Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management
Istanbul, Turkey, July 3 – 6, 2012

Automatic New Topic Identification in Search Engine Transaction Log Using Goal Programming

Fatih Çavdur, Seda Özmutlu, H. Cenk Özmutlu, Duygu Yilmaz Eroğlu and Burcu Çağlar Gençosman
Industrial Engineering Department
Uludag University, Gorukle, Bursa 16059, Turkey

Abstract

Automatic new topic identification is a key issue in transaction log analysis of Web search engines. This study proposes a goal programming model to automatically perform new topic identification in search engine transaction logs by using statistical characteristics of queries, such as time intervals and query reformulation patterns. Sample data log from the Excite search engine is selected, and then the goal program is used to identify topic changes in the data log. Our results show that, the goal programming approach overestimates the topic changes in the sample dataset compared to some other approaches of automatic new topic identification; however, doing so also significantly improves its performance in terms of the number of correctly estimated topic changes. Some performance measures yielded more satisfactory results than the ones in previous studies, showing that the goal programming approach could be a successful tool for automatic new topic identification.

Keywords
Information retrieval, search engine user behavior analysis, automatic new topic identification, query clustering, goal programming

1. Introduction
Search engine transaction log analysis has been an important branch of information science for the last decade and a half. Previous studies [14, 19] have shown that users might be interested in multiple topics during a search session. With focus on the estimation of the topic changes in a search session, *new topic identification* is a sub-area of transaction log analysis, and an essential procedure in investigating user patterns of search engine transaction logs.

In this study, a goal programming model to automatically identify topic changes in search engine transaction logs is proposed. This is a mathematical model that aims to obtain best results with respect to multiple objectives under some constraints. Artificial intelligence based and some statistical methods have been frequently used for transaction log analysis and automatic new topic identification previously; however, mathematical programming approach has not yet been considered for similar studies. Therefore, the research objective of this study is to investigate the performance of goal programming approach as an automatic new topic identification tool and also to introduce the use of mathematical programming based approaches to the information science literature. To establish these objectives, the transaction log queries are first categorized with respect to their statistical characteristics. Then, automatic new topic identification is performed using a goal programming formulation. A real search engine transaction log is used for the study.

The organization of the paper is as follows: We initially present the literature review related to automatic new topic identification, followed by the description of the methodology. The results obtained are given in the following section, and we finally present the conclusion.

2. Related Research
A session is a sequence of queries issued by a user (or an application) to achieve a certain searching task [3]. Consequently, session identification is defining the boundaries of queries submitted by a single user. Time-out methods have been the most popular method for session identification [3]. In time-out methods, a session is accepted to end after a predefined time threshold is exceeded. Search engine researchers have also used cookies and IP

addresses, besides the time threshold, for user and session identification. In case of the use of common-access computers centers such as libraries and computer labs, a single IP address might not mean a single user, due to dynamic IP applications and use of search engines at these common places [4].

Query clustering is also relevant to the new topic identification problem, since it is the segmentation of queries into several topic clusters. It is applied to use search engine query logs including click-through data [1, 20]; an automatic classification methodology is developed to classify search queries into broad subject categories [17], and Support Vector Machines (SVMs) are used to classify queries in different groups [5,6]. More detailed literature review on query clustering and text categorization can be seen in [9].

Most query clustering methods are focused on interpretation of keywords or understanding the topic or the contents of the query, which significantly complicates the process of query clustering and increases the potential noise of the results of the study. Defining topic boundaries by relying on semantics of query terms are "dangerously circular" and conceal persistence of users' long-term information needs [7]. Another obstacle about topic identification or context extraction of search engine transaction logs is that there are few keywords in each query; some of which could be meaningless or subject-specific words, such as model numbers of electronic devices [15]. It is difficult to extract context from only few keywords. On the other hand, the indications of the relation between statistical characteristics of queries and topic change were shown in [2,8,19].

Hence, using statistical characteristics of queries instead of their semantic properties is an alternative approach to use for new topic identification. In such an approach, queries can be categorized in different topic groups with respect to their statistical characteristics, such as the time intervals between subsequent queries or the reformulation of queries. To adapt such an approach our research team applied Dempster-Shafer Theory [18, 10, 12], neural networks [13, 11] and conditional probabilities [9]. In this study, we intend to improve the performance of automatic new topic identification using an alternative methodology; goal programming.

Goal programming is a multi-criteria optimization method, and it aims to obtain the best solution by optimizing two or more objectives that might conflict with each other. The new topic identification problem can be formulated with conflicting objectives of precision (P) and recall (R), and solved to obtain the best possible values of these two conflicting objectives. Hence, to the best of our knowledge, this paper is the first study that introduces the usage of mathematical programming (more specifically, goal programming) for automatic new topic identification.

3. Methodology

3.1. Research design

We use a dataset from Excite search engine's transaction log. The Excite search engine (http://www.excite.com) provided a query log of 1.7 million for our analysis. All queries were submitted on May 4, 2001 to the Excite search engine. In the Excite data log structure, the entries are given in the order of arrival. Excite assigns a new user ID to every new user because it is possible to identify them through the IDs. In addition, each query is given a time stamps in hours, minutes and seconds. A sample of 10,256 queries was selected using Poisson sampling.

The number of queries in the sample size was kept at that level since manual processing of the queries was also necessary for the performance analysis of our approach. The sample was selected using Poisson sampling [16] to provide a sample dataset that is both statistically representative of the entire data set and small enough to be conveniently analyzed. Poisson sampling provides a basis for sampling from large-scale data logs, such as the Excite search query log, while preserving the characteristics of the main dataset.

3.2. Notation

The following notation was used in this study:

Topic shift: A change from one topic to another between queries within a single user session.

Topic continuation: Staying on the same topic from one query to another within a single user session.

N_{shift} : Number of queries labeled as topic shifts by our automatic new topic identification approach using goal programming.

N_{contin} : Number of queries labeled as topic continuation by our automatic new topic identification approach using goal programming.
$N_{trueshift}$: Number of queries labeled as topic shifts by manual examination of the human expert.
$N_{truecontin}$: Number of queries labeled as topic continuation by manual examination of the human expert.
$N_{shift\& \ correct}$: Number of queries labeled as topic shifts by our automatic new topic identification approach using goal programming and by manual examination of the human expert.
$N_{contin\&correct}$: Number of queries labeled as topic continuation by our automatic new topic identification approach using goal programming and by manual examination of human expert.
Type A error: This error occurs when queries on same topics are considered as separate topic groups.
Type B error: This error occurs when queries on different topics are grouped into a single topic group.

We use the performance measures *precision* (*P*) and *recall* (*R*) which are commonly used in automatic new topic identification studies. The focus of precision and recall are both on correctly estimating the number of topic shifts and continuations. Interpreted in terms of topic shifts, as in Ozmutlu and Cavdur (2005b) and Ozmutlu *et al*. (2006), precision (P_{shift}) is the correctly estimated number of shifts by the automatic new topic identification approach among all the shifts marked by the automatic new topic identification approach (Eq.1). Recall (R_{shift}) is the correctly estimated number of shifts by the automatic new topic identification approach among all the shifts marked by the human expert (Eq.2). The third measure, F_{β_shift} combines P_{shift} and R_{shift} to provide a single parameter to compare different results as used in previous studies. As shown in (Eq.3) β is chosen as 1.3 in this study to be consistent with the previous studies on automatic new topic identification. The formulations of these performance measures used in demonstrating the performance of the proposed automatic new topic identification approach are as follows:

$$P_{shift} = \frac{N_{shift\ \&\ correct}}{N_{shift}} \tag{1}$$

$$R_{shift} = \frac{N_{shift\ \&\ correct}}{N_{true\ shift}} \tag{2}$$

$$F_{\beta\_shift} = \frac{(1+\beta^2)P_{shift}R_{shift}}{\beta^2 P_{shift} + R_{shift}} \tag{3}$$

3.3. Proposed Methodology

In this study, we propose a goal programming formulation to be used in identifying topic changes in the Excite search engine query log. The general steps of our approach applied in this paper are explained next.

3.3.1. Preprocessing of the transaction logs

The transaction logs are preprocessed before using the automatic new topic identification approach. The preprocessing steps of the methodology are similar to those used in [9-12].

Evaluation by human expert: The actual topic shifts and continuations in the approximately 10,000-query in Excite dataset is identified and marked by a team of human experts. This step is necessary for the performance analysis of our approach, i.e. how successful the proposed approach in determining topic shifts and continuations.

Separating the data into two sets: The dataset is divided into two approximately equal sized sections. The first section of the dataset is used to determine the label (topic shift or continuation) of each query category using the goal programming model and the second section is used to test the performance of the approach. The two data sections do not contain the same number of queries to keep the entirety of the user session containing the median query. The size of the dataset is given in Table 1 [9-12]:

Table 1: Size of the dataset used in the study

Search engine	Excite
Entire dataset	1,7 million
Sample set	10,256
1st half of the sample set	5128 queries
2nd half of the sample set	5128 queries

Identification of search pattern and time interval of each query in the dataset: Each query in the dataset is categorized in terms of its search pattern and time interval. The classification of the search patterns is based on terms of the consecutive queries within a user session. The time interval is the difference of the arrival times of two consecutive queries. The categories of time intervals are determined with respect to the length of the difference of the arrival times of two consecutive queries. The categorization of time interval and search pattern remains similar to those of [9-12]. We use seven categories of search patterns in this study, which are as follows:

- Unique (New): the second and first queries have no common terms.
- Next Page (Browsing): the second query requests another set of results on the first query.
- Generalization: all of the terms of second query are also included in the first query but the first query has some additional terms.
- Specialization: all of the terms of the first query are also included in the second query but the second query has some additional terms.
- Reformulation: some of the terms of the second query are also included in the first query but the first query has some other terms that are not included in the second query. This means that the user has added and deleted some terms of the first query. Also if the user enters the same terms of the first query in a different order, it is also considered as a reformulation.
- Relevance feedback: the second query has zero terms (empty) and it is generated by the system when the user selects the choice of ‘‘related pages ’’.
- Others: If the second query does not fit any of the above categories, it is labeled as other. Any non-empty query listed after an initial empty query, such as relevance feedback belongs to the ‘others’ category. Note that if this pattern observed on an intermediate query in a user session, this property does not hold.

The search patterns are automatically identified by a computer program. The logic for the automatic search pattern identification and the distribution of the queries with respect to search pattern can be found in previous studies [9-12].

The categories of time intervals are determined with respect to the length of the difference of the arrival times of two consecutive queries and are similar to those used in previous studies [9-12]. We use seven categories of time intervals for a query: 0-5 minutes, 5-10 minutes, 10-15 minutes, 15-20 minutes, 20-25 minutes, 25-30 minutes and 30+ minutes as seen in these studies [9-12].

3.3.2. Session identification using goal programming

The goal programming formulation: As explained in the previous subsection, there are two statistical characteristics of the transaction log queries; time interval and search pattern; each with seven levels. We propose a combination of these two statistical characteristics, which yields 49 categories of queries, and any query in the transaction log falls into one of these categories. For example: Time Interval 1 & Search Pattern 1; Time Interval 1 & Search Pattern 2; …; Time Interval 7 & Search Pattern 7. The objective is *to decide whether a query that belongs to any one of these categories should be labeled as a topic shift or a continuation*. Consequently, there are 49 binary decision variables. Each binary decision variable x_{ij} is given in Eq.4. Since the decision variable x_{ij} can be assigned values independently of each other, the mathematical formulation has no constraints except the ones with binary variable restrictions.

$$x_{i,j} = \begin{cases} 1 & \text{if the query of category time interval i, search pattern j. are assigned as topic shift} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1\ldots 7,\ j = 1\ldots 7 \tag{4}$$

Previous studies [9-12] have shown that identifying shifts is more problematic compared to identifying continuations. Therefore, we chose to use $F_{ß\_shift}$ as the objective function. Consequently, the objective of the methodology is to maximize $F_{\beta\text{-shift}}$ (Eq. 5).

$$F_{\beta\_shift} = \frac{(1+\beta^2) P_{shift} R_{shift}}{\beta^2 P_{shift} + R_{shift}} \tag{5}$$

The upper bound for $F_{ß\_shift}$ is 1, which can only be attained when both P_{shift} and R_{shift} are 1. On the other hand, increasing P_{shift} and R_{shift} are conflicting objectives. In order to increase R_{shift}, $N_{shift\&correct}$ should be increased. The algorithms used in previous studies [9-12] increased $N_{shift\&correct}$ by increasing N_{shiftt}. Unfortunately N_{shift} increases faster than $N_{shift\&correct}$, thus increasing *Type A* errors even faster, resulting a decrease in P_{shift}. Therefore, there are two separate objective functions in the mathematical formulation, which requires multi-criteria decision making. Hence, the two conflicting objective functions are as follows:

$$\begin{aligned} &\text{Max } G_1 = R_{shift} \\ &\text{Max } G_2 = P_{shift} \end{aligned} \tag{6}$$

The ß value in the $F_{ß\_shift}$ formulation is 1.3. This value gives more weight to the optimization of the R_{shift} value. Recalling that , R_{shift} (Eq.7) is necessary to maximize $N_{shift\&correct}$, since $N_{trueshift}$ is a given value for each dataset.

$$R_{shift} = \frac{N_{shift\&correct}}{N_{true\ shift}} \tag{7}$$

Consequently, G_1 becomes:

$$\text{Max } G_1 = N_{shift\&correct} \tag{8}$$

The second objective formulation is given in Eq.9.

$$P_{shift} = \frac{N_{shift\&correct}}{N_{shift}} = \frac{N_{shift\&correct}}{N_{shift\&correct} + N_{shift\&incorrect}} \tag{9}$$

Recall P_{shift} we see that $N_{shift\&correct}$ is both in the numerator and the denominator of the P_{shift} equation. It is necessary to minimize $N_{shift\&incorrect}$ to maximize P_{shift}. Consequently, G_2 becomes:

$$\text{Min } G_2 = N_{shift\&incorrect} \text{ or Max } G_2 = -N_{shift\&incorrect} \tag{10}$$

The final goal programming formulation is as follows:

Max $G_1 = N_{shift\&correct}$

Max $G_2 = -N_{shift\&incorrect}$

s.t.:

$x_{ij} = 0$ or 1 ($i = 1, \ldots, 7$, $j = 1, \ldots, 7$)

Using the non-preemptive goal programming approach, the objective function can be written as:

Max z = $N_{shift\&correct} - \alpha N_{shift\&incorrect}$

s.t.:

$$x_{ij} = 0 \text{ or } 1 \ (i = 1, \ldots, 7, j = 1, \ldots, 7) \tag{11}$$

Hence, by using the non-preemptive goal programming approach, the formulation has been reduced to a binary integer mathematical programming model. The objective of the model is to maximize the number of correctly assigned shifts. An important point here is to assign the correct α value which weighs the second objective showing the relative importance of it compared to the first one. We performed a set of preliminary studies to determine the correct α value to maximize $F_{ß\_shift}$, when ß equals to 1.3. Table 2 shows the values of P_{shift}, R_{shift} and $F_{ß\_shift}$ for different values of α for the Excite dataset. As seen in Table 2, the highest $F_{\beta\_sh\S ft}$ value has been attained when α equals to 0.25, 0.30 or 0.35. Therefore, we chose the median of these values ($\alpha = 0.30$).

Table 2: P_{shift}, R_{shift} and $F_{ß\_shift}$ for different values of α

α	P_{shift}	R_{shift}	$F_{ß\_shift}$
0.05	0.371957	0.976982	0.608832
0.10	0.371957	0.976982	0.608832
0.15	0.371957	0.976982	0.608832
0.20	0.392593	0.948849	0.621495
0.25	0.394036	0.946292	0.622144
0.30	0.394036	0.946292	0.622144
0.35	0.394036	0.946292	0.622144
0.40	0.540359	0.616368	0.585739
0.45	0.540359	0.616368	0.585739
0.50	0.540359	0.616368	0.585739
0.55	0.591036	0.539642	0.557669
0.60	0.591036	0.539642	0.557669
0.75	0.604167	0.519182	0.547829
1.00	0.604167	0.519182	0.547829
1.25	0.630081	0.396419	0.459809
1.50	0.630081	0.396419	0.459809

The goal programming model is solved using the first half of the dataset where each one of the 49 query categories is labeled as either topic shifts or continuations to maximize the combined objective of two conflicting criteria. We used MS Excel to solve the goal programming formulation. Then, using the labels assigned in the training dataset, the queries in the test dataset are labeled as topic shifts and continuations, i.e., if a query category in the first half of a dataset is marked as a topic shift, then, it is also marked as a topic shift in the second half or vice versa.

Comparison of the results from the human expert and automatic new topic identification approach: The results of the automatic new topic identification approach based on the goal programming formulation are compared to the actual topic identifications (topic continuations and shifts) determined by the human expert. Human judgment is the golden standard in information science and topic identification, and is therefore used to calculate the performance measures, such as precision (*P*) and recall (*R*). Correct and incorrect estimates of topic shifts and continuations are marked, and the other statistics given in section 3.2 are calculated and used in the evaluation of results.

4. Results

When the human expert evaluated the 10,256 Excite query dataset, 6,664 queries were included in the analysis (excluding the last query of each user session). The last query of each session cannot be included in the analysis since they have no subsequent queries to identify topic continuations or shifts. Out of 6,664 queries 6,001 topic continuations (90%) and 663 topic shifts (10%) were found. The results of the evaluation of the human expert can be seen in Table 3. In the subset used for solving the goal programming model (first half of the dataset – 5,128 queries), there were 1,858 user sessions. Thus, 3,270 queries of the first half of the dataset were used to solve the goal programming model to mark each query category in the first half as a topic continuation or shift. Out of 3,270 queries, there were 2,879 topic continuations (88%) and 391 topic shifts (12%). In the second half of the dataset, 5128 queries were considered in 1734 user sessions. Eliminating the last query of each session leaves 3,394 queries to be included in the analysis. Out of 3,394 queries, 3,122 (92%) were topic continuations, whereas 272 (8%) were topic shifts.

After each query in the second half of the dataset is marked as a topic continuation or shift based on the solution of the goal programming model from the first half, the results in Table 4. For comparison, we also include the results of human evaluation on the second half of the dataset. We were obtained that the automatic new topic identification approach based on the goal programming model solution marked 2,447 queries as topic continuation, whereas the human expert identified 3,122 queries as topic continuation. Similarly, our approach marked 947 queries as topic shifts, whereas the human expert identified 272 queries as topic shifts.

Table 3: Topic shifts and continuations in the dataset as evaluated by human expert

	Total number of queries	Number of sessions	No. of queries considered by the approach	Total no. of shifts marked by the human expert	Total no. of continuations marked by the human expert
1st half of dataset used for training	5128	1858	3270	391	2879
2nd half of dataset used for testing	5128	1734	3394	272	3122
Entire Dataset	10256	3592	6664	663	6001

Table 4: Topic shifts and continuations on the second half of the dataset

Origin of results	No. of queries in analysis	No. of topic shifts N_{shift}	No. of topic contins N_{contin}	Correctly Estimated number of shifts $\boldsymbol{N_{shift\ \&\ correct}}$	Correctly estimated number of contins $\boldsymbol{N_{contin\ \&\ correct}}$	Type A error	Type B error	P_{shift}	R_{shift}	$F_{ß(shift)}$
Results from approach	3394	947	2447	263	2438	684	9	0.278	0.967	0.503
Results from human expert	3394	$N_{true\ shift}$ 272		-----	-----	----	----	----	----	----

An important result we observe in Table 4 is that, our approach identified all of the topic changes correctly except 9 giving a Type B error value of 9, which yields an R_{shift} equals 0.967, a satisfactory result. In addition, 2,438 topic continuations out of 3,122 continuations were estimated correctly. These results denote a high level of estimation of topic shifts. On the other hand, the automatic new topic identification approach yielded 947 topic shifts when actually there are 272, giving a value of 0.278 for P_{shift}. The result shows that the new topic identification approach based on the solution of the goal programming model overestimates the number of topic shifts.

5. Conclusion

Content information of search engine user queries is an important dimension of user behavior analysis in information retrieval. This study proposes an automatic new topic identification approach based on a goal programming formulation to automatically identify topic changes in a user session by using statistical characteristics of queries, such as time intervals and query reformulation patterns. As a result, all the performance measures yielded satisfactory results, especially for topic continuations. Hence, we conclude that the goal programming approach can be successfully used in automatic new topic identification of search engine data logs, and we believe that it has a promising application potential in information science in general. Future work directions include testing the suggested automatic new topic identification with goal programming approach on other datasets, and integrating it into query clustering and recommendation algorithms for more effective information retrieval systems.

Acknowledgements

This research has been funded by TUBITAK, a National Young Researchers Career Development Project 2005: Fund Number: 105M320: “Application of Web Mining and Industrial Engineering Techniques in the Design of New Generation Intelligent Information Retrieval Systems”.

References

1. Beeferman, D. and Berger, A. (2000). "Agglomerative clustering of a search engine query log", *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining***,** Boston, MA, pp. 407 - 416.
2. Goker, A., and He, D.(2000). "Analyzing Intranet logs to determine session boundaries for user-oriented learning", *Proceedings of AH2000: The international conference on adaptive hypermedia and adaptive web-based systems*, Trento, Italy, pp. 319-322.
3. Huang, X., Yao, Q. and An, A. (2006). "Applying language modeling to session identification from database trace logs", *Knowledge and Information Systems*, Vol. 10, pp. 473–504.
4. Jansen, B.J., Spink, A., Blakely, C. and Koshman, S. (2007). "Defining a session on Web search engines", *Journal of the American Society for Information Science and Technology*, Vol. 58, pp. 862-871.
5. Liu, X., Croft, W. B., Oh, P. and Hart, D. (2004). "Automatic recognition of reading levels from user queries*", Proceedings of the 27th ACM ACM international conference on research and development in information retrieval (SIGIR '04)*, pp. 548–549.
6. Metzler, D. and Croft, W.B. (2005). "Analysis of Statistical Question Classification for Fact-based Questions", *Information Retrieval*, Vol. 8, pp. 481-504.
7. Murray, G.C., Lin J. and Chowdhury, A. (2006). "Identification of user sessions with hierarchical agglomerative clustering", *Proceedings of ASIST 2006: Annual Meeting of the American Society for Information Sciences and Technology*.
8. Ozmutlu, S. (2006). "Automatic New Topic Identification Using Multiple Linear Regression", *Information Processing and Management*, Vol. 42 No 4, pp. 934-950.
9. Ozmutlu, S., Buyuk, B. and Ozmutlu, H.C. (2007). "Using Conditional Probabilities for automatic new topic identification", *Online Information Review*, Vol. 31, pp. 491-515.
10. Ozmutlu, H.C. and Cavdur, F. (2005a). "Application of automatic topic identification on excite web search engine data logs", *Information Processing and Management*, Vol. 41, pp. 1243-1262.
11. Ozmutlu, S and Cavdur, F. (2005b). "Neural Network Applications for Automatic New Topic Identification*", Online Information Review*, Vol. 29, pp. 35-53.
12. Ozmutlu, H.C., Cavdur, F. and Ozmutlu, S. (2006). "Automatic New Topic Identification in Search Engine Datalogs", *Internet Research*, Vol. 26 No3, pp. 323-338.
13. Ozmutlu, H.C., Cavdur, F., Ozmutlu, S. and Spink, A., (2004a). "Neural Network Applications for Automatic New Topic Identification on Excite Web search engine datalogs", *Proceedings of ASIST 2004, Annual Meeting of the American Society for Information Science and Technology*, Providence, RI, pp. 310-316.
14. Ozmutlu, S., Ozmutlu, H.C. and Spink, A. (2003). "Multitasking Web searching and implications for design", *Proceedings of ASIST 2003, Annual Meeting of the American Society for Information Science and Technology*, Long Beach, CA, pp. 416-421.
15. Ozmutlu, S., Ozmutlu, H.C. and Spink, A. (2008). "Topic analysis and identification of queries". Jansen, B. J., Spink, A. and Taksa, I. (Eds.), *Handbook of Web log analysis*, IGI, Hershey, PA.
16. Ozmutlu, S., Spink, A. and Ozmutlu, H.C. (2002). "Analysis of large data logs: an application of Poisson sampling on excite web queries", *Information Processing and Management*, Vol. 38, pp. 473-490.
17. Pu, H.T., Chuang, Shui-Lung and Yang, C. (2002). "Subject Categorization of Query Terms for Exploring Web Users' Search Interests", *Journal of the American Society for Information Science and Technology*, Vol. 53, pp. 617–630.
18. Shafer, G. (1976). *A mathematical theory of evidence*, Princeton University Press, Princeton, NJ.
19. Spink, A., Ozmutlu, H. C., & Ozmutlu, S. (2002). "Multitasking information seeking and searching processes", *Journal of the American Society for Information Science and Technology*, Vol. 53, pp. 639-652.
20. Wen, J.R., Nie, J.Y. and Zhang, H.J. (2002). "Query Clustering Using User Logs", *ACM Transactions on Information Systems*, Vol. 20, pp. 59–81.