Analyzing Web Multimedia Query Reformulation Behavior

Liang-Chun Jack Tseng
Faculty of Science and Technology
Queensland University of Technology
Brisbane, QLD 4001, Australia
ntjack.au@hotmail.com

Dian Tjondronegoro
Faculty of Science and Technology
Queensland University of Technology
Brisbane, QLD 4001, Australia
dian@qut.edu.au

Amanda Spink
Faculty of Science and Technology
Queensland University of Technology
Brisbane, QLD 4001, Australia
ah.spink@qut.edu.au

Abstract *Current multimedia Web search engines still use keywords as the primary means to search. Due to the richness in multimedia contents, general users constantly experience some difficulties in formulating textual queries that are representative enough for their needs. As a result, query reformulation becomes part of an inevitable process in most multimedia searches. Previous Web query formulation studies did not investigate the modification sequences and thus can only report limited findings on the reformulation behavior. In this study, we propose an automatic approach to examine multimedia query reformulation using large-scale transaction logs. The key findings show that search term replacement is the most dominant type of modifications in visual searches but less important in audio searches. Image search users prefer the specified search strategy more than video and audio users. There is also a clear tendency to replace terms with synonyms or associated terms in visual queries. The analysis of the search strategies in different types of multimedia searching provides some insights into user's searching behavior, which can contribute to the design of future query formulation assistance for keyword-based Web multimedia retrieval systems.*

Keywords Web log analysis, multimedia search, query reformulation, search strategy

1. Introduction

The prevalence of multimedia information on the Web has changed user's information need from textual to multi-modal (i.e. audio, image, and video) searching. Multimedia search is more complex compared to general Web searches as evidenced by the longer session times and query lengths [11, 18]. Web multimedia search users also perform many query modifications, and have more difficulties in finding the appropriate terms to represent their needs. In addition, image search has the longest session length (i.e. more queries per session) [19] and more terms per query than video and audio searches [18]. Therefore, it is important to investigate user's multimedia query formulation behavior in order to better understand the characteristics and obstacles in different types of multimedia searches.

Existing studies have attempted to understand user's information searching behavior and the search trends from Web log analysis [7, 15, 19]. These studies have shown that users submit relatively short search queries, typically around three terms per query [15]. Most users do not review many results, typically only the first result page [9, 11]. Such little contextual information and brief interaction between the user and the search engine limited the understanding of user's searching behavior, especially when the analysis is based on individual transaction records [13, 16]. Thus, it is necessary to investigate multiple queries in order to provide more contextual information for Web log analysis.

The current study aims to discover multimedia query reformulation behavior and search strategies by applying novel log analysis procedures. To our knowledge, this is the first study to automatically analyze contextual information beyond two consecutive transaction logs. This approach also allows us to compare the search strategy characteristics among different types of multimedia searches and provide insights for future system development.

2. Related studies

2.1 Limitations of current Web log analysis

Web logs can be effectively used to understand general users' online searching behavior on a large scale [6, 8, 9, 11, 15] and are generally more objective and non-intrusive than other data collection methods [6]. Such unique characteristics make Web log data representative of user's unaltered behavior and thus regarded as the most convenient way to study real users [6]. However, the findings from individual transaction log are usually limited to the descriptive data without explanatory information about user's

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

searching behavior [16]. Recent studies have begun extracting contextual information from consecutive query modifications [5, 13, 14]. However, most studies are limited in the amount of queries that can be investigated because of the need for manual reviewing processes [12, 16, 20]. Other studies using large-scale data only examine searching behavior based on two consecutive query modifications. Thus, the full potential of contextual Web log analysis has yet to be discovered.

The analyses of contextual search information mainly focus on identifying new search sessions based on query modifications between consecutive queries [5, 10, 13]. He, Goker, and Harper [5] compared the effectiveness of using the time interval between two clicks and query modification patterns in detecting new search sessions. While the combination of these two methods produced the best results, query modification patterns accounted for the majority of the improvements. Ozmutlu and Cavdur [13] applied this method to Excite search engine logs. Their findings supported the usefulness of query modification patterns. Query modification pattern and time interval also have a significant effect on judging topic shifts [14]. In the comparison with several Support Vector Machine methods, the use of query modification pattern achieved at least 95% precision and recall in topic continuation, and 35% or more in topic shift cases, far better than its SVM counterparts. Similarly, Lau and Horvitz [12] used query modification pattern and time intervals between two consecutive queries to successfully predict user's upcoming search behavior based on a Bayesian probability model. Query modification has also been used for studying the uptake and effectiveness of terminology feedback provided by retrieval systems [1].

2.2 Web query reformulation behavior and search strategies

The term "modification" and "reformulation" have been used interchangeably in many Web log analysis studies without explicit clarification of the differences [10, 16]. In this study, we use "reformulation" to refer to user's overall behavior of formulating different versions of related queries in a session, whereas "modification" represents each change to the query. Thus query modifications can be classified into certain patterns and the overall query reformulation behavior implies user's search strategies.

Bruza and Dennis [4] investigated user's query reformulation behavior by manually classifying more than one thousand queries into one of the eleven types of query modifications. With the exception of the repeating queries, term substitution was found to be the most dominant type of query modifications, followed by term addition and deletion. A similar finding was also reported from the study on a meta-search engine Dogpile.com [10]. Jansen, Spink, and Narayan [10] investigated query reformulation behavior among large-scale Web log data. Despite the large proportion of formulating new search queries, query reformulation (which is equivalent to substitution in [4]) accounted for more than 15% of all eight types of modifications, with specialization (i.e. addition) occurring more than twice of generalization (i.e. deletion). They also concluded that major search content transitions were between Web and image collections.

Currently, only limited query modification studies have investigated more than two consecutive queries to infer user's search strategies [16] or tactics [2]. Rieh and Xie [16] manually investigated 313 sessions of five modifications or more to classify the overall query reformulation approach into one of the eight distinct strategies, including: *generalized*, *specified*, *dynamic*, *parallel*, *block-building*, *multi-tasking*, *recurrent*, and *format* (details in Section 4.6). Although they did not report the frequency for each strategy, they concluded that the first four (i.e. generalized, specified, dynamic, and parallel) are the most popular strategies. A similar categorization of search strategies can also be found in [2].

3. Research questions

Our focus is user's Web multimedia searching behavior which can be revealed by consecutive query modifications. We investigate the entire session of user's query modifications to infer the searching behavior. The three main questions that we attempt to answer are:

1. What are the frequent modifications in Web multimedia queries and do they differ among the multimedia searches?

2. What can the sequence of query modifications tell us about user's query reformulation behavior?

3. What search strategies can be inferred from query modification sequences? How can they contribute to the improvement of keyword-based Web multimedia retrieval systems?

4. Methodology

4.1 Dogpile log aggregation and query modification records

Dogpile is one of the leading online meta-search engines, which incorporates the indices of top search results from Google, Yahoo!, MSN Live, and Ask.com. For this study, a total of 1,228,310 records taken on May 15th, 2006 have been used in our analysis. The original Dogpile transaction log contains five fields that we use for our analysis:

IP: the IP address of the computer submitting the query.

Cookie: the unique identifier which Dogpile system sends to a particular computer with a pre-defined valid period.

Time: the time of the day when user submits the query.

Query: the original search text submitted to the system.

Vertical: the search type option which user selected on Dogpile's search page. In this study, we separate the logs with "images", "video", and "audio" option selected and replicate the analyses for comparison.

4.2 Browsing record aggregation

The Dogpile transaction logs are sorted based on different user identification (i.e. a unique combination of Internet Protocol (IP) and cookie) in a chronological order. Consecutive transaction logs with identical queries represent browsing records and are aggregated with the foremost record. If the last record has the same time stamp as the first record in current browsing aggregation, the duration will be logged as zero length (e.g. the last aggregation in Table 1 and Table 2). Table 1 and 2 illustrate the original log and aggregated record respectively.

IP	Cookie	Time	Query	Vertical
64.105.73.70	2187RDPA47YLJOB	6:05:33 PM	pod of dolphins	Images
64.105.73.70	2187RDPA47YLJOB	6:06:03 PM	group of dolphins	Images
64.105.73.70	2187RDPA47YLJOB	6:06:18 PM	group of dolphins	Images
64.105.73.70	2187RDPA47YLJOB	6:06:18 PM	group of dolphins	Images
64.105.73.70	2187RDPA47YLJOB	6:08:56 PM	dolphins	Images
64.105.73.70	2187RDPA47YLJOB	6:08:56 PM	dolphins	Images
64.105.73.70	2187RDPA47YLJOB	6:08:56 PM	dolphins	Images

Table 1. Original Dogpile search logs with browsing records

Session No.	Current query	Modified terms	Modification pattern	Duration
1	pod of dolphins		I	0:00:30
1	group of dolphins	pod,group	R	0:02:53
1	bottlenose dolphins	group of,bottlenose	R	0:00:00

Table 2. Query modification table with aggregated modification records

4.3 Modification pattern classification

Each aggregated transaction record is classified into a query modification pattern based on the content of the current query and the previous query. We use four modification patterns for our classification. The definitions for each modification pattern are:

Initial query (I): current query has no terms in common with the previous query

Addition modification (A): current query contains all search terms from the previous query, as well as some new terms

Deletion modification (D): current query omits some terms from the previous query

Replacement modification (R): deletion and addition of terms happen simultaneously to form the current query

Thus an initial query represents a new search topic since no search terms are carried over from previous query. Some studies also classify replacement modification as "reformulation" [5, 13, 14]. However, as users can freely reformulate the query by changing the order of search terms without affecting the search results, we use the term "replacement" to clearly indicate such modification. Details of the classification algorithm can be found in [5]. We built a program to automatically classify queries by their modification patterns and aggregate consecutive browsing records in Table 2.

4.4 Session aggregation

A search session is a series of related queries submitted by same user. In addition to being defined by a unique combination of IP and cookies, a query with no terms in common with its preceding query is regarded as the beginning of a new session, thus classified as the "initial query". By calculating the number of sessions with same IP and cookie combination, we are able to identify the average search topics submitted by a user.

4.5 Modification sequence

Once the query modification records have been generated, consecutive modifications within each session can be classified into several predetermined modification sequences. We used our program to identify the occurrence of thirty-six types of modification sequences, incorporating two or three predetermined modifications. Sessions with less than two modifications are discarded as they provide little information about user's behavior. The two–modification-sequences comprise one initial query (I), followed by two query modifications which can be either of the replacement, addition, or deletion modification. Thus, nine patterns (3*3) can be formulated for the two-modification sequences. Similarly, twenty-seven patterns of three-modification-sequences (3*3*3) can be formulated. The main purpose of this analysis is to discover the frequent patterns of modification sequences that users follow, thus revealing user's preference for consecutive query modifications and providing in-depth information for search strategy analysis.

4.6 Search strategies based on modification sequence analysis

When typical modification sequences emerge from our analysis, we calculate the changes in the number

of query terms within each sequence. Such changes can determine if users adopt some particular search strategies. We construct our strategy classification based on the higher level categorization in [16]. The list of modification strategies used in this study, as well as the detail descriptions of our analysis assumptions are as follows:

Generalized reformulation

A user may begin with several search terms and subsequently drop some of the terms to include more results. This generalized reformulation is often manifested by consecutive term deletion changes [16]. It can also be characterized by replacing the query with fewer terms. Modification sequences in which subsequent queries always have fewer or equal terms to the precedent queries belongs to this category.

Specified reformulation

When a user persistently specifies a query by adding more terms or changing to more specific phrases, we classify this approach as specified reformulation. In our analysis, modification sequences in which a subsequent query always has more or equal terms to its preceding query belongs this category.

Dynamic reformulation

When a user inconsistently switches between generalized and specified reformulation, we characterize such approach as dynamic reformulation. Such modification pattern manifests the unplanned nature of user's search process. Users who adopt this search strategy generally have the most unconsolidated search problems, and require more interaction with the retrieval system. Modification sequences in which subsequent queries can have either fewer or more terms than precedent queries exhibit dynamic search strategy.

Constant reformulation

Constant search occurs when a user modifies terms of the same concept level which shares some common characteristics, for example when substituting with related objects (e.g. from PC to Mac) or synonyms. This strategy is characterized by having a constant number of query terms across the entire modification sequence, regardless of the existence of replacement modifications. The same query specificity suggests a one-to-one relationship between the original and new terms. We used the term "constant reformulation" to reflect this unique characteristic.

5. Results

5.1 Query modification

From Table 3, image searches are the dominant type of multimedia search in our dataset with more than 50% of sessions and users attributed to image searches. Audio is the second popular type of multimedia search whereas video is the least popular. As Table 4 shows, initial queries are the majority of query modification across all multimedia searches. Replacement modification is more than twice of the addition modification in visual searches (i.e. image and video searches) but much less in audio searches. Deletion is the least type of modification in all searches.

Comparing the distribution of the four modifications in multimedia searches, audio search users are more likely to formulate new search topics as they have larger proportion of initial queries and more topics per user than image and video searches. The number of topics submitted by both image and audio users varies a lot (SD=21.60 and 22.39 respectively) while video users shows a much uniformity pattern (SD=8.33). For the number of modifications, image and video users have the same amount of modifications (1.71 modifications on average) while audio users show slightly fewer modifications per session (1.63 on average). Overall, image and video search users are very similar in terms of query modifications.

	Image	%	Video	%	Audio	%	Total
Log records	597,760	48.7	231,941	18.9	398,609	32.5	1,228,310
Sessions	183,825	52.9	52,405	15.1	110,945	32.0	347,175
Users	60,701	52.1	21,677	18.6	34,088	29.3	116,466

Table 3. Statistics of image, video, and audio search logs in Dogpile dataset

	Image	%	Video	%	Audio	%
Initial	183,825	58.6	52,405	58.5	110,945	61.2
Replacement	82,292	26.2	22,225	24.8	33,645	18.6
Addition	30,716	9.8	8,817	9.8	22,553	12.4
Deletion	16,757	5.3	6,124	6.8	14,176	7.8
Total	313,590	100.0	89,571	100.0	181,319	100.0
Topics per user						
Average	3.03		2.42		3.25	
SD	21.60		8.33		22.39	
Modifications per session						
Average	1.71		1.71		1.63	
SD	1.68		1.68		1.42	

Table 4. Statistics of query modification records

5.2 Modification sequence

Two-modification-sequence analysis

The frequencies of each modification sequence pattern (in percentages) are presented in Table 5 and 6. Table 5 signifies the popularity of replacement modification in all types of multimedia searches, as evidenced by the dominance of *I-R-R* and *I-A-R* sequences. On the contrary, the unlikelihood of consecutive deletion modification is manifested by the low occurrence of *I-D-D* sequences (i.e. less or equal to 1% in all multimedia searches).

Figure 1 shows that both image and video searches have prominently more *I-R-R* sequences than audio searches. The audio searches have much more *I-A-D* sequences than the other two types of searches, thus

making it more evenly distributed in the top three modification sequence patterns. All multimedia searches show a similar distribution beyond the top three patterns.

	Image	Video	Audio
I-R-R	41.2%	39.2%	26.8%
I-A-R	24.9%	22.2%	23.6%
I-A-D	10.9%	13.4%	21.9%
I-R-A	5.6%	5.3%	6.2%
I-R-D	5.4%	6.8%	6.9%
I-D-A	5.1%	6.2%	7.1%
I-A-A	3.9%	3.3%	4.3%
I-D-A	2.3%	2.6%	2.1%
I-D-D	0.6%	1.0%	1.0%

Table 5. Comparison of the frequencies in two-modification-sequence patterns

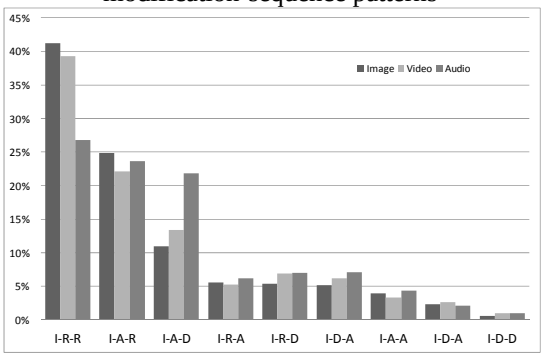

Figure 1. The distribution of the two-modification-sequence patterns among Image, Video, and Audio searches

Three-modification-sequence analysis

The dominance of replacement and addition modification continued in the analysis of three–modification-sequences. As shown in the high frequencies of both *I-R-R-R* and *I-A-R-R* sequences in Table 6, about 50% of all three modification sequences in image and video searches are associated with replacement and addition modifications. Similarly to the distribution in two-modification-sequence analysis, both image and video searches have much higher proportion of consecutive replacement modifications (i.e. *I-R-R-R* sequences) than audio searches. The top three modification sequence patterns distribute more evenly in audio searches with a slightly more *I-A-D-A* sequences than the other two types of searches. For modification sequence patterns beyond the top five, all multimedia searches demonstrate similar distribution, thus provide little information for characterizing different types of multimedia searches. Figure 2 shows that *I-R-R-R* sequences are more prominent in both image and video searches as the distribution decreased more in the top three modification sequence patterns than audio searches. The top five patterns account for over half of the three-modification–sequence in all multimedia searches and only one pattern contains the deletion modification. When we further differentiate *I-R-R* sequence into *I-R-R-R*, *I-R-R-A*, and *I-R-R-D* sequences, the prevalence of replacement over addition and addition over deletion continued (*I-R-R-D* not shown in Table 5). Hence, user's preference for replacing terms and the unlikelihood of deletion modification in the early stage of query modification can be confirmed.

	Image	Video	Audio
I-R-R-R	35.3%	32.4%	21.6%
I-A-R-R	20.1%	17.3%	16.7%
I-A-D-A	5.3%	6.3%	11.1%
I-R-A-R	4.5%	4.1%	3.6%
I-R-R-A	3.9%	3.5%	2.6%
Total	69.1%	63.7%	55.6%

Table 6. Comparison of the top 5 frequencies in three-modification-sequence patterns

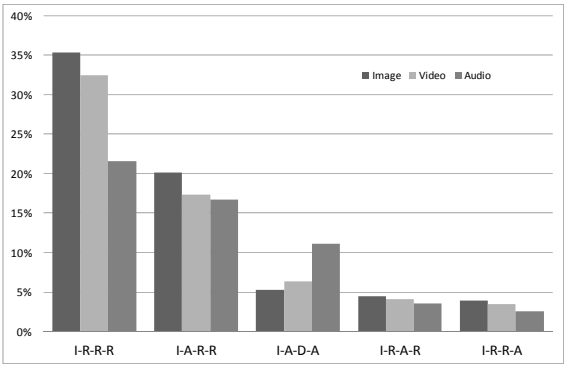

Figure 2. The distribution of the top 5 three-modification-sequence patterns among Image, Video, and Audio searches

5.3 Search strategies based on modification sequence

We investigated search strategies based on the consecutive replacement sequences (i.e. the *I-R-R* and *I-R-R-R* sequences) because of their prominence in the modification sequence analysis. As Table 7 shows, about 40% of all *I-R-R* sequences exhibit a dynamic search strategy. From Table 8, the proportion of dynamic search increases to more than 50% in *I-R-R-R* sequences. While this large proportion of dynamic search can be anticipated, constant search which accounts for nearly one-third of all consecutive replacement sequences is more revealing. Because the query length is held at constant within each session in constant searches, it appears to be a one-to-one relationship between the replaced term pairs. A reasonable explanation is the interchange of synonyms or associated terms of the same construct (e.g. "PC" to "Mac", "UK" to "USA", or "girls" to "boys"). Both Table 7 and 8 show more specified searches than generalized searches, but the difference is only

noticeable in image searches. This indicates that image users are more prone to adopt specified strategy (i.e. gradually adding more search terms as the searching progresses) than other types of multimedia users. In other words, image users progressively consolidate or learn more information about their problems through the interaction with the Web search engine. The percentage for the search strategy analysis from *I-R-R* and *I-R-R-R* sequences are presented in Table 7 and Table 8 respectively.

	Image	**Video**	**Audio**
Dynamic	40.1%	40.6%	42.7%
Constant	34.8%	34.3%	28.5%
Specified	15.2%	12.8%	15.2%
Generalized	9.9%	12.3%	13.5%

Table 7. The percentage of each search strategy from *I-R-R* modification sequences

	Image	**Video**	**Audio**
Dynamic	52.9%	52.9%	58.3%
Constant	27.2%	25.7%	22.1%
Specified	11.9%	11.3%	10.1%
Generalized	8.0%	10.1%	9.4%

Table 8. The percentage of each search strategy from *I-R-R-R* modification sequences

As shown in Figure 3 and 4, both strategy analyses from *I-R-R* and *I-R-R-R* sequences suggested the highest constant search strategy in image searches. Thus image searches require most synonym or related term replacement modification than other types of multimedia searches, and such characteristic should benefit image searches more from term suggestion functionalities when refining the search queries. While video searches have slightly less proportions of constant searches than in image searches, they shared very similar distribution across the four types of search strategies. On the other hand, audio search users are more prone to adopt a dynamic search strategy.

In order to verify the replaced terms in constant search sequences, we implemented a Brill tagger[1] [3] to identify the part-of-speech of the replaced term pairs (i.e. the terms from the original query paired with the terms from the replacement query). Among the randomly selected 1465 constant *I-R-R-R* modification sequences, a total of 3003 replaced term pairs have been successfully tagged using the Brill tagger. More than 70% of these term pairs (2125 in total) have same part-of-speech, reassuring our explanation of interchanging between synonyms or associated terms of same construct in these constant search sequences.

[1] Details on the tagger implementation can be found in [17].

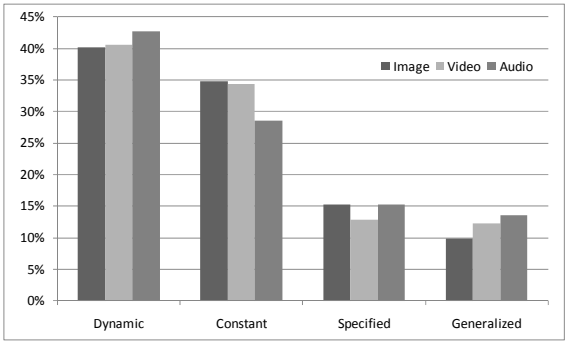

Figure 3. Comparison of the search strategies from *I-R-R* modification sequences

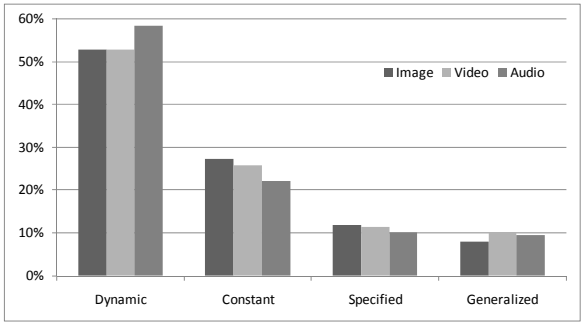

Figure 4. Comparison of the search strategies from *I-R-R-R* modification sequences

6. Discussion and future work

The statistics of query modification revealed that all multimedia search users shift their search topics more than refining their queries. Such phenomenon is most evident in audio searches as initial queries are more than triple of the replacement queries. The replacement queries are more than twice the addition queries in both image and video searches, whereas audio searches have notably more addition queries. Deletion queries are the least type of modifications in all multimedia searches, especially in image searches. Overall, when users do modify their queries, they tend to replace their search terms rather than adding or removing them. Such modification tendency is more prominent for visual searches (i.e. image and video searches). Although the number of topics searched by one user varies a lot, users searched around two to three different topics on average. In terms of in-session modification analysis, the majority of users only perform little modifications to their queries and visual search users modify their queries slightly more than audio users.

The analysis of modification sequence pattern suggests the tendency to replace and add search terms when modifying visual queries. The distribution of modification sequences shows a tendency toward the consecutive replacement modification sequences (i.e. the *I-R-R* and *I-R-R-R* sequences) in visual searches. This tendency also distinguishes visual searches from audio search and suggests the need for interchanging related search terms. In other words, visual search

users are more willing to interact with the system than audio search users.

In terms of search strategies, the changes in number of terms within *I-R-R* and *I-R-R-R* sequences reveal that about 40%-50% of users engage in dynamic searches. This typically reflects the unplanned nature of Web multimedia searching, which manifests the need for initiating several guessing runs to consolidate user's problem, or to find the appropriate search terms. Nevertheless, about one third of users adopt constant search strategy in which they replace search terms with an equal number of terms, suggesting the high likelihood of interchanging with synonyms or related terms. This constant search strategy also differentiates visual searches from audio searches. While visual searches always have higher proportion of constant searches, image users adopt most constant search strategy among all multimedia searches. Hence it can be assumed that image users should benefit most from knowledge or ontology based query expansion or term suggestion assistance. The reason of less constant search strategy in audio searches may be that audio searchers tend to use the song title or singer's names in their queries [19], resulting the replacement of these proper nouns other than interchanging similar terms in visual searches.

When the change of terms shows a unidirectional pattern, all multimedia searchers are more prone to adopt the specified approach. This finding is consistent with prior study's conclusion on user's primary concern of retrieval precisions [9]. In particular, image search users show a stronger preference for adopting this approach than other types of multimedia users. Typical scenario would be that image search users need to see widely before they know exactly what they are searching for or how their target images should look like. This characteristic implies the importance of a browsing tool that helps users compare different results and thus consolidate their problems quicker. A hierarchy arrangement of the results or term suggestions should also be useful.

Compared with general Web search studies, the current study findings are consistent with Jansen and Spink's [8] conclusion on the complexity of user's Web search behavior as one-query session increased over the years and users modify their queries less and less. This is to say that general Web users share the same characteristics with our user pool. Hence the effectiveness of the interaction between the user and the system is substantial to the improvement of query modification process. Future work should include a user study to understand the reasons behind each modification, as well as the corresponding search strategies. A semantic level analysis of replaced terms would also help discover the aspects of multimedia content that users modify most, such as the visual descriptors or the semantic meanings of the retrieved objects.

The prevalence of consecutive replacement modifications implies the need for an effective relevance feedback mechanism that would help users refine the importance of their query terms, perhaps with advanced search term suggestions based on the replaced terms (e.g. automatically displays synonyms or associated terms when user deletes a term). In terms of search strategies, the current study confirms the preference for the specified approach among image searchers. An interactive retrieval system that can gradually obtain more information about user's image problem would be helpful in guiding the user to explore the entire collection, and hence improve the query reformulation effectiveness.

7. Limitations

Due to the aim of using automatic approach to discover user's query modification behavior, this study only perform the part-of-speech analysis of replaced terms in constant search sequences. This limits our understanding of the types of terms being modified during the reformulation process. However, user's overall search strategy can still be inferred from our analysis. Although we have successfully discovered some unique characteristics among different types of multimedia searches, these findings are yet to be compared with general Web searches to address the differences in terms of query modification behavior and search strategies.

8. Conclusion

The current study investigated users' multimedia searching behavior based on their query modification methods. Our analysis showed that around 60% of query modifications are to formulate new search topics. Image and audio users searched more topics on average than video users. Our approach to analyze Web multimedia query modifications went beyond two consecutive queries. The analysis of session modifications revealed that visual search users (i.e. both image and video users) modify their queries slightly more than audio users. Visual search users also tend to replace search terms with other related terms rather than merely narrowing or broadening their searches. Generally speaking, visual searches showed similar modification patterns with much more consecutive replacement modifications than in audio searches. In terms of search strategies, the relatively high proportion of constant search strategy in visual searches indicates the importance of term suggestion assistance that helps user find the synonyms or related terms more easily. Our search strategy analysis also showed the tendency of adopting a specified approach in image searches, which suggests a need for query formulation assistance to help users gradually specify of their problems.

We present an automatic analysis procedure in this study, thus maximizing the ability to apply the same

analysis to different data sets, as well as allowing comparisons with general Web user's searching behavior. By adopting the analysis procedure, it is possible to extract more information about user's query modification behavior, especially the search strategies based on the statistical evidence. Future multimedia retrieval systems can utilize these different search characteristics to improve query formulation process and search efficiency.

9. References

[1] P. Anick. Using terminological feedback for web search refinement: a log-based study. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 88-95, 2003.

[2] M. J. Bates. Information search tactics. *Journal of the American Society for Information Science*, Volume 30, pages 205-214, 1979.

[3] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, pages 112-116, 1992.

[4] P. D. Bruza and S. Dennis. Query Reformulation on the Internet: Empirical Data and the Hyperindex Search Engine. In *Proceedings of the RIAO 97 Conference*, pages 488-499, 1997.

[5] D. He, A. Goker, and D. J. Harper. Combining evidence for automatic Web session identification. *Information Processing & Management*, Volume 38, pages 727-742, 2002.

[6] B. J. Jansen. Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, Volume 28, pages 407-432, 2006.

[7] B. J. Jansen, A. Goodrum, and A. Spink. Searching for multimedia: analysis of audio, video and image Web queries. *World Wide Web*, Volume 3, pages 249-254, 2000.

[8] B. J. Jansen and A. Spink. An analysis of Web searching by European AlltheWeb. com users. *Information Processing and Management*, Volume 41, pages 361-381, 2005.

[9] B. J. Jansen and A. Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, Volume 42, pages 248-263, 2006.

[10] B. J. Jansen, A. Spink, and B. Narayan. Query Modifications Patterns During Web Searching. In *Fourth International Conference on Information Technology (ITNG'07)*, pages 439-444, 2007.

[11] B. J. Jansen, A. Spink, and J. Pedersen. An analysis of multimedia searching on AltaVista. In *5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 186-192, 2003.

[12] T. Lau and E. Horvitz. Patterns of Search: Analyzing and Modeling Web Query Refinement. In *Proceedings of the 7th international conference on user modeling*, pages 119-128, 1999.

[13] H. C. Ozmutlu and F. Cavdur. Application of automatic topic identification on Excite Web search engine data logs. *Information Processing & Management*, Volume 41, pages 1243-1262, 2005.

[14] S. Ozmutlu. Automatic new topic identification using multiple linear regression. *Information Processing & Management*, Volume 42, pages 934-950, 2006.

[15] S. Ozmutlu, A. Spink, and H. C. Ozmutlu. Multimedia web searching trends: 1997-2001. *Information Processing & Management*, Volume 39, pages 611-621, 2003.

[16] S. Y. Rieh and H. Xie. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, Volume 42, pages 751-768, 2006.

[17] T. Simpson and T. Dao. WordNet-based semantic similarity measurement. *The Code Project.com*, Oct. 1, 2005. [Online]. Available: http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx. [Accessed: Jun. 10, 2009].

[18] A. Spink and B. J. Jansen. Searching multimedia federated content web collections. *Online Information Review*, Volume 30, pages 485-495, 2006.

[19] D. Tjondronegoro, A. Spink, and B. J. Jansen. A study and comparison of multimedia Web searching: 1997-2006. *Journal of the American Society for Information Science and Technology*, Volume 60, pages 1756-1768, 2009.

[20] M. Zhang, B. J. Jansen, and A. Spink. Information Searching Tactics of Web Searchers. In *69th Annual Meeting of the American Society for Information Science and Technology*, Austin, USA, 2006.