On the usage of conditional probabilities for automatic new topic identification

*Fatih Cavdur and H. Cenk Ozmutlu
Department of Industrial Engineering, Uludag University, Gorukle, Bursa, Turkey

Abstract:

Using conditional probabilities for automatic new topic identification is an efficient approach compared to the other studies of new topic identification due to its significant performance as well as relatively easy implementation. In this paper, we analyze the usage of conditional probabilities approach for automatic new topic identification, and extend the approach by considering the position of a query, namely query number, as an input for the computation of conditional probabilities besides the other -mostly used- parameters (time interval and search pattern). Specifically, we consider four different settings where these three parameters together as well as their 2-combinations are used for the conditional probability computations. The performance analysis of the approach with these settings is also presented.

Keywords: information retrieval, search engine user behavior analysis, automatic new topic identification, statistical analysis

1. Introduction

With the rapid expansion of the Internet and making it a non-trivial task to search for information on the Web, search engines became the most used information retrieval tools, and efficient use of them hence is important. Personalization of search engines might help us to achieve that efficiency, but it is difficult to do so due to the changes of human information needs as well as their search engine usage patterns. On the focus of some of these efforts to improve the efficiency of search engine usage is the analysis and understanding of the behaviors of search engine users. Identification or at least estimation of topic changes from the queries of a user can be considered as one of the directions, namely *automatic new topic identification*. A well-defined algorithm for automatic new topic identification using some query characteristics was introduced by He *et al*. [4]. Since then, several other studies suggested new approaches to the problem of automatic new topic identification to achieve a better performance and / or to allow an easier implementation. He *et al*. [4] used two "information sources" as the query characteristics in their study to be used to estimate the topic changes, namely *time interval* and *search pattern*, which correspond to the time difference between two consecutive query submissions and some textual pattern relationships (explained in detail in the following sections of the paper) between two consecutive queries, respectively. In other studies later on, these two information sources were mostly considered as well to be used by different approaches. Ozmutlu [10], however, considered the possibility that the position of a query might be related to the topic change behavior of a user, and verified that relationship using a regression model.

*Corresponding author: Address: Department of Industrial Engineering Uludag University, 16059, Bursa TURKEY. E-mail address: fatihcavdur@uludag.edu.tr, Phone: +902242942077 Fax: +902244428021

Using conditional probabilities was also a suggested approach of automatic new topic identification in another study [13]. According to this approach, the conditional probabilities of topic changes are computed given the time intervals and search patterns of the queries, and then, used for automatic new topic identification. The performance of the approach was significant, and it was easier to implement compared to the other suggested approaches of automatic new topic identification.

In our study, we analyze the usage of conditional probabilities approach for automatic new topic identification introduced by Ozmutlu *et al*. [13], and extend the approach by considering query position as an information source in addition to time interval and search pattern using the fact that the position of a query might also be related to the topic change behavior of a user as shown by Ozmutlu [10]. We can thus investigate if using query position in addition to time interval and search pattern for the computation of conditional probabilities, contributes to the performance of the automatic new topic identification.

More research efforts have been focused on the analysis of search engine transaction logs with the growth of Internet and increased use of search engines for information retrieval needs. Some researchers performed large-scale data analysis on search engine transaction logs. Among them are Jansen *et al*. [5], Silverstein, *et al*. [17], Spink *et al*. [18] and Ozmutlu *et al*. [6], [12], [15]. These studies mostly focused on the statistical characteristics of the queries. Some other studies on search engine transaction logs focused on topic changes occurred during search sessions. It has been shown that some of the statistical characteristics of search engine queries can be used for the identification or estimation of topic changes. In fact, He *et al*. [4] proposed an automatic new topic identification algorithm, where they used the query characteristics time interval and search pattern to estimate topic changes. This well-designed algorithm later was used by Ozmutlu *et al*. in some studies on different datasets [7], [8]. Some new approaches for automatic new topic identification were also presented by the researchers. Some other automatic new topic identification approaches were also suggested, such as, using neural networks, support vector machines and multiple linear regression for automatic new topic identification [10], [11], [14]. In another study, Ozmutlu *et al*. suggested using conditional probabilities [13]. In this approach, the conditional probabilities of topic changes (given time intervals and search patterns of the corresponding queries) are computed and used for new topic identification. The approach stands out with being as successful as the others and with its relatively easy implementation.

2. Materials and Method

Based on Ozmutlu *et al*.'s [13] approach, the probability that a topic continuation (and shift) occurs given a particular time interval and search pattern, is computed for each query, and the probability is then used to mark the corresponding query as a topic continuation or shift, where a topic shift simply means a topic change occurs whereas the term topic continuation is used to state that no topic change occurs in two consecutive queries in a search engine user session.

The conditional probabilities used in our study here are computed using time intervals, search patterns and query numbers as well as their 2-combinations (i.e., (i) time interval – search pattern, (ii) time interval – query number and (iii) search pattern – query number). We compute the

conditional probability of a topic continuation (and shift) given a particular time interval, search pattern and query number as well as their above mentioned 2-combinations. Hence, the probabilities are computed by conditioning on the four different settings of these parameters, i.e., (i) time interval and search pattern (the same as Ozmutlu *et al.*'s study [13]), (ii) time interval and query number, (iii) search pattern and query number, and finally all three, (iv) time interval, search pattern and query number. We therefore obtain four different conditional probabilities of topic continuation (and shift) for a particular query as seen in Table 1. These probabilities are then used to mark each query as a topic continuation or topic shift as it is done in Ozmutlu *et al.* [13].

Table 1: Four Different Parameter Settings for Computing Conditional Probabilities

Setting	Query Characteristics Used to Compute Conditional Probabilities
Setting 1	Time Interval and Search Pattern
Setting 2	Time Interval and Query Number
Setting 3	Search Pattern and Query Number
Setting 4	Time Interval, Search Pattern and Query Number

Two datasets are used for the evaluation of our approach. The first dataset comes from the logs of Excite search engine. 1,025,910 queries were collected on December 20, 1999, and approximately 10,000 of them are sampled to use for our evaluations. The second dataset is obtained from the FAST search engine data logs. The data were collected from 12.00 a.m. on February 6, 2001 to 12.00 a.m. on February 7, 2001 (Norwegian time). Again, approximately 10,000 of 1,257,891 queries are sampled for the analysis. In both datasets, the entries are ordered based on their arrival times. Each user is identified with an anonymous user id in the datasets, and each record has the following three fields: (i) an anonymous user id assigned by the search engine, (ii) time of day, in hours, minutes and seconds, (iii) search query.

The idea behind the usage of conditional probabilities for automatic new topic identification is to compute the probability of a topic continuation (and shift) for a particular query given the corresponding characteristics of that particular query based on the above mentioned settings. Therefore, for the four settings considered here, we compute the conditional continuation and shift probabilities of $P(c|TI,SP)$, and $P(s|TI,SP)$, $P(c|TI,QN)$ and $P(s|TI,QN)$, $P(c|SP,QN)$ and $P(s|SP,QN)$, and finally, $P(c|TI,SP,QN)$ and $P(s|TI,SP,QN)$, respectively, where TI, SP and QN shows the time interval, search pattern and query number of the corresponding query, and c and s correspond to the continuation and shift, respectively.
The same notation as Ozmutlu *et al.*'s study [13] is also used here as summarized below.

Session: A sequence of queries submitted to the search engine by a single user.
Topic Shift: A change from one topic to another between queries within a session.
Topic Continuation: Staying on the same topic from one query to another within a session.
N_{contin} (N_{shift}): Number of queries marked as topic continuation (shift) by the approach.
$N_{\text{true-contin}}$ ($N_{\text{true-shift}}$): Number of queries marked as topic continuation (shift) by the human expert.
$N_{\text{contin\&correct}}$ ($N_{\text{shift\&correct}}$): Number of queries marked as topic continuation (shift) by the approach and human expert.

Type A Error: A Type A Error occurs when a topic continuation is considered as a topic shift.
Type B Error: A Type B Error occurs when a topic shift is considered as a topic continuation.

Similar to Ozmutlu *et al.*'s work [13], performance measures precision (*P*), recall (*R*) and a combination of these two (F_β) are used to evaluate the performance of the approach. These performance measures are computed as shown below. A more detailed explanation on these quantities can be found in Ozmutlu *et al.*'s paper [13].

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

$$P_{\text{shift}} = \frac{N_{\text{shift\&correct}}}{N_{\text{shift}}} \tag{2}$$

$$R_{\text{contin}} = \frac{N_{\text{contin\&correct}}}{N_{\text{true-contin}}} \tag{3}$$

$$R_{\text{shift}} = \frac{N_{\text{shift\&correct}}}{N_{\text{true-shift}}} \tag{4}$$

$$F_{\beta-\text{contin}} = \frac{\left(1+\beta^2\right)P_{\text{contin}}R_{\text{contin}}}{\beta^2 P_{\text{contin}} + R_{\text{contin}}} \tag{5}$$

$$F_{\beta-\text{shift}} = \frac{\left(1+\beta^2\right)P_{\text{shift}}R_{\text{shift}}}{\beta^2 P_{\text{shift}} + R_{\text{shift}}} \tag{6}$$

3. Results

Using the three parameters to compute the conditional probabilities; namely, time interval, search pattern and query number, as well as the 2-combination of these, we obtain four different settings where the conditional continuation and shift probabilities of (i) $P(c|TI,SP)$ and $P(s|TI,SP)$, (ii) $P(c|TI,QN)$ and $P(s|TI,QN)$, (iii) $P(c|SP,QN)$ and $P(s|SP,QN)$, and finally, (iv) $P(c|TI,SP,QN)$ and $P(s|TI,SP,QN)$ are computed. The two datasets are first examined by a human expert with fluent English to mark the true topic continuations and true topic shifts. The human expert is also provided consultation by the native speakers of some other languages (German, Italian and Turkish). The data in both datasets are divided into two approximately equal-sized parts as seen in Table 2 where the first one is used for the computation of the probabilities and the second one for the test of the approach. The reason that the parts do not have the same number of records in the first and second halves of the datasets is to be able to include a session in either the first or the second parts of the datasets (not to divide a session into two parts with some queries in the first half and the others in the second).

Since the conditional probabilities are computed using time interval, search pattern and query number, each query in the datasets is categorized in terms of these query characteristics. Time interval is simply the time difference between two consecutive query submissions. The second parameter, search pattern, shows some textual pattern relationships between the terms of two

consecutive queries. Both for time interval and search pattern, the same categorizations are used as of Ozmutlu & Cavdur's study [7]. Finally, query number is just the number or the order of the query in a session starting from one to the number of queries in that particular session.

Table 2: Number of queries in each dataset

	Excite	FAST
Entire dataset	1,025,910	1,257,891
Sample dataset	10,003	10,007
1st half of the sample dataset	5,014	4,997
2nd half of the sample dataset	4,989	5,010

To categorize time intervals, seven classes of time intervals are used; namely 0-5 minutes, 5-10 minutes,10-15 minutes, 15-20 minutes, 20-25 minutes, 25-30 minutes and 30+ minutes. The distribution of queries with respect to time interval for both datasets is shown in Table 3.

Table 3: Distribution of Queries with respect to Time Intervals

Time Interval (min)	Intra-Topic (Excite)	Inter-Topic (Excite)	Intra-Topic (FAST)	Inter-Topic (FAST)
0-5	3001	77	3466	95
5-10	219	18	283	27
10-15	84	14	112	24
15-20	47	7	56	19
20-25	22	13	33	17
25-30	20	5	24	10
30+	151	135	200	194
Total	3544	269	4174	386

Seven categories of search pattern as in Ozmutlu & Cavdur's study [7] are used (see the study [7] for more details on the categories). The search patterns are automatically identified by a computer program using a modified version of the search pattern identification algorithm by He *et al*. [4]. Figure 1 shows the algorithm first implemented by Ozmutlu & Cavdur [7], and we obtain the distribution of queries with respect to search patterns as in Table 4 by running this algorithm.

Finally, the distribution of queries with respect to query numbers is given in Table 5. Again, seven classes of query numbers are used; namely 0-10 queries, 10-20 queries, 20-30 queries, 30-40 queries, 40-50 queries, 50-60 queries and 60+ queries. After characterizing each query in terms of these three quantities, we can now compute the conditional probabilities of topic shifts and topic continuations given any combination of time interval, search pattern and query number. Hence, each query in the dataset is categorized with respect to the corresponding class combinations of time interval, search pattern and query numbers, i.e., (time interval, search pattern), (time interval, query number), (search pattern, query number), and (time interval, search pattern, query number). Therefore, a total of 49 (7^2) categorizations for the first three settings and 343 (7^3) categorizations for the last setting are obtained. We use the breakdown of shifts and continuations with respect to the query categories to compute the conditional probabilities. The

conditional probability of a topic continuation and shift is computed by dividing the number of topic continuations and topic shifts in a certain category by the total number of queries in that category. For example, in the Excite dataset, considering our 4th setting, there are 333 topic continuations and 59 topic shifts in the query category (1, 5, 1). That is, time interval class 1, search pattern class 5 and query number class 1. Hence, the conditional probability of a topic continuation and shift are computed as 333 / (333 + 59) = 0.8495 and 59 / (333 + 59) = 0.1505 (or 1 – 0.8495 = 0.1505), respectively. All conditional probability computations of topic continuations and shifts for each query categorization are available upon request.

Input: Queries Q_{i-1}, Q_i, Q_{i+1} // set of three consecutive queries
Local:
Q_c (current query) as string
Q_n (next query) as string
$B = \{t | t \in Q_c \text{ and } t \in Q_n\}$ // set of terms that are included in both queries
$C = \{t | t \in Q_c \text{ and } t \notin Q_n\}$ // set of terms that are included in current query only
$D = \{t | t \notin Q_c \text{ and } t \in Q_n\}$ // set of terms that are included in next query only
Output: SP (Search Pattern)

begin
if $(Q_i == \emptyset)$ **then**
if $(i == 1)$ **then**
$SP =$ Other;
else
$Q_c = Q_{i-1}$
$Q_n = Q_{i+1}$
end if
else
$Q_c = Q_i$
$Q_n = Q_{i+1}$
end if
$SP =$ Other // default value
if $(Q_n = \emptyset)$ **then** $SP =$ Relevance Feedback **end if**
if $(Q_n = Q_c)$ **then** $SP =$ Next Page **end if**
if $(B \neq \emptyset \text{ and } C \neq \emptyset \text{ and } D == \emptyset)$ **then** $SP =$ Generalization **end if**
if $(B \neq \emptyset \text{ and } C == \emptyset \text{ and } D \neq \emptyset)$ **then** $SP =$ Specialization **end if**
if $(B \neq \emptyset \text{ and } C \neq \emptyset \text{ and } D \neq \emptyset)$ **then** $SP =$ Reformulation **end if**
if $(Q_c \neq Q_n \text{ and } B \neq \emptyset \text{ and } C == \emptyset \text{ and } D == \emptyset)$ **then** $SP =$ Reformulation **end if**
if $(Q_c \neq \emptyset \text{ and } B == \emptyset)$ **then** $SP =$ New **end if**
end

Figure 1: Search Pattern Identification Algorithm

Summary of the results of the human expert evaluation is shown in Table 6. We note that there are 10,003 and 10,007 queries in the Excite and FAST datasets, respectively. The first half of the Excite dataset contains 5,014 queries, and there are 4,989 queries in the second half. The numbers

of queries of the two halves of the FAST dataset are 4,997 and 5,010, respectively. Based on the human expert's evaluation, we note that the number of topic continuations and shifts are larger in the FAST dataset than they are in the Excite dataset. Also, the number of Excite sessions is larger than the number of FAST sessions in our samples.

Table 4: Distribution of Queries with respect to Search Patterns

Search Pattern	Intra-Topic (Excite)	Inter-Topic (Excite)	Intra-Topic (FAST)	Inter-Topic (FAST)
Browsing	2371	0	3100	5
Generalization	58	0	39	0
Specialization	166	0	136	2
Reformulation	327	1	276	5
New	622	268	551	370
Relevance Feedback	0	0	70	2
Others	0	0	2	2
Total	3544	269	4174	386

Table 5: Distribution of Queries with respect to Query Numbers

Query Number	Intra-Topic (Excite)	Inter-Topic (Excite)	Intra-Topic (FAST)	Inter-Topic (FAST)
0-10	2620	227	2691	256
10-20	523	27	1097	96
20-30	165	10	295	25
30-40	63	2	77	7
40-50	40	0	14	2
50-60	33	1	0	0
60+	100	2	0	0
Total	3544	269	4174	386

Table 6: Summary of the results of the human expert evaluation

	Dataset	Total # of Queries	# of sessions	# of queries considered for analysis	# of continuations (human expert)	# of shifts (human expert)
1st half	Excite	5,014	1,201	3,813	3,544	269
1st half	FAST	4,997	437	4,560	4,174	386
2nd half	Excite	4,989	1,322	3,667	3,515	152
2nd half	FAST	5,010	526	4,484	4,174	310
Entire Dataset	Excite	10,003	2,523	7,480	7,059	421
Entire Dataset	FAST	10,007	963	9,044	8,348	696

Now, using the computed conditional probabilities computed from the first parts of the datasets, we can estimate the topic continuations and shifts in the second parts. The larger probability value (between the conditional continuation and shift probabilities) is used for the estimation, i.e., if the corresponding conditional topic continuation probability is greater than 0.5, then, we

estimate the corresponding query as a topic continuation or vice versa. The results obtained from our four approaches are compared to the actual topic continuations and shifts determined by the human expert. To evaluate the performance of our settings, correct and incorrect estimates are identified, and the performance measures, precision (P), recall (R) and F_β are computed. The results are summarized in Table 7 where the best results for the corresponding performance measures and related results (Type A and B Errors, P_{contin}, R_{contin}, $F_{\beta\text{-contin}}$, P_{shift}, R_{shift}, $F_{\beta\text{-shift}}$) are highlighted.

Table 7: Results of the four different settings

Dataset	Excite				FAST			
Setting	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
N_{contin}	3,438	3,660	3,659	3,426	4,207	4,230	4,472	4,182
N_{shift}	226	2	0	222	276	253	11	291
$N_{\text{contin\&correct}}$	3,366	3,511	3,508	3,353	4,044	4,020	4,167	4,024
$N_{\text{shift\&correct}}$	80	2	0	77	146	100	5	150
Type A Error	146	0	0	145	130	153	6	141
Type B Error	72	149	151	73	163	210	305	158
P_{contin}	0.9791	0.9559	0.9587	0.9787	0.9613	0.9504	0.9318	0.9622
R_{contin}	0.9576	0.9989	0.9980	0.9539	0.9689	0.9631	0.9983	0.9641
$F_{\beta\text{-contin}}$	96.55%	98.38%	98.30%	96.30%	96.60%	95.83%	97.25%	96.34%
P_{shift}	0.3540	1.0000	#DIV/0!	0.3468	0.5290	0.3953	0.4545	0.5155
R_{shift}	0.5263	0.1320	0.0000	0.5066	0.4710	0.3226	0.0161	0.4839
$F_{\beta\text{-shift}}$	44.57%	2.08%	#DIV/0!	43.25%	49.10%	34.62%	2.51%	49.52%

In general, setting 1, setting 2, setting 3 and setting 4 give 5, 4, 3 and 4 of the best results out of 16, respectively. We note that setting 1 (based on time interval and search pattern) gives the best Type B Error, P_{contin}, R_{shift} and $F_{\beta\text{-shift}}$ values for the Excite dataset, and the best P_{shift} value for the FAST dataset. Setting 1 and setting 4 are the more similar to each other in terms of their performance measure values. Setting 4 (based on time interval, search pattern and query number) gives the best Type B Error, P_{contin}, R_{shift} and $F_{\beta\text{-shift}}$ values for the FAST dataset. We note that setting 1 produce better results on the Excite dataset while setting 4 is more successful on the FAST dataset. From the results, we can also make the following observations. In terms of continuation based performance measures (especially R_{contin} and $F_{\beta\text{-contin}}$), setting 2 and 3 are the most successful approaches. Also, another interesting point is that we obtain the best results on the Excite dataset using setting 2 (based on time interval and query number) whereas setting 3 is the better (based on search pattern and query number) on the FAST dataset.

4. Discussion

In terms of the contributions of this study, comparing setting 1 introduced by Ozmutlu *et al*. [13] and the others (settings 2, 3 and 4), we note that settings 2, 3 and 4 produce some of the best results (using these, we are able to obtain 11 of the best performance measure values out of 16), which illustrates the usage of the suggested settings has a potential to improve the results.

In general, we note that the different settings might produce better results on different datasets as well as in terms of different performance measures. In the context of this study, that implies the differences between some of the characteristics of the queries (datasets) and their relationships to the performance measures defined. For example, for some search engine users, it might be more descriptive to use time interval as an information source to estimate topic continuations or shifts whereas for some others that might be search pattern, query order or a combination of those. Moreover, some particular quantities might be more useful to be used for the estimation of topic continuations rather than shifts whereas some others might just be the opposite.

Conclusions

In this paper, we analyze the usage of conditional probabilities approach for automatic new topic identification, and extend it by considering the position of a query, namely query number, as an input for the computation of conditional probabilities besides the other parameters, time interval and search pattern. Two sample datasets with approximately 10,000 queries from search engines Excite and FAST are used for the implementations. We start by the implementation of the approach suggested by Ozmutlu *et al*. [13] where time interval and search pattern are used to compute the conditional probabilities, and extend it by adding query number as well as considering the 2-combinations of these three parameters to be used to compute the conditional probabilities. Ozmutlu *et al*.'s [13] approach of using conditional probabilities for automatic new topic identification is successful (as successful as the ones previously suggested in the literature), and yet, it is relatively easier to implement compared to the other studies of automatic new topic identification. In our study, by the different settings we employ, we are able to observe some improvements over the initial setting implemented by Ozmutlu *et al*. [13].

In general, our results indicate the differences between the datasets, and hence, search engine usage patterns, and also the differences between the effects of using different information sources for the implementation of the approach on the performance. Extending our implementations beyond this study might significantly contribute to the automatic new topic identification literature; however, more experiments using larger and different datasets need to be performed before generalizing our results.

References

[1] Cooley, R., Mobasher, B. & Srivastava, J. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems 1999; 1(1): 5-32.
[2] Haykin, S. Neural Networks. Macmillan College Publishing, Englewood Cliffs, NJ; 1994
[3] He, D. & Goker, A. Detecting session boundaries from web user logs. Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research, Cambridge UK; 2000, 57-66.
[4] He, D., Goker, A. & Harper, D. J. Combining evidence for automatic Web session identification. Information Processing and Management 2002; 38(5): 727-742.
[5] Jansen, B. J., Spink, A. & Saracevic, T. Real life, real users, and real needs: A study and analysis of user queries on the Web. Information Processing and Management 2000; 36(2): 207-227.
[6] Ozmultu, H. C. & Spink, A. Characteristics of question format web queries: An exploratory study. Information Processing and Management 2002; 38(4): 453-471.
[7] Ozmutlu, H. C. & Cavdur, F. Application of automatic topic identification on excite web search engine data logs. Information Processing and Management 2005; 41(5): 1243-1262.
[8] Ozmutlu, H. C., Cavdur, F. & Ozmutlu, S. Automatic New Topic Identification in Search Engine Datalogs. Internet Research 2006; 16: 323-338.
[9] Ozmutlu, H. C., Cavdur, F. & Ozmutlu, S. Cross Validation of Neural Network Applications for Automatic New Topic Identification. Journal of the American Society of Information and Society 2008; 59(3): 339-362.
[10] Ozmutlu, S. Automatic new topic identification using multiple linear regression. Information Processing and Management 2006; 42(4): 934-950.
[11] Ozmutlu, S. & Cavdur, F. Neural network applications for automatic new topic identification. Online Information Review 2005; 29(1): 35-53.
[12] Ozmutlu, S., Ozmutlu, H. C. & Spink, A. A day in the life of Web searching: An exploratory study. Information Processing and Management 2004; 40(2): 319-345.
[13] Ozmutlu, S., Ozmutlu, H. C. & Buyuk, B. Using conditional probabilities for automatic new topic identification. Online Information Review 2007; 31(4): 491-515.
[14] Ozmutlu, S., Ozmutlu, H. C. & Spink, A. Using Support Vector Machines for Automatic New Topic Identification. Proceedings of ASIST 2007: 70th Annual Meeting of the American Society for Information Science and Technology, Madison, WI, November 2007.
[15] Ozmutlu, S., Spink, A. & Ozmutlu, H. C. Analysis of large data logs: An application of Poisson sampling on excite web queries. Information Processing and Management 2002; 38(4): 473-490.
[16] Shafer, G. A Mathematical Theory of Evidence. Princeton University Press, Princeton, NJ 1976.
[17] Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. Analysis of a very large Web search engine query log. AGM SIGIR Forum 1999, 6-12.
[18] Spink, A., Bateman, J. & Jansen, B. J. Searching heterogeneous collections on the Web: A survey of Excite users. Internet Research: Electronic Networking Applications and Policy 1999; 9(2): 117-128.