), “Characteristics of streaming media stored on the Web”, ACM Transactions on Internet Technology, Vol 5 No 4, pp.601-26.

Park, S, Bae, H, and Lee, J. (2005), “End User Searching: A Web Log Analysis of NAVER, a Korean Web Search Engine”, Library & Information Science Research, Vol 27 No 2, pp.203-21.

Rieh, S. Y, and Xie, H. (2006), “Analysis of multiple query reformulations on the web: The interactive information retrieval context”, Information Processing and Management, Vol 42 No 3, pp.751-68.

Silverstein, C, Henzinger, M, Marais, H, and Moricz, M. (1999), “Analysis of a Very Large Web Search Engine Query Log”, SIGIR Forum, Vol 33 No 1, pp.6-12.

Spink, A, and Jansen, B. J. (2004), Web Search: Public Searching of the Web, Kluwer, New York.

Spink, A, Jansen, B. J, Blakely, C, and Koshman, S. (2006), “A Study of Results Overlap and Uniqueness among Major Web Search Engines”, Information Processing & Management, Vol 42 No 5, pp.1379-91.

Stojanovic, N. (2005), “On the conceptualization of the query refinement task”, Library Management, Vol 26 No 4/5, pp.231-93.

Wang, P, Berry, M, and Yang, Y. (2003), “Mining Longitudinal Web Queries: Trends and Patterns”, Journal of the American Society for Information Science and Technology, Vol 54 No 8, pp.743-58.

Zhang, M, Jansen, B. J, and Spink, A. (2006), “Information Searching Tactics of Web Searchers”, Proceedings of Sixty-ninth Annual Meeting of the American Society for Information Science & Technology, E-LIS, Austin, TX. Retrieved February 27, 2007, from http://eprints.rclis.org/archive/00008101/01/ZhangASIST2006.pdf