Cross Validation Of Neural Network Applications For Automatic New Topic Identification

H. Cenk Özmutlu, Corresponding Author

Department of Industrial Engineering, Uludag University, Gorukle Kampusu, Bursa, TURKEY Tel: (++90-224) 442-8176 Fax: (++90-224) 442-8021. hco@uludag.edu.tr

Fatih Çavdur

Department of Industrial Engineering, Uludag University, Gorukle Kampusu, Bursa, TURKEY Tel: (++90-224) 442-8176 Fax: (++90-224) 442-8021. fcavdur@uludag.edu.tr

Amanda Spink

School of Information Sciences, University of Pittsburgh, 610 IS Building, 135 N. Bellefield Ave, Pittsburgh, PA 15260 Tel: (412) 624-9454 Fax: (412) 648-7001. aspink@sis.pitt.edu

Seda Özmutlu

Department of Industrial Engineering, Uludag University, Gorukle Kampusu, Bursa, TURKEY Tel: (++90-224) 442-8176 Fax: (++90-224) 442-8021. seda@uludag.edu.tr

Abstract

There are recent studies in the literature on automatic topic-shift identification in Web search engine user sessions; however most of this work applied their topic-shift identification algorithms on data logs from a single search engine.

The purpose of this study is to provide the cross-validation of an artificial neural network application to automatically identify topic changes in a web search engine user session by using data logs of different search engines for training and testing the neural network. Sample data logs from the Norwegian search engine FAST (currently owned by Overture) and Excite are used in this study. Findings of this study suggest that it could be possible to identify topic shifts and continuations successfully on a particular search engine user session using neural networks that are trained on a different search engine data log.

Keywords: search engine, topic identification, session identification, neural networks

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. A search engine, which is able to understand or at least estimate the user interests or the topics user is interested in, will be a significant improvement in developing intelligent search engines.

One of the main elements in developing an intelligent search engine 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 query clustering algorithms. With more effective query clustering algorithms, search engines can increase the quality of the results presented to the user and better satisfy users' information needs.

There are few studies on new topic identification. The studies generally analyzed the queries semantically. Some researchers, such as Silverstein, et al. (1999), Jansen, et al. (2000), Spink, et al. (2001) have performed content analysis of search engine data logs at the term level. Besides term analysis, Jansen, et al. (2000) and Spink, et al. (2001, 2002a) have also 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, closely followed by sex, pornography and preferences. Ozmutlu, et al. (2004b) and Beitzel, et al. (2004) have done hourly

statistical and topical analysis search engine query logs, and have found that the popularity of topics vary throughout the day.

Besides studies analyzing search engine queries for content information, another research area 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 using subject taxonomies. Muresan and Harper (2004) propose a topic modeling system for developing mediated queries. Beeferman and Berger (2000) and Wen, et al. (2002) applied query clustering that uses search engine query logs including clickthrough data, which provides the documents that the user have selected as a result of the search query. Query similarities are proposed based on the common documents that users have selected.

During a search session, some users are interested in multiple topics. Multitasking is performing more than one task simultaneously. In terms of information retrieval, multitasking information seeking and searching processes or, in short, multitasking is defined as "the process of searches over time in relation to more than one, possibly evolving, set of information problems (including changes or shifts in beliefs, cognitive, affective, and/or situational states" (Spink, et al., 2002b). Miwa (2001) and Spink, et al. (2002b) show that users' searches have multiple goals or topics. Spink, et al. (1999) found that 3.8% of Excite users responding to a Web-based survey reported multitasking searches. In other studies, it was observed that in a datalog of the Excite search engine collected for a day in 1999, 11.4% of users performed multitasking searches and in a datalog of the FAST search engine collected for a day in 2001, 31.8% of users performed multitasking searches (Ozmutlu, *et al.*, 2003).

Most query clustering methods, as explained above, are focused on interpretation of keywords or understanding the topic or the contents of the query, which significantly complicates the process of query clustering and increases the potential noise of the results of the study. 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. There are few studies adapting such an approach. He, et al. (2002) proposed a topic identification algorithm that uses Dempster-Shafer Theory (Shafer, 1976) and genetic algorithms. Their algorithm automatically identifies topic changes using statistical data from Web search logs. He et al. (2002) used the search pattern and duration of a query for new topic identification. He, et al.`s (2002) approach was replicated on Excite search engine data (Ozmutlu and Cavdur, In Press,a). Main finding of Ozmutlu and Cavdur (In Press,a) was that the idea of using query patterns and time intervals in identifying topic shifts is valuable, but there are indications of some problems in

the application of this idea. Ozmutlu, et al. (2004a) and Ozmutlu and Cavdur (in press,b) have shown that neural networks are successful in automatically identifying topic shifts. In these studies the neural network was tested on the datasets that it was trained on.

In this study, the main research question is to see whether neural networks continue to provide successful estimates of topic shifts given that they are tested on datasets other than the one they are trained on. Hence, if the neural network is trained with dataset A, we will test it on dataset B and vice versa. The results are compared to the cases in the previous studies (Ozmutlu, et al. (2004a) and Ozmutlu and Cavdur (in press,b)), where the neural network's training and testing datasets are from the same search engine.

Methodology

The datasets

The first 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. Approximately the first 10,000 queries of the dataset were selected as a sample from the Excite dataset. 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 second dataset comes from the FAST search engine (http://www.alltheweb.com) and contains a query log of 1,257,891 queries. Queries were collected from 12:00 AM (Norwegian time) on February 6, 2001 for 24 hours until 12:00 AM February 7, 2001. In both data log structures, the entries are given in the order they arrive. It is possible to identify new sessions through a user ID and each query contains three fields: 1) Identification: anonymous code assigned by Excite server to a user 2) Time of day: in hours, minutes, and seconds (in US West coast time) 3) Query: user terms as entered. We selected a sample of 10,007 queries from 963 users from a total of 1,257,891 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 (Details on Poisson sampling can be seen in Ozmutlu, et al. (2002)).

Notation

The notation used in this study is below:

N_{shift}: Number of queries labeled as shifts by the neural network
N_{contin}: Number of queries labeled as continuation by the neural network
$N_{true\ shift}$: Number of queries labeled as shifts by manual examination of human expert
$N_{true\ contin}$: Number of queries labeled as continuation by manual examination of human expert
$N_{shift\ \&\ correct}$: Number of queries labeled as shifts by the neural network and by manual examination of human expert
$N_{contin\ \&\ correct}$: Number of queries labeled as continuation by the neural network 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:

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

The 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 (In Press,a) and He et al. (2002), precision (P_{shift}) is the correctly estimated number of shifts by the neural network among all the shifts marked by the neural network (Eq.5), and recall (R_{shift}) is the correctly estimated number of shifts by the neural network among all the shifts marked by the human expert (Eq.6). On the other hand, interpreted in terms of topic continuations, precision (P_{contin}) is the correctly estimated number of continuations by the neural network among all the continuations marked by the neural

network (Eq.7) and recall (R_{contin}) is the correctly estimated number of continuations by the neural network among all the continuations marked by the human expert (Eq.8). These performance measures are used to demonstrate the performance of the proposed artificial neural network. The formulation of these measures are as follows:

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

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

Proposed Algorithm and Research Design

The general steps of the methodology and research design applied in this paper are explained in detail in the following paragraphs.

Evaluation by human expert:

A human expert goes through the 10,003 query set for Excite and 10,007 query set for FAST and marks the actual topic changes and topic continuations. This step is necessary for training the neural network and also for testing the performance of the neural network.

Dividing the data into two sets:

For both datasets, approximately, first half of the data is used to train the data and the second half is used to test the performance of the neural network. 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 size of the datasets to train and test the neural network is seen in Table 1.

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

Each query in the dataset is categorized in terms of its search pattern and time interval. The time interval is the difference of the arrival times of two consecutive queries. The classification of the search patterns is based on terms of the consecutive queries within

a session. The categorization of time interval and search pattern is selected similar to those of He *et al*(2002), Ozmutlu, et al., 2004a, Ozmutlu and Cavdur (in press,a) and Ozmutlu and Cavdur (in press,b) to avoid any bias during comparison.

Table 1: Size of the datasets used in the study

Search engine	Excite	Fast
Entire dataset	1,025,910	1,257,891
Sample set	10,003	10,007
1st half of the sample set used for training the neural network	5014 queries	4989 queries
2nd half of the sample set used for training the neural network	4997queries	5010 queries

We use seven categories of time intervals for a query: 0-5 minutes, 5-10 minutes, 10-15 minutes, 15-20 minutes, 20-25 minutes, 25-30 minutes, 30+ minutes. See Table 2 for distribution of the queries with respect to time interval in the Excite and Fast training datasets. It should be noted that not all of 5014 queries in Excite and 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 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, there were 437 user sessions, so excluding the last query of each session, the test dataset is reduced to 4560 queries from 4997 queries. For the Excite dataset, 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 FAST dataset, 4174 topic continuations and 386 topic shifts were identified within the 4560 queries.

We also use seven categories of search patterns in this study, which are as follows (Ozmutlu, et al., 2004a, Ozmutlu and Cavdur, in press,a, Ozmutlu and Cavdur, in press,b):

- 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. This means that the user has added and deleted some terms of the first query. Also if the user enters the same terms of the first query in different order, it is also considered as reformulation.
- Relevance feedback: the second query has zero terms (empty) and it is generated by the system when the user selects "related pages ".
- Others: If the second query does not fit any of the above categories, it is labeled as other.

Table 2: Distribution of time interval of queries

Time Interval (min)	Excite Intra-topic	Excite Inter-topic	FAST Intra-topic	FAST Inter-topic
0-5	3001	77	3466	95
5-10	218	18	283	27
10-15	85	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

Table 3: Distribution of search pattern of queries

Search Pattern	Excite Intra-topic	Excite Inter-topic	FAST Intra-topic	FAST Inter-topic
Browsing	2371	0	3100	5
Generalization	58	0	39	0
Specilization	166	0	136	2
Reformulation	327	1	276	5
New	622	268	551	370
Relevance feedback	0	0	70	2
Other	0	0	2	2
Total	3544	269	4174	386

The search patterns are automatically identified by a computer program. The pattern identification algorithm is adapted from He et al. (2002), but is considerably altered. The logic for the automatic search pattern identification can be summarized as in Figure 1.

Also see Table 3 for distribution of queries with respect to search patterns in the training datasets.

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

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

Forming the neural network:

In this study, we propose a neural network with three layers; an input layer, one hidden layer and an output layer. There are two neurons in the input layer. One neuron corresponds to categories of search patterns and the other corresponds to the categories of time interval of queries. Each neuron can get the value 1 through 7 according to its search pattern or time interval (Note that there are seven search pattern types and seven time intervals). The output layer has only one neuron, which can get the values 1 or 2, referring to a topic shift or continuation. The hidden layer has five neurons. The number of hidden layers and the number of neurons in each hidden layer are determined after a series of pilot experiments. The neural network applied in this study is a feedforward neural network, which is trained using backward propagation. See Figure 2, for the structure of the neural network.

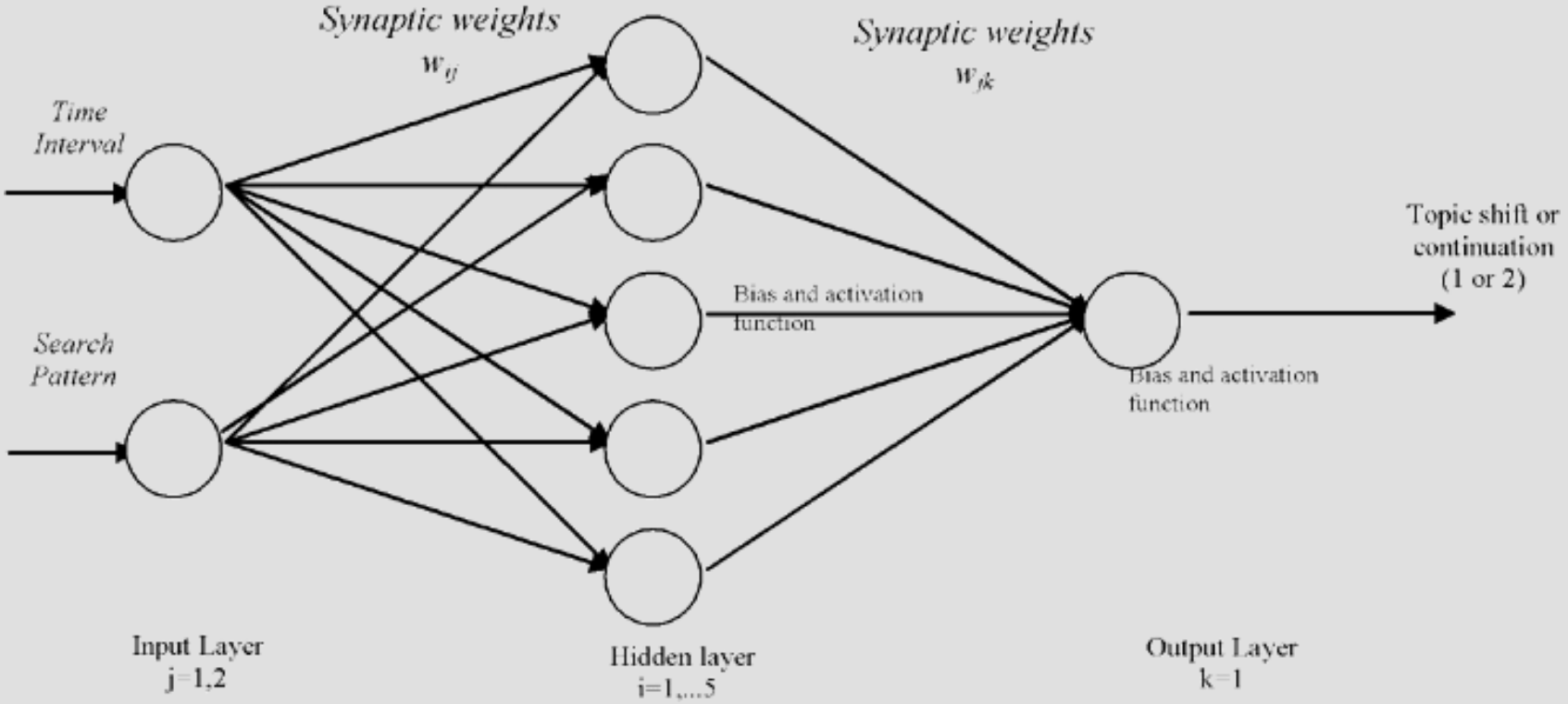

Figure 2: The structure of the proposed neural network

Training the neural network:

We obtain two neural networks by training the neural network with the first half of Excite and FAST datasets. See Table 1. (Excite dataset: 5014 queries or the first half of the 10,003 queries, FAST dataset: 4989 queries or the first half of the 10,007 queries). The values for the input and output layers are provided to the neural network, so that it can train itself. The values for the input layer are the search pattern and time interval of the query. The values for the output layer are the label of each query as topic shift or continuation. The neural network trains the weights so that the output layer yields the correct label (the topic shift or continuation) as much as possible, by minimizing the total error. We used the software MATLAB to create and train the neural network.

Applying the neural network to the test data sets:

Using the information from training, the neural network is used to identify topic changes in the second half of the datasets. The neural network trained on the Excite dataset is tested on the Excite and FAST datasets, and the neural network trained on the FAST dataset is tested on FAST and Excite datasets. The research design is in Table 4. The four cases are investigated to see, whether training the neural network on a dataset other than the one it is tested on affects the performance of the neural network.

Table 4: Research design for training and testing the neural network.

	Training dataset Excite (Neural Network A)	Training dataset FAST (Neural Network B)
Testing dataset Excite	Case 1	Case 4
Testing dataset FAST	Case 3	Case 2

The output layer of the neural network design yields a result between 1 and 2 depending on the input parameters. However, the result of the neural network should be either 1 (continuation) or 2 (shift). A threshold value is used to round any number between 1 and 2 to 1 or 2. Given there is no priority, it is reasonable to set the threshold value to 1.5 (any value over 1.5 is considered as 2, and under 1.5 is considered as 1). He *et al.*(2002), and Ozmutlu and Cavdur (in press,a) gave a greater priority to Type B errors by increasing βvalue in their fitness functions. To create similar priority effect in the results of the neural network application, we use a threshold value of 1.2 (any value over 1.2 is considered as 2, and under 1.2 is considered as 1), thus lowering the risk of Type B errors.

Comparison of results from human expert and the neural network:

The results of the neural network tested on the FAST and Excite datasets are compared to the actual topic shifts and identifications determined by the human expert. 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 neural network is evaluated in terms of precision (P) and recall (R). Higher P and R values mean higher success in topic identification.

Results and discussion

In this section, we present the results of the methodology described in the previous section. We present the results of the cases as shown in Table 4. We name the neural networks trained on the Excite and FAST datasets as neural network A and neural network B respectively.

Case 1: Neural network A trained with first half of the Excite dataset and tested with the second half of the Excite dataset

When the human expert evaluated the 10,003 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 training (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. The results of the evaluation of the human expert can be seen in Table 5.

After running the neural network, we obtain the results in Table 6. We observe that the neural network marked 3268 queries as topic continuation, whereas the human expert identified 3515 queries as topic continuation. Similarly, the neural network marked 399 queries as topic shifts, whereas the human expert identified 152 queries as topic shifts. During the topic identification process, we observed 283 Type A errors and 36 Type B errors.

Using the neural network approach, 116 out of 152 topic shifts are identified correctly, yielding an R_{shift} value of 0.76 and 3232 out of 3515 topic continuations are identified correctly, yielding an R_{contin} value of 0.92. These results show that most of the topic shifts and

continuations were estimated correctly by the neural network. On the other hand, the neural network yielded 399 topic shifts, when actually there are 152 topic shifts, giving a value of 0.29 for P_{shift}. This results means that the neural network overestimates the number of topic shifts. This result could be due to the assumption stated in the previous section. Since the previous studies gave greater weight to identifying topic shifts, we kept the threshold value in the neural network as 1.2, therefore increased the probability of erring on the preferred side, hence overestimating the number of topic shifts. Changing the threshold value of the neural network is subject to further study. In terms of topic continuations $P_{contin}$was 0.99, 3232 topic continuations out of 3268 topic continuations were estimated correctly, i.e. almost all, but 1%, of the topic continuations marked by the neural network were correct.

Table 5: Topic shifts and continuations in the entire Excite dataset as evaluated by the human expert

	Total number of queries	Number of sessions	Number of queries considered by the neural network	Total number of shifts marked by the human expert	Total number of continuations marked by the human expert
First half of dataset used for training	5014	1201	3813	269	3544
Second half of dataset used for testing	4989	1322	3667	152	3515
Entire dataset	10,003	2523	7480	421	7059

Table 6: Results of training the neural network on Excite and testing it on Excite

Origin of results	Total number of queries included in analysis	Number of topic shifts	Number of topic continuations	Correctly estimated number of shifts	Correctly estimated number of continuations	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
Results of neural network	3667	N_{shift} = 399	N_{contin} = 3268	$N_{shift\ \&\ correct}$ =116	$N_{contin\ \&\ correct}$ =3232	283	36	0.29	0.76	0.99	0.92
Results of human expert	3667	$N_{true\ shift}$ = 152	$N_{true\ contin}$ = 3515	----	----	----	----	----	----		

In this study, we prioritize Type B errors over Type A errors, since we apply the assumptions of the previous studies (He *et al.*(2002) and Ozmutlu and Cavdur (in press,a)) to avoid any bias in comparison of our methods with previous approaches.. The worth of Type A errors in terms of Type B errors is an important issue to discuss but it is left as future work.

<u>Case 2: Neural network B trained with first half of the FAST dataset and tested with the second half of the FAST dataset</u>

Out of 9044 queries, 8348 topic continuations (92.3%) and 696 topic shifts (7.7%) were found. In the subset used for training (first half of the dataset- 4997 queries), there were 437 user sessions, thus 4560 queries of the first half of the dataset are used for training the neural network. Out of 4560 queries, there are 4174 topic continuations (91.5%) and 386 topic shifts (8.5%). 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 (93.1%) were topic continuations, whereas 310 (6.9%) were topic shifts. The results of the evaluation of the human expert can be seen in Table 7.

After running the neural network on the second half of the dataset, we obtain the results in Table 8. For comparison, we also include the results on the second half of the dataset as evaluated by the human expert. We observe that the neural network marked 3619 queries as topic continuation, whereas the human expert identified 4174 queries as topic continuation. Similarly, the neural network

marked 865 queries as topic shifts, whereas the human expert identified 310 queries as topic shifts.

An important result in Table 8 is that the neural network identified all the topic changes except five, giving a value for Type B error of 5. This yields a R_{shift} value of 0.984, which is a very satisfactory result. In addition, 3614 topic continuations out of 4174 continuations were estimated correctly, yielding a R_{contin} value of 0.866. These results denote a high level of estimation of topic shifts and continuations. On the other hand, the neural network yielded 865 topic shifts, when actually there are 310 topic shifts, giving a value of 0.353 for P_{shift}. This results means that the neural network overestimates the number of topic shifts. This result could be due to the assumption stated in the previous section. In terms of topic continuations P_{contin} was 0.999, 3614 topic continuations out of 3619 topic continuations were estimated correctly, i.e. almost all topic continuations marked by the neural network were correct.

Table 7: Topic shifts and continuations in the entire FAST dataset as evaluated by the human expert

	Total number of queries	Number of sessions	Number of queries considered by the neural network	Total number of shifts marked by the human expert	Total number of continuations marked by the human expert
First half of dataset used for training	4997	437	4560	386	4174
Second half of dataset used for testing	5010	526	4484	310	4174
Entire dataset	10007	963	9044	696	8348

Table 8: Results of training the neural network on FAST and testing it on FAST

Origin of results	Total number of queries included in analysis	Number of topic shifts	Number of topic continuations	Correctly estimated number of shifts	Correctly estimated number of continuations	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
Results of neural network	4484	N_{shift} = 865	N_{contin} = 3619	$N_{shift\ \&\ correct}$ = 305	$N_{contin\ \&\ correct}$ = 3614	560	5	0.353	0.98	0.999	0.866
Results of human expert	4484	$N_{true\ shift}$ = 310	$N_{true\ contin}$ = 4174	-----	-----	----	----	----	----	----	----

Case 3: Neural network A trained with first half of the Excite dataset and tested with the second half of the FAST dataset

The number of topic shifts and continuations as evaluated by the human expert are given in Tables 5 and 7 for Excite and FAST datasets, respectively. After training the neural network with the first half of the Excite dataset and running it on the second half of the FAST dataset, we obtain the results in Table 9. For comparison, we also include the results on the second half of the dataset as evaluated by the human expert. We observe that the neural network marked 3953 queries as topic continuation, whereas the human expert identified 4174 queries as topic continuation. Similarly, the neural network marked 531 queries as topic shifts, whereas the human expert identified 310 queries as topic shifts. 3953 topic continuations out of 4174 continuations were estimated correctly, yielding a R_{contin}value of 0.93. 226 topic shifts out of 310 were also estimated correctly, giving an R_{shift}value of 0.73. As with the previous cases, these results show that there is a high level of estimation of topic shifts and continuations. However, again as with the previous cases the neural network overestimates the number of topic shifts (531 instead of 310). The potential reason for this result was explained in the previous paragraphs. In terms of topic continuations $P_{contin}$was 0.98, 3869 topic continuations out of 3953 topic continuations were estimated correctly, i.e. almost all the topic continuations marked by the neural network were correct.

Case 4: Neural network B trained with first half of the FAST dataset and tested with the second half of the Excite dataset

The number of topic shifts and continuations as evaluated by the human expert are given in Tables 5 and 7 for Excite and FAST datasets, respectively. After training the neural network with the first half of the FAST dataset and running it on the second half of the Excite dataset, we obtain the results in Table 10. For comparison, we also include the results on the second half of the dataset as evaluated by the human expert.

By comparing the results in Table 6 and Table 10, we observe an interesting finding. The results in Table 10 are same as the results in Table 6. Even though neural network A is trained on the Excite dataset in Case 1, and neural network B is tested on the FAST dataset in Case 4, their application on the second half of the Excite dataset yields identical results. Also considering the similarity of the results of Cases 2 and Case 3, there is an indication that no matter the training dataset for the neural network, the application results of the neural network are successful. These results suggest that the estimation power of topic shifts and continuations by the neural network is independent of the training dataset for the neural network. This finding is also indicating that the multitasking behavior of Excite and FAST users could be similar. Based on these indications, further studies should be performed to see whether there are common characteristics of multitasking behavior of Web users, regardless of the search engine they are using. However, in order to support these indications, more replications on different search engine data logs are necessary

Table 9: Results of training the neural network on Excite and testing it on FAST

Origin of results	Total number of queries included in analysis	Number of topic shifts	Number of topic continuations	Correctly estimated number of shifts	Correctly estimated number of continuations	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
Results of neural network	4484	N_{shift} = 531	N_{contin} = 3953	$N_{shift\ \&\ correct}$ =226	$N_{contin\ \&\ correct}$ =3869	305	84	0.43	0.73	0.98	0.93
Results of human expert	4484	$N_{true\ shift}$ = 310	$N_{true\ contin}$ = 4174	-----	-----	----	----	----	----	----	----

Table 10: Results of training the neural network on FAST and testing it on Excite

Origin of results	Total number of queries included in analysis	Number of topic shifts	Number of topic continuations	Correctly estimated number of shifts	Correctly estimated number of continuations	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
Results of neural network	3667	N_{shift} = 399	N_{contin} = 3268	$N_{shift\ \&\ correct}$ =116	$N_{contin\ \&\ correct}$ =3232	283	36	0.29	0.76	0.99	0.92
Results of human expert	3667	$N_{true\ shift}$ = 152	$N_{true\ contin}$ = 3515	-----	-----	----	----	----	----	----	----

Conclusion

This study shows that neural networks can be successfully used to automatically identify topic changes in the search engine query logs, even when the neural network is tested on a dataset other than the one it is trained on. The search query logs used in this study comes from two search engines; the Excite and FAST search engines. Samples of approximately 10,000 queries were selected from both datasets. Two neural networks were trained with approximately half the data sets. The neural network trained on the Excite dataset was tested on both the Excite and FAST datasets, and the neural network trained on the FAST dataset was tested on both the Excite and FAST datasets. The results were compared to those of a human expert.

In all the cases considered, topic shifts and continuations were estimated successfully, yielding high recall values for topic shifts and continuations. As a result of this study, we conclude that neural networks are successful in automatic identification of topic shifts and continuations in search engine data logs. However, the number of topic shifts was overestimated due to the structure of the neural network used in the study. Future work includes developing more enhanced and refined neural network structures and determining which network structure is the most successful in automatic topic identification.

In addition, there is an indication that no matter the training dataset for the neural network, the application results of the neural network are successful. These results suggest that the estimation power of topic shifts and continuations by the neural network is independent of the training dataset for the neural network. Based on these indications, further studies should be performed to see whether there are common characteristics of multitasking behavior of Web users. However, in order to support these indications, more replications on different search engine data logs are necessary.

References

Beeferman, D. and Berger, A. (2000), Agglomerative clustering of a search engine query log. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA (pp. 407 - 416).

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D. and Frieder, O. (2004). Efficiency and Scaling: Hourly Analysis of a Very Large Topically Categorized Web Query Log. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval*, Sheffield, UK (pp. 321-328).

He. D., Goker, A. and Harper, D.J. (2002). Combining evidence for automatic Web session identification, *Information Processing and Management, 38*(5), 727-742.

Jansen, B.J., Spink A. and Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web, *Information Processing and Management, 36*, (pp.207-227).

Miwa. (2001). User situations and multiple levels of users goals in information problem solving processes of AskERIC users. In *Proceedings of the 2001 Annual Meeting of the American Society for Information Sciences and Technology*, 38, (pp. 355-371).

Muresan, G. and Harper, D.J. (2004). Topic Modeling for Mediated Access to Very Large Document Collections", *Journal of the American Society for Information Science and Technology, 55*(10), 892–910.

Ozmutlu, H.C. and Cavdur, F. (in press, a). Application of automatic topic identification on excite web search engine data logs, *Information Processing and Management*

Ozmutlu, S and Cavdur, F. (in press, b). Neural Network Applications for Automatic New Topic Identification, *Online Information Review*.

Ozmutlu, H.C., Cavdur, F., Ozmutlu, S. and Spink, A., (2004a). Neural Network Applications for Automatic New Topic Identification on Excite Web search engine datalogs, In *Proceedings of ASIST 2004, Annual Meeting of the American Society for Information Science and Technology*, Providence, RI, (pp. 310-316).

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

Ozmutlu, S., Ozmutlu, H.C. and Spink, A. (2003). Multitasking Web searching and implications for design, *In Proceedings of ASIST 2003, Annual Meeting of the American Society for Information Science and Technology*, Long Beach, CA, (pp. 416-421).

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

Pu, H.T., Chuang, Shui-Lung &Yang, C. (2002). Subject Categorization of Query Terms for Exploring Web Users'Search Interests, *Journal of the American Society for Information Science and Technology, 53*(8), 617–630.

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

Silverstein, C., Henzinger, M., Marais, H., &Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum, 33*(1), 6-12.

Spink, A., Bateman, J., &Jansen, B.J. (1999). Searching Heterogeneous Collections on the Web: A survey of Excite users. *Internet Research: Electronic Networking Applications and Policy, 9*(2): 117-128.

Spink, A., Wolfram, D., Jansen, B.J., &Saracevic, T. (2001). Searching the Web: The public and their queries, *Journal of the American Society for Information Science and Technology, 53*(2), 226–234.

Spink, A., Jansen, B. J., Wolfram, D., &Saracevic, T. (2002a). From e-sex to e-commerce: Web search changes, *IEEE Computer, 35*(3), 133-135.

Spink, A., Ozmutlu, H. C., &Ozmutlu, S. (2002b). Multitasking information seeking and searching processes, *Journal of the American Society for Information Science and Technology, 53*(8), 639-652.

Wen, J.R., Nie, J.Y. and Zhang, H.J. (2002). Query Clustering Using User Logs, *ACM Transactions on Information Systems, 20*(1), 59–81.