Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

USING MONTE-CARLO SIMULATION FOR AUTOMATIC NEW TOPIC IDENTIFICATION OF SEARCH ENGINE TRANSACTION LOGS

Seda Ozmutlu
Huseyin C. Ozmutlu
Buket Buyuk

Industrial Engineering Department, School of Engineering and Architecture
Uludag University
Gorukle, Bursa, 16059, TURKEY

ABSTRACT

One of the most important dimensions of search engine user information seeking behavior and search engine research is content-based behavior, and limited research has focused on content-based behavior of search engine users. The purpose of this study is to present a simulation application on information science, by performing automatic new topic identification in search engine transaction logs using Monte Carlo simulation. Sample data logs from FAST and Excite are used in the study. Findings show that Monte Carlo simulation for new topic identification yields satisfactory results in terms of identifying topic continuations, however the performance measures regarding topic shifts should be improved.

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 (Spink, et al. 2002, Ozmutlu, et al. 2003). 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 (Ozmutlu, et al. 2003).

There are many large scaled studies on search engine datalogs, such as those of Silverstein et al.(1999), Spink, et. al. (2001), and Ozmutlu, et al.(2004). 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 (Pu, et al., 2002). Some researchers, such as Silverstein, et al. (1999) and Spink, et al. (2001) 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. (2001) and Ozmutlu, et al. (2004) 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. (2002) developed an automatic classification methodology to classify search queries into broad subject categories. Muresan and Harper (2004) and Beeferman and Berger (2004) 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 inter-

vals between subsequent queries or the reformulation of queries. Ozmutlu and Cavdur (2005a) and Ozmutlu, et al. (2006) used Dempster-Shafer Theory (Shafer, 1976) for automatic new topic identification. They automatically identified topic changes using statistical data from Web search logs. In other studies, Ozmutlu et al. (2004) and Ozmutlu and Cavdur (2005b) 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 propose to use Monte Carlo simulation application to automatically identify topic changes in search engine transaction logs. The methodology does not contain semantic analysis, and relies on the statistical characteristics of the queries, such as the time between query submissions and the reformulation of the queries. The initial indications of the relation between statistical characteristics of queries and topic change were shown in Spink et al. (2002) and He and Goker (2000). In this study, the conditional probabilities used in the Monte-Carlo simulation is calculated using a sample from real search engine datalogs and simulation is used to identify topic changes on a separate dataset from the same search engine. The success of Monte Carlo simulation is determined comparing its results to that of the human expert.

The layout of the paper is as follows: We initially present the literature review related to topic identification, followed by the description of the methodology, results and the conclusion.

2 METHODOLOGY

2.1 Research Design

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. The transaction log might seem outdated to the unfamiliar eye, however it is very difficult to obtain real datasets form search engines, hence analyses are made on available datasets only. 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 (Ozmutlu, et al., 2002) to provide a sample dataset that is both statistically representative of the entire data set and small enough to be analyzed conveniently.

The second search query log used in this study comes from the Excite search engine (http://www.excite.com) located in the U.S.A. The data was collected on December 20, 1999 and consists of 1,025,910 search queries. 10,003 queries of the dataset were selected as a sample. The datalog structures of Excite and FAST are similar. The third query log also comes from the Excite search engine, but was collected on May 4, 2001 and consists of 1,7 million for our analysis. We selected a sample of 10,256 queries using Poisson sampling.

We use the following concepts in the study (Spink, et al. 2001):

(1) Query: a set of one or more search terms; it may include advanced search features, such as logical operators and modifiers.

(2) Session: the entire set of queries by the same user over time. A session could be as short as one query or contain many unique and repeat queries.

2.2 Notation

The notation used in this study is below:

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 Monte-Carlo simulation

N_{contin} : Number of queries labeled as topic continuation by Monte-Carlo simulation

$N_{true\ shift}$: Number of queries labeled as topic shifts by Monte-Carlo simulation

$N_{true\ contin}$: Number of queries labeled as topic continuation by Monte-Carlo simulation

$N_{shift\&correct}$: Number of queries labeled as topic shifts by the human expert and by Monte-Carlo simulation

$N_{contin\&correct}$: Number of queries labeled as topic continuation by Monte-Carlo simulation and by manual examination of human expert

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:

$$N_{true\ shift} = N_{shift\&correct} + \text{Type B error} \quad (1)$$

$$N_{true\ contin} = N_{contin\&correct} + \text{Type A error} \quad (2)$$

$$N_{shift} = N_{shift\&correct} + \text{Type A error} \quad (3)$$

$$N_{contin} = N_{contin\&correct} + \text{Type B error} \quad (4)$$

The commonly used performance measures of Precision (P) and Recall (R) are used in this study. 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 (2005a, 2005b) precision (P_{shift}) is the correctly estimated number of shifts by the Monte-Carlo simulation among all the shifts marked by the Monte-Carlo simulation (5), and recall (R_{shift}) is the correctly estimated number of shifts by the Monte-Carlo simulation among all the shifts marked by the human expert (6). On the other hand, interpreted in terms of topic continuations, precision (P_{contin}) is the correctly estimated number of continuations by the Monte-Carlo simulation among all the continuations marked by the Monte-Carlo simulation (7) and recall (R_{contin}) is the correctly estimated number of continuations by the Monte-Carlo simulation among all the continuations marked by the human expert (8). The fifth measure, F_{β_shift} (9) combines P_{shift} and R_{shift} to provide a single parameter to compare different results and to be consistent with previous studies. β is chosen as 1.3 in this study to be consistent with the previous studies on automatic new topic identification. The sixth measure, fitness function F_{β_contin} (10), combines P_{contin} and R_{contin} into a single value. These performance measures are used to demonstrate the performance of the proposed Monte-Carlo simulation. The formulations of these measures are as follows:

$$P_{shift} = \frac{N_{shift\,\&\,correct}}{N_{shift}} \tag{5}$$

$$P_{contin} = \frac{N_{contin\,\&\,correct}}{N_{contin}} \tag{6}$$

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

$$R_{contin} = \frac{N_{contin\&correct}}{N_{truecontin}} \tag{8}$$

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

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

2.3 Proposed methodology

In this study, we propose a Monte-Carlo simulation application to identify topic changes in the Excite and FAST search engine query logs. The general steps of the methodology applied in this paper are explained next.

Evaluation by human expert: The actual topic shifts and continuations in the 10,000 query FAST and Excite data sets are identified and marked by a human expert. This step is necessary for determining the conditional probabilities that would be used for Monte-Carlo simulation, and also for testing the performance of Monte-Carlo simulation in identifying topic shifts and continuations.

Separating the data into two sets: The data is separated to two approximately equal sized sections. The first section of the datasets is used to determine the conditional probabilities that would be used for Monte-Carlo simulation and the second section is used to test the performance of Monte-Carlo simulation. The two data sections do not contain the same number of queries to keep the entirety of the user session containing the median query. Sizes of the datasets are given in Table 1.

Table 1: Size of the datasets used in the study

Search engine	Excite 2001	FAST	Excite 1999
Entire dataset	1,7 million	1,257,891	1.025,910
Sample set	10,256	10,007	10,003
1st half of the sample set	5128 queries	4997queries	5014 queries
2nd half of the sample set	5128 queries	5010 queries	4989 queries

Identification of search pattern and time interval of each query in the dataset:

Application of Monte-Carlo simulation for automatic new topic identification relies on the statistical characteristics of the queries, such as the time between query submissions and the reformulation of the subsequent queries. Therefore, each query in the datasets 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 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 Ozmutlu and Cavdur (2005a, 2005b) and Ozmutlu, et al. (2004, 2006).

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. See Table 2 for distribution of queries in the training dataset with respect to search patterns. Note that the total number of queries in Table 2 is not equivalent to the total number of queries in the training dataset. The reason is that it is impossible to identify the search pattern or the time interval of the last query in each session, since there are no subsequent queries after the last query of each session. It should be noted that not all of 5014 queries in Excite 1999, 5128 queries in Excite 2001 and 4997 queries in FAST can be used for training. 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 Excite, there were 1201 user sessions, so excluding the

last query of each session, the test dataset is reduced to 3813 queries from 5014 queries. In the training dataset for FAST, excluding the last query of each session, the training dataset is reduced to 4560 queries from 4997 queries. Similary, in the Excite 2001 dataset, the training dataset was reduced to 3270 from 5128. After the human expert identified the topic shifts and continuations, 3544 topic continuations and 269 topic shifts were identified within the 3813 queries for the Excite 1999 dataset; 4174 topic continuations and 386 topic shifts were identified within the 4560 queries for the FAST dataset, and 2879 topic continuations and 391 topic shifts were identified within the 3270 queries for the Excite 2001 dataset.

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) then take the preceding query // (Q_{i-1}) to analyze the 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 ($Q_n \neq Q_c$ and $B \neq \phi$ and $C = = \phi$ and $D = = \phi$) **then** *SP = Reformulation* **endif**
  if ($Q_c \neq \phi$ and $B = = \phi$) **then** *SP = New* **endif**
end

Figure 1: Search pattern identification algorithm

We use seven categories of search patterns in this study, which are as follows (Ozmutlu and Cavdur 2005a, 2005b, Ozmutlu, et al. 2004, 2006). We also provide examples for each search pattern.

Unique (New): the second query has no common term compared to the first query.

Example: Query j : Automobile
Query j+1 : Harry Potter

Next Page (Browsing): the second query requests another set of results on the first query.

Example: Query j : Automobile
Query j+1 : Automobile

Generalization: all of the terms of second query are also included in the first query but the first query has some additional terms.

Example: Query j : Red Automobile
Query j+1 : Automobile

Specialization: all of the terms of the first query are also included in the second query but the second query has some additional terms.

Example: Query j : Automobile
Query j+1 : Red Automobile

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.

Example: Query j : Red Automobile Toyota
Query j+1 : Automobile Corolla

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 ''.

Example: Query j : Automobile
Query j+1 : ""

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 session, this property does not hold.

Example: Q1 : ""
Q2 : Toyota Car

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 Ozmutlu, et al.(2004) and Ozmutlu and Cavdur (2005) We use seven categories of time intervals for a query: 0-5 seconds, 5-10 seconds, 10-15 seconds, 15-20 seconds, 20-25 seconds, 25-30 seconds, 30+ seconds. The distribution of the queries with respect to time interval in the training datasets for the FAST and Excite search engine datalogs can be seen in Table 3.

Computing conditional probabilities for topic shifts and continuations: Each query in the datasets is categorized with respect to its time interval and search pattern combination. This step involves 49 categories of queries, which can be listed as 1_1, 1_2; etc. The categories can be seen in Table 4. Since, all the queries have previously been tagged by the human expert as shifts and continuations, it is possible to determine the breakdown of shifts and continuations with respect to the 49 topic categories. See Table 4 for the breakdown of shifts and continuations with respect to topic categories in the first half of the datasets Using the breakdown of the shifts and continuations with respect to query categories, the conditional probability of a topic shift and continuation given the query category is computed by dividing the number of shifts in a certain category to the total number of queries in that category. For example, in query category 1-5 in the Excite dataset, there are 403 continuations and 76 shifts, totaling to 479 queries. Consequently, the conditional probability for

Table 2: Distribution of search pattern of queries

Search Pattern	Excite 2001 topic continuation	Excite 2001 topic shift	FAST topic continuation	FAST topic shift	Excite 1999 topic continuation	Excite 1999 topic shift
Browsing	1627	0	3100	5	2371	0
Generalization	88	1	39	0	58	0
Specilization	186	0	136	2	166	0
Reformulation	412	21	276	5	327	1
New	566	369	551	370	622	268
Relevance feedback	0	0	70	2	0	0
Other	0	0	2	2	0	0
Total	2879	391	4174	386	3544	269

Table 3: Distribution of time interval of queries

Time Interval (min)	Excite 2001 topic continuation	Excite 2001 topic shift	FAST topic continuation	FAST topic shift	Excite 1999 topic continuation	Excite 1999 topic shift
0-5	2265	137	3464	95	3001	77
5-10	229	38	285	27	218	18
10-15	109	28	112	24	85	14
15-20	42	8	56	19	47	7
20-25	34	13	33	17	22	13
25-30	25	9	24	10	20	5
30+	175	158	200	194	151	135
Total	2879	391	4174	386	3544	269

continuation given the query category 1_5 is 0.84; whereas the conditional probability for shift is 0.16.

Notice that there are no observations in some of the query categories. For example, in Table 4, there were no queries of the category 2_2 in the first half of the Excite 1999 dataset. In such cases, it might always be possible to encounter that category in the second half. Then, the methodology should blindly estimate whether there is a topic change or not. To stay on the safe side, we have adjusted the conditional probabilities to provide an estimate of "continuation", where there are no observations in a given category in the first half of the dataset. The resulting conditional probability values for each query type for the Excite and FAST datasets are in Table 4.

Automatic new topic identification using Monte-Carlo simulation: Arrivals of topic shifts and continuations are estimated in the second portion of the datasets using the conditional probabilities calculated in the first half of the datasets. The estimation of topic shifts and continuations is made by using Monte-Carlo simulation for each query in the test dataset. Initially, the time interval and search pattern category of the query is determined. Then, a random number is generated using Uniform (0,1). If the random number is between 0 and the conditional probability value for continuation, the query is labeled as continuation. If the random number exceeds the conditional probability value for continuation, then the query is labeled as shift. For, example consider query category 3_5 in the FAST dataset. If a query in the test dataset belongs to category 3_5, after generating the random number, the query is labeled as continuation if the random number is between 0 and 0.63; and it is labeled as shift otherwise. The study was made using MS Excel.

Since, it is not adequate to rely on results based on one simulation run only, the Monte-Carlo simulation is replicated ten times, and the average of the results of the ten replications is considered to evaluate the performance of the methodology.

Comparison of results from human expert and Monte-Carlo simulation: The results of Monte-Carlo simulation are compared to the actual topic shifts and identifications determined by the human expert for the test dataset. 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.

Table 4: Distribution of queries with respect to time interval_search pattern combination and the conditional probabilities of topic shifts and identifications given query category

		Query distribution						Conditional probabilities					
		Excite 2001		FAST		Excite 1999		Excite 2001		FAST		Excite 1999	
Time Interval	Search Pattern	Continuation (C)	Shift (S)	C	S	C	S	C	S	C	S	C	S
1	1	1358	0	2822	4	2120	0	1	0	0,999	0,001	1	0
1	2	80	0	31	0	54	0	1	0	1	0	1	0
1	3	160	0	114	2	148	0	1	0	0,983	0,017	1	0
1	4	306	9	192	2	276	1	0,9714	0,028	0,99	0,01	0,996	0,0036
1	5	361	128	244	86	403	76	0,7382	0,261	0,74	0,26	0,841	0,159
1	6	0	0	61	1	0	0	1	0	0,984	0,016	0	0
1	7	0	0	0	0	0	0	1	0	0	0	0	0
2	1	110	0	168	0	133	0	1	0	1	0	1	0
2	2	3	1	3	0	0	0	0,75	0,25	1	0	0	0
2	3	12	0	10	0	10	0	1	0	1	0	1	0
2	4	45	7	38	0	21	0	0,8653	0,134	1	0	1	0
2	5	59	30	63	27	54	18	0,6629	0,337	0,7	0,3	0,75	0,25
2	6	0	0	2	0	0	0	1	0	1	0	0	0
2	7	0	0	1	0	0	0	1	0	1	0	0	0
3	1	56	0	50	0	46	0	1	0	1	0	1	0
3	2	1	0	2	0	1	0	1	0	1	0	1	0
3	3	5	0	6	0	4	0	1	0	1	0	1	0
3	4	23	0	11	0	5	0	1	0	1	0	1	0
3	5	24	28	41	24	29	14	0,4615	0,538	0,63	0,37	0,674	0,326
3	6	0	0	2	0	0	0	1	0	1	0	0	0
3	7	0	0	0	0	0	0	1	0	1	0	0	0
4	1	16	0	18	1	20	0	1	0	0,95	0,05	1	0
4	2	2	0	2	0	0	0	1	0	1	0	0	0
4	3	4	0	3	0	1	0	1	0	1	0	1	0
4	4	7	0	9	0	6	0	1	0	1	0	1	0
4	5	13	8	23	18	20	7	0,6190	0,380	0,56	0,44	0,74	0,26
4	6	0	0	0	0	0	0	1	0	0	0	0	0
4	7	0	0	1	0	0	0	1	0	1	0	0	0
5	1	15	0	10	0	5	0	1	0	1	0	1	0
5	2	1	0	0	0	0	0	1	0	0	0	0	0
5	3	2	0	1	0	1	0	1	0	1	0	1	0
5	4	6	1	5	0	2	0	0,8571	0,142	1	0	1	0
5	5	10	12	17	17	14	13	0,4545	0,545	0,5	0,5	0,519	0,481
5	6	0	0	0	0	0	0	1	0	0	0	0	0
5	7	0	0	0	0	0	0	1	0	0	0	0	0
6	1	12	0	4	0	6	0	1	0	1	0	1	0
6	2	0	0	0	0	1	0	1	0	0	0	1	0
6	3	0	0	0	0	0	0	1	0	0	0	0	0
6	4	5	1	3	0	2	0	0,8333	0,166	1	0	1	0
6	5	8	8	17	10	11	5	0,5	0,5	0,63	0,37	0,687	0,3125
6	6	0	0	0	0	0	0	1	0	0	0	0	0
6	7	0	0	0	0	0	0	1	0	0	0	0	0
7	1	60	0	28	0	41	0	1	0	1	0	1	0
7	2	1	0	1	0	2	0	1	0	1	0	1	0
7	3	3	0	2	0	2	0	1	0	1	0	1	0
7	4	20	3	18	3	15	0	0,8695	0,130	0,857	0,143	1	0
7	5	91	155	146	188	91	135	0,3699	0,630	0,44	0,56	0,397	0,603
7	6	0	0	5	1	0	0	1	0	0,833	0,167	0	0
7	7	0	0	0	2	0	0	1	0	0	1	0	0
				4174	386	3544	269						

3 RESULTS

When the human expert evaluated the 10,256 Excite 2001 query dataset, 6664 queries were included in the analysis (excluding the last query of each user session). Out of 6664 queries 6001 topic continuations (90%) and 663 topic shifts (10%) were found. The results of the evaluation of the human expert can be seen in Table 5. In the subset used for training (first half of the dataset- 5128 queries), there were 1858 user sessions, thus 3270 queries of the first half of the dataset are used for computing the conditional probabilities. Out of 3270 queries, there are 2879 topic continuations and 391 topic shifts. In the second half of the dataset, 5128 queries were considered. Eliminating the last query of each session leaves 3394 queries to be included in the analysis. Out of 3394 queries, 3122 were topic continuations, whereas 272 were topic shifts.

When the human expert evaluated the 10,007 FAST query dataset, 9044 queries were included in the analysis. Out of 9044 queries 8348 topic continuations and 696 topic shifts were found (Table 5). 4560 queries of the first half of the dataset are used for computing conditional probabilities. Out of 4560 queries, there are 4174 topic continuations and 386 topic shifts. In the second half of the dataset, there were 5010 queries and 526 user sessions. Eliminating the last query of each session leaves 4484 queries to be included in the analysis. Out of 4484 queries, 4174 were topic continuations, whereas 310 were topic shifts.

When the human expert evaluated the 10,003 Excite query dataset, 7059 topic continuations and 421 topic shifts were found. Eliminating the last query of each session leaves 7480 queries to be included in the analysis. In the subset used for computing conditional probabilities (first half of the dataset (5014 queries), there are 3544 topic continuations and 269 topic shifts, and in the second half of the dataset (4989 queries), there are 3515 topic continuations and 152 topic shifts.

After applying Monte-Carlo simulation on the second half of the datasets to estimate topic shifts and continuations, we obtain the results in Table 6 for the Excite and FAST datasets. The results are the average of the 10 Monte-Carlo simulation runs rounded to the nearest integer. For comparison, we also include the results on the second half of the dataset as evaluated by the human expert. For the Excite 2001 dataset, we observe that Monte-Carlo simulation marked 2995 queries as topic continuation, whereas the human expert identified 3122 queries as topic continuation. Similarly, Monte-Carlo simulation marked 399 queries as topic shifts, whereas the human expert identified 272 queries as topic shifts. The number of Tye B errors were 126, and this yields a R_{shift} value of 0.54, which is a satisfactory result. In addition, 2869 topic continuations out of 3122 continuations were estimated correctly, yielding a R_{contin} value of 0,96. These results denote a high level of estimation of topic shifts and continuations. On the other hand, Monte-Carlo simulation yielded 399 topic shifts, when actually there are 272 topic shifts, giving a value of 0.37 for P_{shift}. This results means that the Monte-Carlo simulation overestimates the number of topic shifts. The reasons for this overestimation should be investigated. In terms of topic continuations P_{contin} was 0.96, 2869topic continuations out of 2995 topic continuations were estimated correctly, i.e. almost all, but 4% the topic continuations marked by the Monte-Carlo simulation were correct. The F_{β} values are also significantly higher compared to previous studies.

Table 5: Topic shifts and continuations in the Excite and FAST datasets as evaluated by human expert

	Total number of queries	Number of sessions	No. of queries considered for Monte-Carlo simulation	Total number of shifts marked by the human expert	Total no. of continuations marked by the human expert
1st half of dataset used for training	5128-Excite 2001 4997-Fast 5014-Excite 1999	1858-Excite 2001 437-Fast 1201-Excite 1999	3270-Excite 2001 4560-Fast 3813-Excite 1999	391-Excite 2001 386-Fast 269-Excite 1999	2879-Excite 2001 4174-Fast 3544-Excite 1999
2nd half of dataset used for testing	5128-Excite 2001 5010-Fast 4989-Excite 1999	1734-Excite 2001 526-Fast 1322-Excite 1999	3394-Excite 2001 4484-Fast 3667-Excite 1999	272-Excite 2001 310-Fast 152-Excite 1999	3122-Excite 2001 4174-Fast 3515-Excite 1999
Entire dataset	10256-Excite 2001 10007-Fast 10003-Excite 1999	3592-Excite 2001 963-Fast 2523-Excite 1999	6664-Excite 2001 9044-Fast 7480-Excite 1999	663-Excite 2001 696-Fast 421-Excite 1999	6001-Excite 2001 8348-Fast 7059-Excite 1999

Table 6: The results of applying Monte-Carlo simulation on the Excite and FAST datasets

Data-set	Origin of results	Queries included in analy-sis	No. of topic shifts	No. of topic contin.s	Correctly Estimated number of shifts	Correctly estimated number of contin.s	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}	$F_{ß(shift)}$	$F_{ß(contin)}$
Excite 2001	Monte-Carlo Simulation	3394	N_{shift} = 399	N_{contin} = 2995	$N_{shift\&correct}$ = 146	$N_{contin\&correct}$ = 2869	253	126	0,37	0,54	0,96	0,92	0,46	0,93
	Human expert	3394	$N_{trueshift}$ = 272	$N_{truecontin}$ = 3122	-----	-----	----	----	----	----	----	----	----	----
FAST	Monte-Carlo Simulation	4484	N_{shift}= 338	N_{contin} = 4146	$N_{shift\&correct}$ = 137	$N_{contin\&correct}$ = 3973	201	173	0,41	0,44	0,96	0,95	0,43	0,95
	Human expert	4484	$N_{trueshift}$= 310	$N_{truecontin}$= 4174	-----	-----	----	----	----	----	----	----	----	----
Excite 1999	Monte-Carlo Simulation	3667	N_{shift} = 283	N_{contin} = 3384	$N_{shift\&correct}$ = 67	$N_{contin\&correct}$ = 3299	216	85	0,24	0,44	0,98	0,94	0,26	0,69
	Human expert	3667	$N_{true\ shift}$ = 152	$N_{true\ contin}$ = 3515	-----	-----	----	----	----	----	----	----	----	----

For the FAST dataset, we observe that the Monte-Carlo simulation marked 4146 queries, whereas the human expert identified 4174 queries as topic continuation. Similarly,Monte-Carlo simulation marked 338 queries, whereas the human expert identified 310 queries as topic shifts. This yields a R_{shift} value of 0,44, which is a moderate result. In addition, 3973 topic continuations out of 4174 continuations were estimated correctly, yielding a R_{contin} value of 0.95, which is a very satisfactory result. These results show a satisfactory level of estimation of topic continuations, but less fo for topic shifts. On the other hand, the Monte-Carlo simulation yielded 338 topic shifts, giving a value of 0,41 for P_{shift}. In the FAST dataset as well as the Excite 2001 dataset, Monte-Carlo simulation overestimates the number of topic shifts. In terms of topic continuations P_{contin} was 0,96, 3973 out of 4146 topic continuations were estimated correctly, i.e. many topic continuations marked by Monte-Carlo simulation were correct.

After applying Monte-Carlo simulation on the Excite 1999 dataset, we observe that Monte-Carlo simulation marked 3383 queries as topic continuation, whereas the human expert identified 3515 queries as topic continuation. Similarly, Monte-Carlo simulation marked 283 queries as topic shifts, whereas the human expert identified 152 queries as topic shifts. The results in terms of the performance measures are similar to those of other datasets. The performance measures are poorer for the Excite 1999 dataset compared to the other datasets, and the reasons should be investigated.

4 CONCLUSION

Content information of search engine user queries is an important dimension of information retrieval and science. This study proposes a Monte-Carlo simulation application to automatically identify topic changes in a user session by using statistical characteristics of queries, such as time intervals and query reformulation patterns. Real transaction logs come from the FAST and Excite search engines. Conditional probabilities given query category were computed and Monte-Carlo simulation was applied to automatically identify topic shifts and continuations. As a result of Monte-Carlo simulation, the performance measures yielded satisfactory results, except the overestimation of topic shifts, especially for the Excite 1999 dataset. Therefore, we conclude that Monte-Carlo simulation is fairly successful in automatic identification of topic shifts and continuations in search engine data logs, and have a promising application potential in information science. Future work includes solving the problem of overestimation of topic shifts, improving the other performance measures, and integrating Monte-Carlo simulation as the automatic new topic identification tool in information retrieval and search engine algorithms.

ACKNOWLEDGMENTS

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

Beeferman D. and A. Berger. 2004. Agglomerative clustering of a search engine query log, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining***.** 407 – 416. Boston, MA, USA.

He, D. and A. Goker. 2000. Detecting session boundaries from Web user logs, *Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research*,. 57-66. Cambridge, UK.

Muresan, G. and D.J. Harper. 2004. Topic Modeling for Mediated Access to Very Large Document Collections, *Journal of the American Society for Information Science and Technology* 55: 892–910.

Ozmutlu H.C. and F. Cavdur. 2005a. Application of automatic topic identification on excite web search engine data logs. *Information Processing and Management* 41: 1243-1262.

Ozmutlu, S. and F. Cavdur. 2005b. Neural Network Applications for Automatic New Topic Identification. *Online Information Review* 29: 34-53.

Ozmutlu, H.C., F. Cavdur, and S. Ozmutlu. 2006. Automatic New Topic Identification in Search Engine Datalogs, *Internet Research: Electronic Networking Applications and Policy* 16: 323-338.

Ozmutlu, H.C., F.Cavdur, A. Spink and S. Ozmutlu. 2004. Neural network applications for automatic new topic identification on excite web search engine data logs. *Proceedings of ASIST 2004: 67th Annual Meeting of the American Society for Information Science and Technology.* 317-323. Providence, RI, USA.

Ozmutlu, S. H.C. Ozmutlu and A. Spink. 2004. A day in the life of Web searching: an exploratory study, *Information Processing and Management* 40: 319-345.

Ozmutlu, S,., 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*. 416-421. Long Beach CA.

Ozmutlu, S., A. Spink and H.C. Ozmutlu. 2002. Analysis of large data logs: an application of Poisson sampling on excite web queries. *Information Processing and Management* 38: 473-490.

Pu, H.T. , S-L Chuang, and C. Yang. 2002. Subject Categorization of Query Terms for Exploring Web Users' Search Interests, *Journal of the American Society for Information Science and Technology* 53: 617–630.

Shafer, G. 1976. *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press

Silverstein, C., M. Henzinger, H. Marais and M. Moricz. 1999. Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33: 6-12

Spink, A., H.C. Ozmutlu and S. Ozmutlu. 2002. Multitasking information seeking and searching processes, *Journal of the American Society for Information Science and Technology* 53: 639-652.

Spink, A., D. Wolfram, B.J. Jansen and T. Saracevic. 2001. Searching the Web: The public and their queries, *Journal of the American Society for Information Science and Technology*, 53: 226–234.

Talja, S, H. Keso, and T.Pietilainen. 1999. The production of 'context 'in information seeking research: a metatheoretical view", *Information Processing and Management* 35: 751-763.

AUTHOR BIOGRAPHIES

SEDA OZMUTLU is an associate professor in Industrial Engineering Department of Uludag University, Turkey. She has a Ph.D. in Industrial Engineering from PennState University. Her research interest lies primarily in the application of numerical and statistical techniques on information science related problems and search engine research, and development of various information systems. Her web page is http://www20.uludag.edu.tr/~seda.

HUSEYIN C. OZMUTLU is an associate professor in Industrial Engineering Department of Uludag University, Turkey. He has a Ph.D. in Industrial Engineering from PennState University. His research interest lies primarily in the application of numerical, statistical and operations research techniques on information science related problems, search engine research and telecommunication systems. His web page is http://www20.uludag.edu.tr/~hco.

BUKET BUYUK is a research assistant and graduate student in Industrial Engineering Department of Uludag University, Turkey.