Automatic New Topic Identification in Search Engine Transaction Logs using Multiple Linear Regression*

Seda Ozmutlu
Department of Industrial Engineering
Uludag University
Muhendislik-Mimarlik Fakultesi,
Gorukle, Bursa, TURKEY
Tel: (++90-224) 294-2085
Fax: (++90-224) 442-8021
E-mail:seda@uludag.edu.tr

H. Cenk Ozmutlu
Department of Industrial Engineering
Uludag University
Muhendislik-Mimarlik Fakultesi,
Gorukle, Bursa, TURKEY
Tel: (++90-224) 294-2082
Fax: (++90-224) 442-8021
E-mail:hco@uludag.edu.tr

Amanda Spink**
School of Information Systems
Faculty of Information Technology
Queensland University of Technology
126 Margaret Street GPO Box 2434
Brisbane QLD 4001 Australia
Phone: +61 7 3138 9583
Email: ah.spink@qut.edu.au

Abstract

Content analysis of search engine user queries is an important task for search engine research, and identification of topic changes within a user search session is a key issue in content analysis of search engine user queries. The purpose of this study is to provide automatic new topic identification of search engine query logs, and estimate the effect of statistical characteristics of search engine queries on new topic identification. By applying multiple linear regression and ANOVA on a sample data log from the FAST search engine, we have reached the following findings: 1) We demonstrated that the statistical characteristics of Web search queries are effective on shifting to a new topic; 2) Multiple linear regression is a successful tool for estimating topic shifts and continuations. This study provides statistical proof for the relationship between the non-semantic characteristics of Web search queries and the occurrence of topic shifts and continuations.

1. Introduction and Related Research

The World Wide Web, and its search tools, the search engines, are becoming the major source of information for many people. It is important, for this reason, to study the behavior of search engine users. One dimension of search engine user profile is content-based behavior. Currently, search engines are not designed to differentiate according to the user's profile and the content that the user is interested in. However, exploiting the user's interest in various topics and developing a search engine, which is able to understand or at least estimate the user interests will be a significant improvement in search engine research.

During a search session, some users are interested in multiple topics. It was observed that 10-30% of search engine users performed multitasking searches [1], [2]. Considering multitasking behavior of search engine users, one of the most important facets of content-based analysis is new topic identification. New topic identification is discovering when the user has switched from one topic to another during a single search session. Estimating the arrival of a new topic from a user will be very useful in developing effective information retrieval algorithms necessary for efficient search engines that would provide better results to the Web users. Besides providing better results to the user, custom-tailored graphical user interfaces can be offered to the Web search engine user, if topic changes were estimated correctly by the search engine [2]

There are many large scaled studies on search engine datalogs, such as those of Silverstein et al. [3], Spink, et al. [4], and Ozmutlu, et al. [5]. The number of studies on content analysis is few, the reason generally being the effort required to manually process the queries for topic identification; however content analysis is a growing area [6]. Some researchers, such as Silverstein, et al. [3] and Spink, et al. [4] have performed content analysis of search engine data logs at the term level, and have observed that the highest ranking terms are related to topics of pornography, entertainment and education. Besides term analysis, Spink, et al. [4] and Ozmutlu, et al. [5] have performed analysis of a sample of queries at the conceptual or topical level and discovered that the top category in subject of queries was entertainment and recreation. Another research area in content-related search engine research is developing query clustering models based on content information. Pu et al. [6]

1530-1605/08 $25.00 © 2008 IEEE

developed an automatic classification methodology to classify search queries into broad subject categories. Muresan and Harper [7] and Beeferman and Berger [8] propose a topic modeling system for developing mediated queries.

Studies on search engine transaction logs usually analyzed the queries semantically. Semantic analysis of queries is a promising line of research, but is a complicated task; hence its current success is ambiguous One promising approach is to use content-ignorant methodologies to the problem of query clustering or new topic identification in a user search session. 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. Ozmutlu and Cavdur [9] used Dempster-Shafer Theory [10] for automatic new topic identification. They automatically identified topic changes using statistical data from Web search logs. In other studies, Ozmutlu et al. [11] and Ozmutlu and Cavdur [12] applied an alternative content-ignorant methodology, namely artificial neural networks, to automatically identify topic changes. In these studies, neural networks also identified topic shifts fairly successfully; however there were still some problems with the estimation of topic shifts.

In this study, we aim to estimate topic shifts in search engine query logs using multiple linear regression and demonstrate the statistical significance of the relationship between non-semantic characteristics of query logs and topic shifts/continuations. Using the characteristics of the search queries as independent factors and the existence of topic shifts as the dependent factor, multiple linear regressions is applied to investigate the relationship between statistical characteristics and topic shifts. We also apply ANOVA to examine the structure of the variance of the topic shifts with respect to the statistical characteristics of the search queries. These studies will be helpful in identifying whether there is a relationship between statistical characteristics of the search queries and topic shifts/continuations. If such a relationship exists, content-ignorant methodologies can be expected to be successful.

We initially present the literature review related to topic identification, followed by the description of the methodology, results and the conclusion.

2. Methodology

3.1. Research question

The research question in this study is to observe whether there is statistical relationship between topic shifts within consecutive queries and characteristics of search engine user queries. In addition, we aim to provide successful estimation of topic shifts in consecutive queries within a user session. In order to perform these tasks, we apply multiple linear regression (Montgomery, 1991) on a search engine query log. We also apply ANOVA to examine the structure of the variance of the topic shifts with respect to the statistical characteristics of the search queries.

3.2. The dataset

The FAST search engine (http://www.alltheweb.com) provided a query log of 1,257,891 for our analysis. Queries were collected from 12:00 AM (Norwegian time) on February 6, 2001 for 24 hours until 12:00 AM February 7, 2001. In the FAST data log structure, the entries are given in the order they arrive. FAST assigns a new user ID to every new user and it is possible to identify new sessions through these user IDs. In addition, FAST gives each query is a time stamps in hours, minutes and seconds. We selected a sample of 10,007 queries from the total of 1,257,891 queries. The sample size was not kept very large, since evaluation of the performance of the algorithm would require a human expert to go over all the queries. The sample was selected using Poisson sampling [13] to provide a sample dataset that is both statistically representative of the entire data set and small enough to be analyzed conveniently.

3.3. Notation

The following notation is used in this study:

N_{shift}: Number of queries labeled as shifts by multiple linear regression (MLR)

N_{contin} : Number of queries labeled as continuation by MLR

$N_{true\ shift}$: Number of queries labeled as shifts by manual examination of human expert (HU)

$N_{true\ contin}$: Number of queries labeled as continuation by manual examination of HU

$N_{shift\ \&\ correct}$: Number of queries labeled as shifts by MLR and by manual examination of HU

$N_{contin\ \&\ correct}$: Number of queries labeled as continuation by MLR and by manual examination of HU

Type A error: This type of error occurs in situations where queries on same topics are considered as separate topic groups.
Type B error: This type of error occurs in situations where queries on different topics are grouped together into a single topic group.

Some useful formulation related to the above notation is as follows:

$$N_{true\ shift} = N_{shift\ \&\ correct} + Type\ B\ error \quad (1)$$
$$N_{true\ contin} = N_{contin\ \&\ correct} + Type\ A\ error \quad (2)$$
$$N_{shift} = N_{shift\ \&\ correct} + Type\ A\ error \quad (3)$$
$$N_{contin} = N_{contin\ \&\ correct} + Type\ B\ error \quad (4)$$

The performance measures of Precision (P), Recall (R) and their combination, a fitness function (F_β) are used in this study to demonstrate the performance of MLR. The focus of precision and recall are both on correctly estimating the number of topic shifts and continuations. P_{shift} is the correctly estimated number of shifts by MLR among all the shifts marked by MLR (Eq. 5). R_{shift} is the correctly estimated number of shifts by MLR among all the shifts marked by the HU (Eq. 6). P_{contin} is the correctly estimated number of continuations by MLR among all the continuations marked by MLR (Eq. 7) and recall R_{contin} is the correctly estimated number of continuations by MLR among all the continuations marked by the HU (Eq. 8). The fifth measure, fitness function F_β (Eq. 9 and 10), combines P and R values into a single value, where β is a parameter to prioritize different types of resulting errors of estimating topic changes. The reason for choosing an F_β combining P and R is to provide a single parameter to compare different results and to be consistent with previous studies [11]. These performance measures are used to demonstrate the performance of MLR for estimating topic shifts and continuations, and their formulation are as follows:

$$P_{shift} = \frac{N_{shift\ \&\ correct}}{N_{shift}} \quad (5) \qquad P_{contin} = \frac{N_{contin\ \&\ correct}}{N_{contin}} \quad (6)$$

$$R_{shift} = \frac{N_{shift\ \&\ correct}}{N_{true\ shift}} \quad (7) \qquad R_{contin} = \frac{N_{contin\ \&\ correct}}{N_{true\ contin}} \quad (8)$$

$$F_\beta = \frac{(1+\beta^2) P_{shift} R_{shift}}{\beta^2 P_{shift} + R_{shift}} \quad (9)$$

$$F_{\beta\_contin} = \frac{(1+\beta^2) P_{contin} R_{contin}}{\beta^2 P_{contin} + R_{contin}} \quad (10)$$

3.4. Statistical Analysis

The research question in this study is to observe whether there is statistical relationship between characteristics of search engine user queries and topic shifts. In order to perform this task, we apply MLR on the dataset. MLR is a model, where more than one regressor variable is involved to explain and estimate a dependent variable [14].

The dependent factor for the multiple regression model in our study is existence of topic shift/topic continuation. Topic continuations are marked as "1" and topic shifts are marked as "2". The independent factors are search pattern of queries (SP), time interval of queries (TI) and the order of the search query (QN) in the user session (QN = k means that the query under consideration is the k^{th} query of the session). Two factor interactions are usually considered to be effective on dependent factors, so we considered the SP-TI interaction, SP-QN interaction and TI-QN interaction. Considering these terms, the hypothesized MLR model is as shown in equation (11):

$$Y = \beta_0 + \sum_{i=1}^{3} \beta_i F_i + \sum_{i=1}^{2} \sum_{j=i+1}^{3} \beta_n F_i F_j + \varepsilon \quad (11)$$

where F_i = factor i, i=1,2,3,
F_j = factor j, j=i+1,….,3,
n = i+j+1 and $\varepsilon \sim iid\ N(\mu, \sigma^2)$.

This model has a total of 7 terms. The coefficient of this linear regression model of considerable size can be determined with the data from the FAST query log. The levels of the main factors will be given in the next section.

Testing the overall significance of the regression model requires hypothesis testing, which is as follows [14]:

$$H_0 : \beta_1 = \beta_2 = \ldots\ldots\ldots \beta_6 = 0 \quad (12)$$
$$H_1 : \beta_i \neq 0 \ for\ at\ least\ one\ i,\ i=1,\ldots,6$$

In addition to MLR, we apply ANOVA (Montgomery, 1991), where we can test the significance of each regressor on the dependent variable. Since the search queries are not designed with respect to a certain experimental design and acquired from the actual query log, sequential sum of squares are used to test for the significance of each factor in the model. [14]. The hypothesis testing for testing the significance of each independent factor is as follows:

$$H_0 : \beta_i = 0 \quad (13)$$
$$H_1 : \beta_i \neq 0 \ ,i=1,\ldots,6$$

This hypothesis test tests whether each regression coefficient is zero or not. If the null hypothesis cannot

be rejected for a coefficient, this means that that certain factor does not have a significant effect on the dependent parameter, i.e. does not effect whether the query is a topic shift or continuation.

3.5. Methodology

In this section, we present the methodology applied in the paper to investigate the relationship between topic shifts and search query characteristics.

Evaluation by human expert

A HU goes through the 10,007 query set for FAST and marks the actual topic changes and topic continuations. This step is necessary to form the dependent factor for the MLR equation and also for testing the performance of the regression equation.

Dividing the data into two sets

Approximately, first half of the data is used to form a regression equation and the second half is used to test the performance of the proposed MLR equation. The two data sections do not contain the same number of queries to keep the entirety of the user session containing the query in the middle of the datasets. The first half of the FAST dataset contains 4997 queries, and the second half of the FAST dataset contains 5010 queries.

Identifying search pattern, time interval and order of each query in the dataset

Each query in the dataset is categorized in terms of its search pattern, time interval, and its order in the user session. The time interval is the difference of the arrival times of two consecutive queries. The search pattern is the change of terms of the consecutive queries within a session. The order of the query in the session is the position of a query in the user session. The categorization of time interval and search pattern is selected similar to those of [11], [12] and [9] to be in harmony with previous studies.

We use seven categories of time intervals for a query: 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30+ minutes. The levels of the time interval factor used in MLR are 1 through 7. See Table 1 for distribution of the queries with respect to time interval in the FAST dataset. It should be noted that not all of 4997 queries in FAST can be used for training, since the last query of each user session cannot be processed for pattern classification and time duration, since there are no subsequent queries after the last query of each session. In the training dataset for FAST, excluding the last query of each session, the test dataset is reduced to 4560 queries from 4997.

Table 1. Distribution of time interval of queries

Time Interval (min)	FAST Intra-topic	FAST Inter-topic
0-5	3464	95
5-10	285	27
10-15	112	24
15-20	56	19
20-25	33	17
25-30	24	10
30+	200	194
Total	4174	386

We also use seven categories of search patterns in this study, which are as follows [9], [11], [12]:

- Unique (New): the second query has no common term compared to the first query.
- 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.
- Relevance feedback: the second query is empty (i.e., contains no terms) and it is generated by the system when the user selects the choice of ‘‘related pages ’’.
- Other: If the second query does not fit any of the above categories, it is labeled as other.

The search patterns are automatically identified by a computer program. The logic for the automatic search pattern identification can be summarized as in Figure 1. Also see Table 2 for distribution of queries with respect to search patterns in the training dataset. The levels of the search pattern factor used in MLR are 1 through 7.

Input: Queries Q_{i-1}, Q_i, Q_{i+1} (set of three subsequent queries)

Local: Q_c, current query (as a string)

Q_n, next query (as a string)

$B = \{t \mid t \in Q_c \text{ and } t \in Q_n\}$, the set of terms (terms determined using "space" as a divider) that are common in both Q_c and Q_n

$C = \{t \mid t \in Q_c \text{ and } t \notin Q_n\}$, the set of terms, which appear in Q_c only

$D = \{t \mid t \notin Q_c \text{ and } t \in Q_n\}$, the set of terms, which appear in Q_n only

Output: Search Pattern, *SP*

begin
  if ($Qi == \phi$) **then**
    if ($i == 1$) **then** *SP = Other*,
    else $Q_c = Q_{i-1}$ //if Q_i is empty (relevance feedback) take preceding query (Q_{i-1}) to analyze relationship
      $Q_n = Q_{i+1}$,
    endif
  else $Q_c = Q_i$,
    $Q_n = Q_{i+1}$,
  endif
  SP = other //default value
  if ($Q_n == \phi$) **then** *SP = Relevance Feedback* **endif** // if the next query is empty then //it is relevance feedback
  if ($Q_n == Q_c$) **then** *SP = Next Page* **endif**
  if ($B \neq \phi$ and $C \neq \phi$ and $D == \phi$) **then** *SP = Generalization* **endif**
  if ($B \neq \phi$ and $C == \phi$ and $D \neq \phi$) **then** *SP = Specialization* **endif**
  if ($B \neq \phi$ and $C \neq \phi$ and $D \neq \phi$) **then** *SP = Reformulation* **endif**
  if ($Q \neq Q$ and $B \neq \phi$ and $C == \phi$ and $D == \phi$) **then** *SP= Reform* **endif**
  if ($Q_c \neq \phi$ and $B == \phi$) **then** *SP = New* **endif**
end

Figure 1. Search pattern identification algorithm

Table 2. Distribution of search pattern of queries

Search Pattern	FAST Intra-topic	FAST Inter-topic
Browsing	3100	5
Generalization	39	0
Specialization	136	2
Reformulation	276	5
New	551	370
Relev. feedback	70	2
Other	2	2
Total	4174	386

The order of a query in the session is the position of that query in that user session, which is assigned with respect to the query's arrival time. The queries are categorized up to query 47 in the user sessions, since the highest number of queries within a session is 47 queries in the entire dataset. The distribution of the queries with respect to query order in the session is given in Table 3. The level of the "query order" factor in the MLR could be 1 through 47.

Table 3. Distribution of order of queries

Order of query in user session	FAST Intra-topic	FAST Inter-topic	Order of query in user session	FAST Intra-topic	FAST Inter-topic
1st query	368	49	24th query	32	5
2nd query	358	34	25th query	28	4
3rd query	321	37	26th query	27	1
4th query	297	26	27th query	22	2
5th query	272	24	28th query	16	3
6th query	256	18	29th query	16	1
7th query	235	16	30th query	14	2
8th query	213	17	31st query	12	1
9th query	195	18	32nd query	11	2
10th query	180	17	33rd query	12	0
11th query	166	13	34th query	9	1
12th query	156	12	35th query	8	1
13th query	139	13	36th query	8	0
14th query	128	14	37th query	6	1
15th query	112	11	38th query	5	0
16th query	100	10	39th query	4	0
17th query	91	7	40th query	2	1
18th query	79	6	41st query	3	0
19th query	66	6	42nd query	3	0
20th query	59	4	43rd query	3	0
21st query	51	2	44th query	3	0
22nd query	46	2	45th query	0	2
23rd query	40	3	46th query	1	0
			47th query	1	0

Identifying coefficients of MLR

The coefficients of the MLR equation are determined using the first half of the dataset. Multiple factor ANOVA is also performed. The details of how these methods can be applied are in Montgomery [14].

Applying MLR on the second half of the dataset

In order to validate the MLR equation, the second portion of the FAST dataset is used. By using time interval, search pattern and query order values as independent factors, we determine whether the query is a topic shift or not using the regression equation. Naturally, the regression equation yields a real number. However, the answer has to be of the form 1 or 2, where 1 is topic continuation and 2 is topic shift. In order to bring the regression equations response to a binary form, we apply a threshold value. We use threshold values between 1.1 and 1.9, since regression was trained to provide answers between 1 and 2.

Comparison of results from HU and the MLR equation

The results of the MLR equation tested on the second half of the FAST dataset are compared to the actual topic shifts and identifications determined by the HU. Correct and incorrect estimates of topic shift and continuation are marked and the statistics in the notation section are calculated, which are used in the evaluation of results.

Evaluation of results

The performance of the MLR equation is evaluated in terms of the performance measures given in the previous section (Equations 5 to 9). Higher performance measure values mean higher success in topic identification.

4. Results and Discussion

The MLR equation, where topic shifts are the dependent factor and the characteristics of the query log are the independent factors is as follows:

$$Y = 1.14 - 0{,}135\ TI - 0{,}0354\ SP - 0{,}000864\ QN + 0{,}0491\ TI*SP - 0{,}0163\ TI*QN + 0{,}00184\ SP*QN \quad (14)$$

Using this regression equation, it may be possible to identify topic shifts and continuations in a Web search query log. To test the validity of the regression equation, hence perform the hypothesis test in Equation 12, the F value for the regression equation is determined. The sum of squares and the calculation of the F value are as in Table 4. The F value for the regression equation is 4893.58. The critical F value is $F_{0.05,6,4990}$, which is equal to 2.10. The regression equation is valid and the relationship between the topic shifts and characteristics of the search queries is statistically significant. This results shows that there is a certain relationship between non-semantic characteristics of search queries and occurrence of topic shifts/continuations. Hence, a Web user might be demonstrating certain querying patterns, when he/she is about to make a topic shift. These results are also indication that content-ignorant topic identification methodologies can be successful for automatic new topic identification.

We applied multiple factor ANOVA to investigate the effect of the characteristics of the search queries on topic shifts. The results of the multiple factor ANOVA are in Table 5. The critical F value for testing the effect of each factor is F0.05,1,4990, which is equal to 3.84. From these results, we observe that the independent main factors TI (time interval), SP (search pattern) and QN (query number) are effective on occurrence of topic shifts/continuations. In earlier studies, the position of a query was not considered as one of the characteristics that could affect topic shifts and continuations. In addition, the interactions of TI-SP and SP-QN are also effective on topic shifts/continuations. The TI-QN interaction does not have an impact on topic shifts/continuations. This might mean that at certain time interval/search pattern and search pattern/query number combinations, the occurrence of topic shifts and continuations increase. It requires further research to identify exactly which time interval and search pattern combinations cause this effect.

When the HU evaluated the 10,007 query dataset, 8348 topic continuations and 696 topic shifts were found. Eliminating the last query of each session leaves 9044 queries to be included in the analysis. In the subset used for training (first half of the dataset (5014 queries), there are 4174 topic continuations and 386 topic shifts, and in the second half of the dataset (4989 queries), there are 4174 topic continuations and 310 topic shifts.

Table 4. ANOVA for testing the significance of the MLR equation

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F value
Regression	1552,45	6	258,74	4893,58
Error	263,84	4990	0,05	
Total	1816,29	4996		

Table 5. Multiple factor ANOVA for testing the significance of independent factors

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F value
TI	1145.28	1	1145.28	22905.6
SP	136.04	1	136.04	2720.8
QN	2.25	1	2.25	45
TI*SP	266.43	1	266.43	5328.6
TI*QN	0.1	1	0.1	2
SP*QN	2.34	1	2.34	46.8
Error	263.85	4990	0.05	
Total	1816.29	4996		

After applying the MLR equation on the second half of the dataset, the results we obtained with different threshold values are summarized in Table 6. From Table 6, considering threshold=1.1; we observe that the MLR equation marked 3541 queries as topic continuation, whereas the HU identified 4174 queries as topic continuation. Similarly, the regression equation marked 943 queries as topic shifts, whereas the HU identified 310 queries as topic shifts. During the topic identification process, we observed 652 Type A errors and 19 Type B errors. Using the MLR approach, an R_{shift} value of 0.526 and an R_{contin} value of 0.958, are observed, which demonstrate that the topic shifts were estimated somewhat correctly and topic continuations were estimated almost entirely correctly by the regression equation. On the other hand, the regression equation yielded a value of 0.309 for P_{shift}. This results means that the regression equation overestimates the number of topic shifts. This result is due to the threshold vale, and increases with increasing threshold values. Overall, we observe that MLR yielded fairly successful results for new topic identification, and very successful results for identification of topic continuations.

5. Conclusion

This study uses multiple linear regression and multiple factor ANOVA to identify the relationships between topic shifts and the non-semantic characteristics of the search queries, and successfully estimate topic shifts and continuations. The non-semantic characteristics of the search queries are the time interval of queries, the search pattern of queries and the order of a query in a search session.

Hypothesis testing showed that the multiple linear regression equation is statistically valid, which demonstrates that there is a valid relationship between non-semantic characteristics of user queries and topic shifts and continuations. Multiple factor ANOVA showed that the non-semantic factors of time interval, search pattern and query position in the user session, as well as the search pattern and time interval interaction, have a statistically significant effect on topic shifts. These results provide statistical proof that Web users demonstrate a certain way of behavior when they are about to make topic shifts or continue on a topic, which is exacerbated when a certain combination of search pattern and time interval occurs. The determination of exact multi-tasking behavior patterns requires future research. These findings also indicate that, since non-semantic characteristics of query logs are relevant to topic shifts/continuations, content-ignorant methodologies for automatic new topic identification can be a promising line of research.

Table 6. Topic shifts and continuations in the entire dataset as result of regression equation and as evaluated by the HU

Origin of results	Number of topic shifts N_{shift}	Number of topic continuations N_{contin}	$N_{shift\&correct}$	$N_{contin\&correct}$	Type A error	Type B error	P_{shift} - R_{shift} $F_{\beta(shift)}$	P_{contin} R_{contin} $F_{\beta(contin)}$
Results from MLR Threshold =1.1	943	3541	291	3522	652	19	0.31 0.94 0.53	0.99 0.84 0.89
Threshold =1.2	600	3884	230	3804	370	80	0.38 0.74 0.55	0.98 0.91 0.93
Threshold =1.3	453	4031	204	3925	249	106	0.45 0.66 0.56	0.97 0.94 0.95
Threshold =1.4	388	4096	183	3969	205	127	0.47 0.59 0.39	0.97 0.95 0.96
Threshold =1.5	342	4142	168	4000	174	142	0.49 0.54 0.52	0.97 0.96 0.96
Threshold =1.6	315	4169	157	4016	158	153	0.49 0.50 0.50	0.96 0.96 0.96
Threshold =1.7	241	4243	119	4052	122	191	0.49 0.38 0.42	0.96 0.97 0.96
Threshold =1.8	12	4472	0	4162	12	310	0 0 --	0.93 0.99 0.97
Threshold =1.9	12	4472	0	4162	12	310	0 0 --	0.93 0.99 0.97
Results from the HU	$N_{trueshift}$= 310	$N_{truecontin}$ = 4174	----	----	----	----	----	-----

We also used the proposed multiple regression equation to estimate topic shifts and continuations on the second portion of the dataset. The regression equation was quite successful in estimating topic shifts and continuations; however the number of topic shifts was overestimated. Future work includes repeating the linear regression approach on more datasets from different search engines to test its performance of automatic new topic identification.

Acknowledgements

This research has been funded by TUBITAK, Turkey and is 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] A. Spink, H.C. Ozmutlu and S. Ozmutlu, "Multitasking information seeking and searching processes", Journal of the American Society for Information Science and Technology, 2002, 53, pp. 639-652.
[2] S. Ozmutlu, H.C. Ozmutlu and A. Spink, "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, 2003, pp. 416-421.
[3] C. Silverstein, M. Henzinger, H. Marais and M. Moricz, " Analysis of a very large Web search engine query log". ACM SIGIR Forum, 1999, 33, pp. 6-12.
[4] A. Spink, D. Wolfram, B.J. Jansen and T. Saracevic, "Searching the Web: The public and their queries", Journal of the American Society for Information Science and Technology, 2001, 53, pp. 226–234.
[5] S. Ozmutlu, H.C. Ozmutlu and A. Spink, "A day in the life of Web searching: an exploratory study", Information Processing and Management, 2004, pp. 40, 319-345.
[6] H.T. Pu, Chuang, L. Shui and C. Yang, « Subject Categorization of Query Terms for Exploring Web Users' Search Interests", Journal of the American Society for Information Science and Technology, 2002, 53,pp. 617–630.
[7] G. Muresan and D.J. Harper, "Topic Modeling for Mediated Access to Very Large Document Collections", Journal of the American Society for Information Science and Technology, 2004, 55, pp. 892–910.
[8] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log", Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, 2000, pp. 407 - 416.
[9] H.C. Ozmutlu and F. Cavdur, "Application of automatic topic identification on excite web search engine data logs", Information Processing and Management, 2005, 41, pp. 1243-1262.
[10] G. Shafer, A mathematical theory of evidence. Princeton University Press, Princeton, NJ, 1976
[11] H.C. Ozmutlu,, F. Cavdur, S. Ozmutlu and A. Spink, "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, 2004, pp. 310-316.
[12] S. Ozmutlu and F. Cavdur, "Neural Network Applications for Automatic New Topic Identification", Online Information Review, 2005, 29, pp. 35-53
[13] S. Ozmutlu, A. Spink, and H.C. Ozmutlu, "Analysis of large data logs: an application of Poisson sampling on excite web queries", Information Processing and Management, 2002, 38, pp. 473-490.
[14] D.C. Montgomery, Design and Analysis of Experiments, 3rd Ed., John Wiley and Sons, New York, 1991.