Investigating the Performance of Automatic New Topic Identification Across Multiple Datasets [1]

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

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

Recent studies on automatic new topic identification in Web search engine user sessions demonstrated that neural networks are successful in automatic new topic identification. However most of this work applied their new topic identification algorithms on data logs from a single search engine. In this study, we investigate whether the application of neural networks for automatic new topic identification are more successful on some search engines than others. 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 query logs with more topic shifts tend to provide more successful results on shift-based performance measures, whereas logs with more topic continuations tend to provide better results on continuation-based performance measures.

Introduction and related research

An important facet of Web mining is to study the behavior of search engine users. One dimension of search engine user profiling 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. One of the main elements in following user topics is new topic identification, which is discovering when the user has switched from one topic to another during a single search session. If the search engine is aware that the user's new query is on the same topic as the previous query, the search engine could provide the results from the document cluster relevant to the previous query, or alternatively, if the user is on a new topic, the search engine could resort to searching other document clusters. Consequently, search engines can decrease the time and effort required to process the query. 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).

Many researchers worked on large scaled studies on search engine datalogs, such as Silverstein et al.(1999), Cooley, Mobasher, & Srivastava (1999), Spink, et. al., (2000, 2001, 2002a), Ozmutlu, et al. (2002b, 2003b, 2003c) and Ozmultu and Spink (2002). There are few studies on query clustering and new topic identification, and the studies generally analyzed the queries semantically. Silverstein, et al.(1999), Jansen, et al.(2000), and Spink, et al. (2001) have performed content analysis of search engine data logs at the term level, and. Jansen, et al. (2000) and Spink, et al. (2001,2002a) at the conceptual or topical level. Ozmutlu, et al. (2004b) and Beitzel, et al. (2004) have done hourly statistical and topical analysis search engine query logs.

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.

Another dimension of topic related Web searching is multitasking, which 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" in terms of information retrieval (Spink, et al, 2002b). Spink, et al. (2002b) and (Ozmutlu, et al, 2003a) have found that 11.4%-31.8% of search engine users

performed multitasking searches.

Most query clustering methods are focused on interpretation of keywords or understanding the topic or the contents of the query, which complicates the process of query clustering and increases the potential noise of the results of the study. One of the promising approaches 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 (2006) applied multiple factor regression to automatically identify topic changes, and showed that there is a valid relationship between non-semantic characteristics of user queries and topic shifts and continuations. Ozmutlu. (2006) 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 provided 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. He, et al. (2002) proposed a topic identification algorithm that uses Dempster-Shafer Theory (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. Their approach was replicated on Excite search engine data (Ozmutlu and Cavdur, 2005a). Ozmutlu and Cavdur (2005b) and Ozmutlu, et al. (2004a) proposed an artificial neural network to automatically identify topic changes, and showed that neural networks successfully provided new topic identification. Application of neural networks for automatic new topic identification does not contain semantic analysis, and relies on the statistical characteristics of the queries.

In this study, we aim to apply the neural network to multiple datasets, and investigate whether the application of neural networks for automatic new topic identification are more successful on some search engines than others. We also aim to see whether there are specific conditions, which affect the performance of neural networks in terms of automatic new topic identification. To conduct this study, neural networks are tested on separate datasets.

In the next section, we provide a detailed discussion of the experimental framework and application of neural networks for automatic new topic identification. Finally, we provide results of the proposed methodology and discussion of the results, and conclude the study.

Methodology

The datasets

The first search query log used in this study comes from the Excite search engine, was

collected on December 20, 1999, and consists of 1,025,910 search queries. The first 10,003 queries of the dataset were selected as a sample. The sample 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, and contains a query log of 1,257,891 queries. Queries were collected on February 6, 2001. We selected a sample of 10,007 queries by using Poisson sampling (Ozmutlu, et al., 2002a) to provide a sample dataset that is both representative of the data set and small enough to be analyzed conveniently.

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:

$N_{true\ shift} = N_{shift\ \&\ correct} + Type\ B\ error$
$N_{true\ contin} = N_{contin\ \&\ correct} + Type\ A\ error$
$N_{shift} = N_{shift\ \&\ correct} + Type\ A\ error$
$N_{contin} = N_{contin\ \&\ correct} + Type\ B\ error$

The commonly used performance measures of Precision (*P*) and Recall (*R*) are used in this study to demonstrate the performance of the proposed neural network. The focus of *P* and *R* are both on correctly estimating the number of topic shifts and continuations. The formulation of these measures are as follows:

$$P_{shift} = \frac{N_{shift\&correct}}{N_{shift}} \quad (1)$$

$$P_{contin} = \frac{N_{contin\&correct}}{N_{contin}} \quad (2)$$

$$R_{shift} = \frac{N_{shift\&correct}}{N_{trueshift}} \quad (3)$$

$$R_{contin} = \frac{N_{contin\&correct}}{N_{truecontin}} \quad (4)$$

Research Design

The main research question in this study is whether automatic new topic identification is equally successfully performed on different datasets. To answer this research question, neural networks need to be tested on several datasets. For testing, the research design in Table 1 is proposed. For application of neural networks for automatic new topic identification, datasets are divided to almost two equal parts, and the neural network is trained on the first half of the datasets. Then, using the information from training, the neural network is used to identify topic changes in the second half of the datasets. In this study, we initially train and test the neural network on Excite and FAST datalogs. Additionally, 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 reason for a cross-application of the neural networks is that we would like to observe any effects that might come from the training dataset.

Table 1: Research design

	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

Proposed Algorithm

The steps of the application of neural networks 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.
- *Divide 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 2.

Table 2: 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 NN	5014 queries	4997 queries
2nd half of the sample set used for training the NN	4989queries	5010 queries

- *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 (Ozmutlu and Cavdur, 2005a, 2005b), Ozmutlu (2006), Ozmutlu, et al. (2004a) to avoid any bias during comparison.

We use seven categories of time intervals for a query: 0-5 min., 5-10 min., 10-15 min., 15-20 min., 20-25 min., 25-30 min., 30+ min. See Table 3 for distribution of the queries with respect to time interval. It should be noted that not all of 5014 queries in Excite 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, 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 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.

Table 3: Distribution of time interval of queries

Time Interval (min)	Excite continuations	Excite shifts	FAST continuations	FAST Shifts
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

We also use seven categories of search patterns in this study, which are as follows:

- 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.

For details on the search patterns, see Ozmutlu and Cavdur (2005a, 2005b), Ozmutlu (2006), Ozmutlu, et al. (2004a). 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 found in Figure 1. See Table 4 for distribution of queries with respect to search patterns in the training datasets.

Table 4: 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
Relev. Feed.	0	0	70	2
Other	0	0	2	2
Total	3544	269	4174	386

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

- *Forming the neural network*: In this study, we propose a feedforward 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.

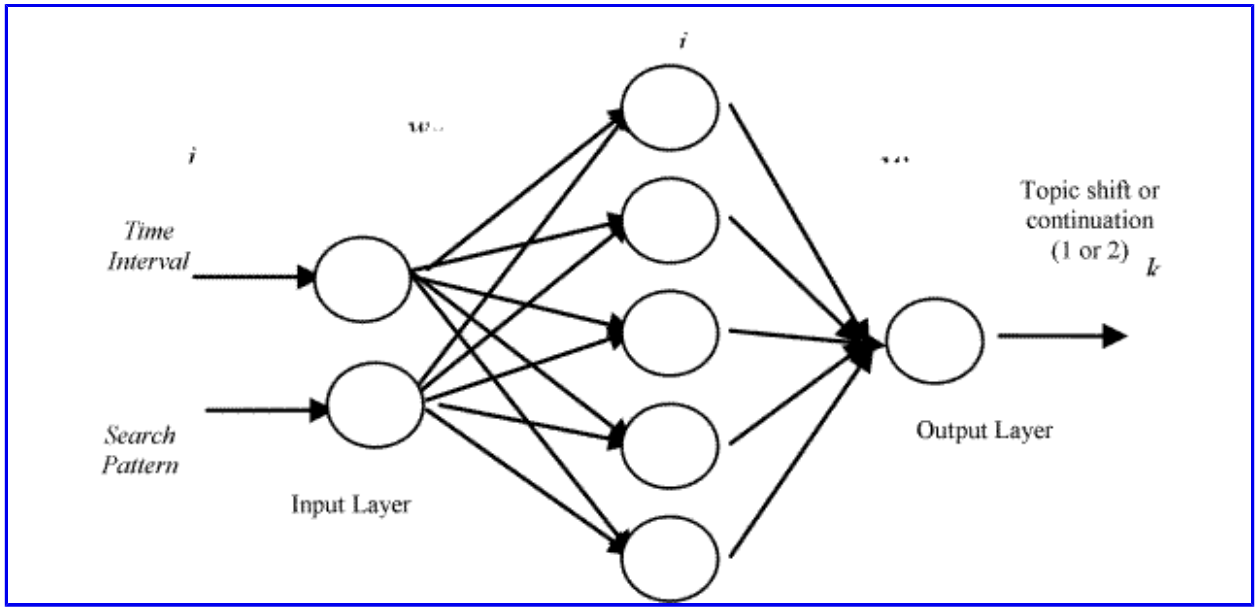

Figure 2: The structure of the proposed neural network

- *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. The values for the input layer, i.e the search pattern and time interval of the query, and output layer, i.e. the label of each query as topic shift or continuation, are provided to the neural network, so that it can train itself. The neural network trains the weights so that the output layer yields the correct label (the topic shift or continuation) as much as possible. 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. To be statistically reliable, each case is repeated 50 times. The output layer of the neural network design yields a result between 1 and 2 depending on the input parameters. To conform with previous studies, we use a threshold value of 1.3. Any value over 1.3 is considered as 2 (shift), and under 1.3 is considered as 1(continuation).
- *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.
- *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.

3. Results and Discussion

In this section, we present the results of the methodology described in the previous section

and provide the discussion of the results. We present the results of the cases in Table 1.

- *Case 1: Neural network A trained with the Excite dataset and tested with 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 structure of the dataset in terms of number of sessions and queries included in the analysis can be seen in Table 5.

To be statistically reliable 50 replications of the neural network is made. The results of the first 10 runs of the 50 runs of the neural network is seen in Table 6. All the results could not be provided due to space considerations.The results are also seen in Figures 3a, 4a, 5a and 6a. The explanation of the figures can be provided as follows: For example in Run 1 in Table 6, we observed that the neural network marked 3375 queries as topic continuation, whereas the human expert identified 3515 queries as topic continuation. Similarly, the neural network marked 292 queries as topic shifts, whereas the human expert identified 152 queries as topic shifts. 86 out of 152 topic shifts are identified correctly, yielding an R_{shift} value of 0.57 for Run 1 (Fig. 4a) and 3309 out of 3515 topic continuations are identified correctly, yielding an R_{contin} value of 0.94 (Fig. 6a). For the first run, these results show that topic shifts and continuations were estimated somewhat correctly by the neural network. On the other hand, the neural network yielded 292 topic shifts, when actually there are 152 topic shifts, giving a value of 0.30 for P_{shift} (Fig. 3a). This results means that the neural network overestimates the number of topic shifts. Since the previous studies gave greater weight to identifying topic shifts, we kept the threshold value in the neural network as 1.3, 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 0.98 (Fig. 5a), 3309 topic continuations out of 3375 topic continuations were estimated correctly, i.e. almost all, but 2%, of the topic continuations marked by the neural network were correct.

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 by the neural network	Total no. of shifts marked by the human expert	Total no. of continuations marked by the human expert

1st half of dataset used for training	5014-Excite	1201-Excite	3813-Excite	269-Excite	3544-Excite
	4997 - Fast	437-Fast	4560-Fast	386-Fast	4174-Fast
2nd half of dataset used for testing	4989-Excite	1322-Excite	3667-Excite	152-Excite	3515-Excite
	5010-Fast	526-Fast	4484-Fast	310-Fast	4174-Fast
Entire dataset	10003-Excite	2523-Excite	7480-Excite	421-Excite	7059-Excite
	10007-Fast	963-Fast	9044-Fast	696-Fast	8348-Fast

Table 6: Results of training the neural network on Excite and testing it on Excite - Case 1

Origin of results	**Total number of queries included in analysis**	**Number of topic shifts**	**Number of topic continuations**	**Correctly estimated no. of.shifts**	**Correctly estimated no. of continuations**	**Type A error**	**Type B error**	***Pshift***	***Rshift***	***Pcontin***	***Rcontin***
Human expert	3667	*Ntrue shift* = 152	*Ntrue contin* = 3515				—		—		
Neural Network	3667	*Nshift*	*Ncontin*	*Nshift & correct*	*Ncontin & correct*	Type A error	Type B error	*Pshift t*	*Rshift*	*Pcontin*	*Rcontir*
NN-Run 1	3667	292	3375	86	3309	206	66	0,295	0,566	0,980	0,941
NN- Run 2	3667	314	3353	92	3293	222	60	0,293	0,605	0,982	0,937
NN- Run 3	3667	292	3375	86	3309	206	66	0,295	0,566	0,980	0,941
NN- Run 4	3667	360	3307	103	3258	257	49	0,286	0,678	0,985	0,927
NN- Run 5	3667	444	3223	117	3188	327	35	0,264	0,770	0,989	0,907
NN- Run 6	3667	312	3355	92	3295	220	60	0,295	0,605	0,982	0,937
NN- Run 7	3667	302	3365	92	3305	210	60	0,305	0,605	0,982	0,940

NN- Run 8	3667	285	3382	86	3316	199	66	0,302	0,566	0,980	0,943
NN- Run 9	3667	446	3221	117	3186	329	35	0,262	0,770	0,989	0,906
NN- Run 10	3667	292	3375	86	3309	206	66	0,295	0,566	0,980	0,941

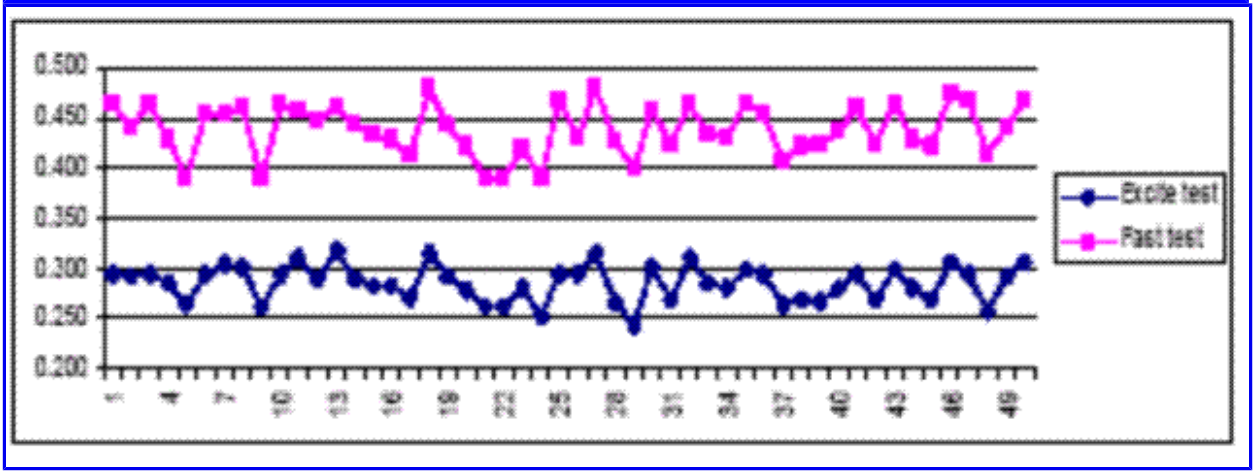

Figure 3a P_{shift} when Excite is the training dataset

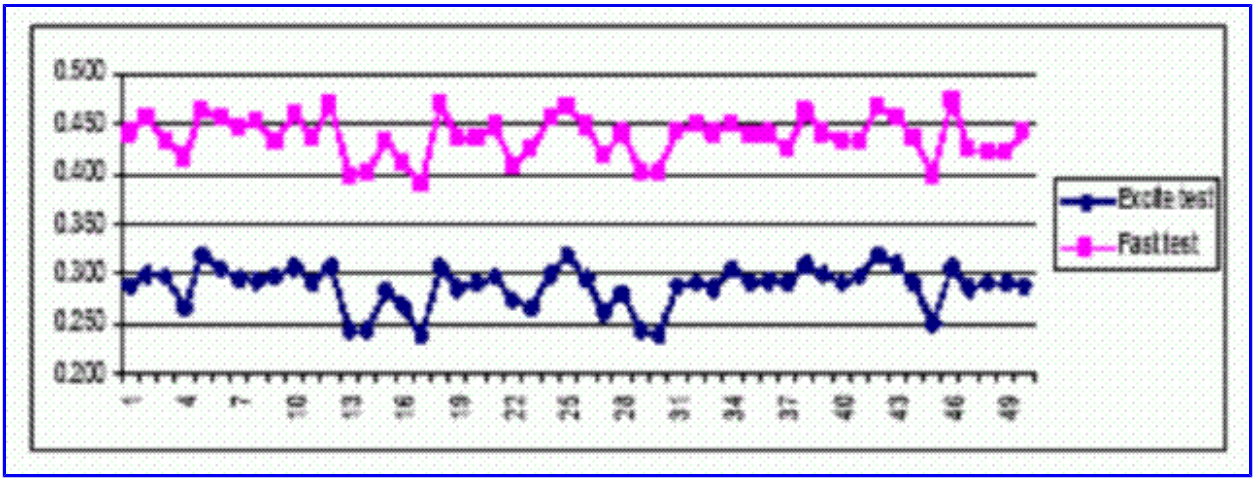

Figure 3b: P_{shift} when FAST is the training dataset

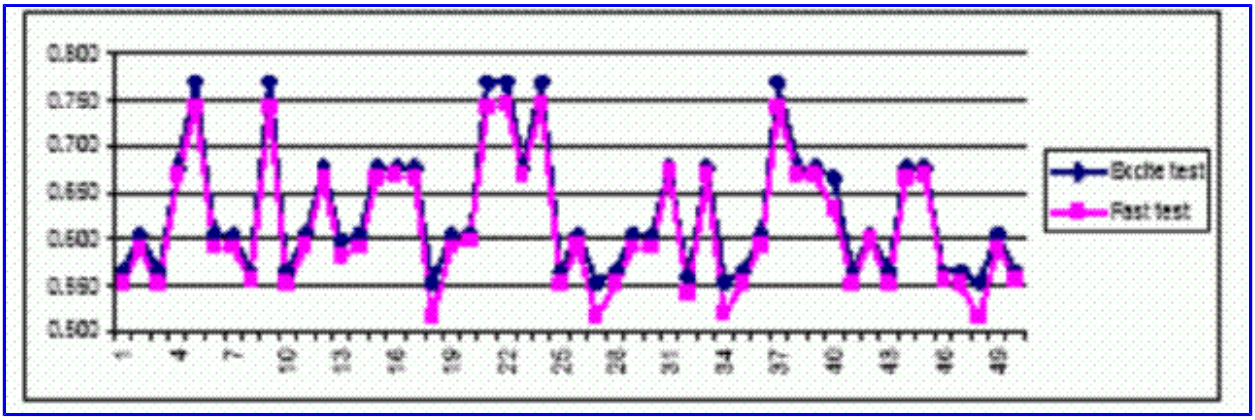

Figure 4a: R_{shift} when Excite is the training dataset

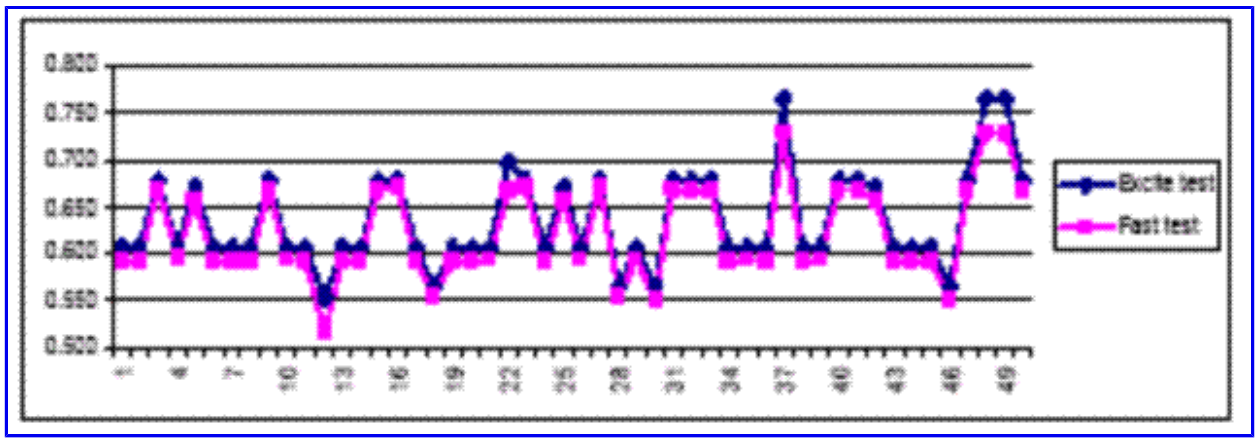

Figure 4b: R_{shift} when FAST is the training dataset

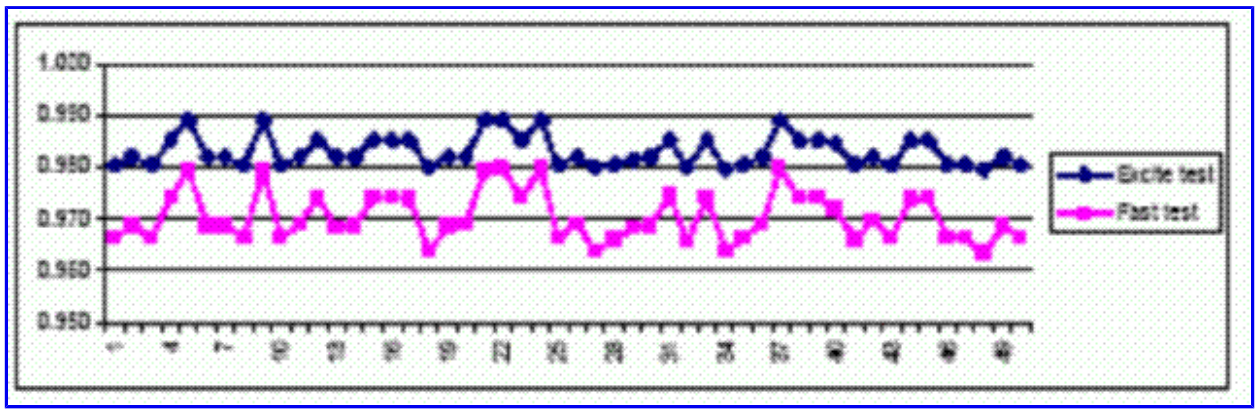

Figure 5a: P_{contin} when Excite is the training dataset

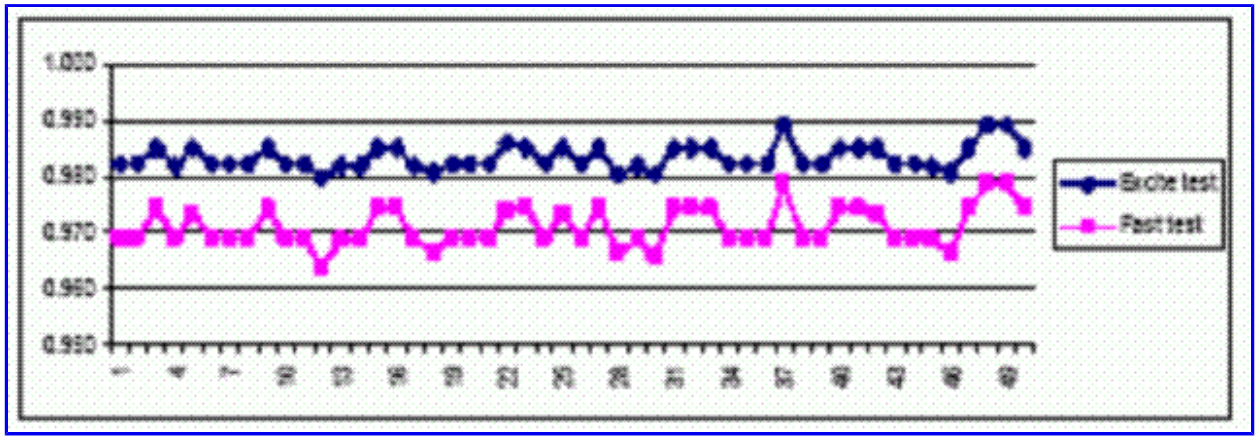

Figure 5b: P_{contin} when FAST is the training dataset

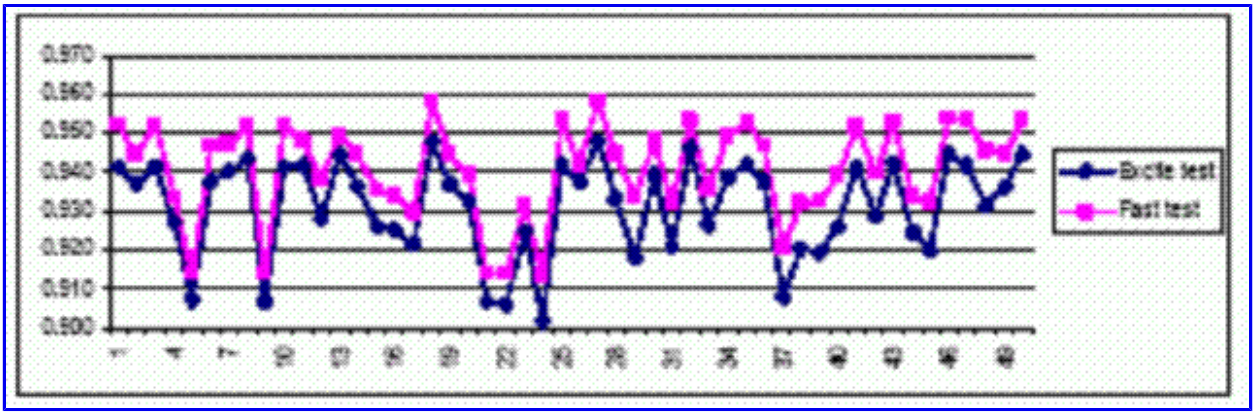

Figure 6a: R_{contin} when Excite is the training dataset

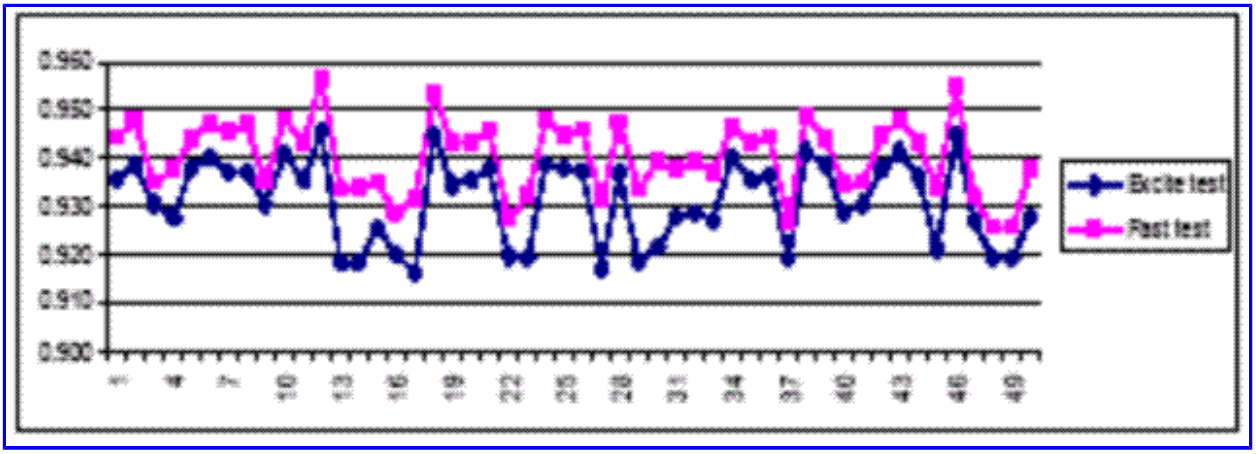

Figure 6b: R_{contin} when FAST is the training dataset

- *Case 2: Neural network B trained with the FAST dataset and tested with the FAST dataset:*

Out of 9044 queries, 8348 topic continuations and 696 topic shifts were found. In the subset used for training (4997 queries), there were 427 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 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. The results of the evaluation of the human expert can be seen in Table 5. 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 7. The results of the first 20 runs of the 10 runs of the neural network is seen in Table 7. All the results could not be provided due to space considerations. In Run 1, we observe that the neural network marked 4069 queries as topic continuation, whereas the human expert identified 4174 queries as topic continuation. Similarly, the neural network marked 415 queries as topic shifts, whereas the human expert identified 310 queries as topic shifts, yielding an R_{shift} value of 0.59 (Fig. 4b). In addition, 3942 topic continuations out of 4174

continuations were estimated correctly, yielding a R_{contin} value of 0.944 (Fig. 6b). These results denote a high level of estimation of topic shifts and continuations. On the other hand, the neural network yielded 415 topic shifts, when actually there are 310 topic shifts, giving a value of 0,441 for P_{shift} (Fig. 3b). 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, i.e. giving greater weight to identifying topic shifts. In terms of topic continuations P_{contin} was 0,969 (Fig. 5b) , 3942 topic continuations out of 4069 topic continuations were estimated correctly, i.e. almost all topic continuations marked by the neural network were correct.

Table 7: Results of training the neural network on FAST and testing it on FAST - Case 2

Origin of results	Total number of queries included in analysis	Number of topic shifts	Number of topic continuations	Correctly estimated no. of shifts	Correctly estimated no.of continuations	Type A error	Type B error	P_{shift}	R_{shift}	P_{contIn}	R_{contIn}
Human expert	4484	*Ntrue shift* = 310	*Ntrue contin* = 4174				—		—		
Neural Network	4484	*Nshift*	*Ncontin*	*Nshift & correct*	*Ncontin & correct*	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
NN-Run 1	4484	415	4069	183	3942	232	127	0,441	0,590	0,969	0,944
NN- Run 2	4484	401	4083	183	3956	218	127	0,456	0,590	0,969	0,948
NN- Run 3	4484	477	4007	207	3904	270	103	0,434	0,668	0,974	0,935
NN- Run 4	4484	445	4039	185	3914	260	125	0,416	0,597	0,969	0,938
NN- Run 5	4484	438	4046	203	3939	235	107	0,464	0,655	0,974	0,944
NN- Run 6	4484	403	4081	183	3954	220	127	0,454	0,590	0,969	0,947
NN- Run 7	4484	411	4073	183	3946	228	127	0,445	0,590	0,969	0,945
NN- Run 8	4484	404	4080	183	3953	221	127	0,453	0,590	0,969	0,947

NN- Run 9	4484	477	4007	207	3904	270	103	0,434	0,668	0,974	0,935
NN- Run 10	4484	400	4084	184	3958	216	126	0,460	0,594	0,969	0,948

- *Case 3: Neural network A trained with the Excite dataset and tested with the FAST dataset:*

The number of topic shifts and continuations as evaluated by the human expert and the structure of the datasets are given in Table 5. 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 8. The results of the first 10 runs of the 50 runs of the neural network are seen in Table 8. All the results could not be provided due to space considerations. For comparison, we also include the results on the second half of the dataset as evaluated by the human expert. In run 1, we observe that the neural network marked 4114 queries as topic continuation, whereas the human expert identified 4174 queries as topic continuation. Similarly, the neural network marked 370 queries as topic shifts, whereas the human expert identified 310 queries as topic shifts. 3975 topic continuations out of 4174 continuations were estimated correctly, yielding a R_{contin} value of 0.952 (Fig. 6a). 171 topic shifts out of 310 were also estimated correctly, giving an R_{shift} value of 0.552 (Fig. 4a). The neural network overestimates the number of topic shifts (370 instead of 310). The potential reason for this result was explained in the previous paragraphs. In terms of topic continuations P_{contin} was 0.966 (Fig. 5a), 3975 topic continuations out of 4114 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 the FAST dataset and tested with the Excite dataset:*

The number of topic shifts and continuations as evaluated by the human expert are given in Table 5. 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 9. The results of the first 10 runs of the 50 runs of the neural network are seen in Table 9. All the results could not be provided due to space considerations. For comparison, we also include the results on the second half of the dataset as evaluated by the human expert. The results in Run 1 are discussed as follows: We observe that the neural network marked 3348 queries as topic continuation, whereas the human expert identified 3515 queries as topic continuation. Similarly, the neural network marked 319 queries as topic shifts, whereas the human expert identified 152 queries as topic shifts. During the topic identification process, we observed 227 Type A errors and 60 Type B errors. Using the neural network approach, in Run 1, 92 out of 152 topic shifts are identified correctly, yielding an R_{shift} value of 0.605 (Fig. 4b) and

3288 out of 3515 topic continuations are identified correctly, yielding an R_{contin} value of 0.935 (Fig. 6b). For the first run, these results show that of the topic shifts and continuations were estimated somewhat correctly by the neural network. On the other hand, the neural network yielded 319 topic shifts, when actually there are 152 topic shifts, giving a value of 0.29 for P_{shift} (Fig. 3b). 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 sections, i.e. giving greater weight to identifying topic shifts by using a threshold value of 1.3 in the neural network. In terms of topic continuations P_{contin} was 0.982 (Fig. 5b), 3288 topic continuations out of 3348 topic continuations were estimated correctly, i.e. almost all, but 2%, of the topic continuations marked by the neural network were correct.

Discussion

In Figures 3, 4, 5 and 6, we see the effects of the training datasets on P_{shift}, R_{shift}, P_{contin}, and R_{contin}, respectively. Figure 3 shows that regardless of the training dataset, the FAST dataset seems to yield better values of P_{shift}. Figures 4 and 6 show that, in terms of R_{shift} and R_{contin}, both datasets are equally successful. Figure 5 demonstrates that the Excite dataset bears better results in terms of P_{contin} compared to the FAST dataset. Consequently, this paper's findings might indicate that the application of neural networks on different search engines does not provide the same results.

Generally, the FAST dataset tends to produce better results on shift based measures, whereas the Excite dataset tends to yield more favorable results in terms of continuation based measures. During the presentation of the results for the cases, we noted that the neural network usually overestimated the number of topic shifts. This result is probably due to keeping the threshold value in the neural network as 1.3, to be consistent with previous studies. The previous studies gave priority to identifying topic shifts, and the probability of erring on the preferred side increases when the threshold is set to 1.3. This choice causes overestimating the number of topic shifts. Since the neural networks overestimate the number of topic shifts, the dataset which has more topic shifts, would be expected to be more successful in terms of shift based measures. The FAST dataset has more topic shifts compared to the Excite dataset, as seen in Table 5. Similarly, it can be deduced that the Excite dataset is more successful in terms of continuation based performance measures, since it has more topic continuations. Since the neural network seems to be biased towards identifying topic shifts, it would be expected to modify the parameters of the neural network to remove the bias. The parameter, which might cause the shift-directed bias is the threshold of the neural network. Testing neural networks with different threshold values is a subject of further research.

Table 8: Results of training the neural network on Excite and testing it on FAST- Case 3

Origin of results	Total no. of queries included in analysis	Number of topic shifts	Number of topic continuations	Correctly estimated no. of shifts	Correctly estimated no. of contin.s	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
Human expert	4484	$N_{true\ shift}$ = 310	$N_{true\ contin}$ = 4174				—		—		
Neural Network	4484	*Nshift*	*Ncontin*	*Nshift & correct*	*Ncontin & correct*	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
NN-Run 1	4484	370	4114	171	3975	199	139	0,462	0,552	0,966	0,952
NN- Run 2	4484	415	4069	183	3942	232	127	0,441	0,590	0,969	0,944
NN- Run 3	4484	370	4114	171	3975	199	139	0,462	0,552	0,966	0,952
NN- Run 4	4484	483	4001	207	3898	276	103	0,429	0,668	0,974	0,934
NN- Run 5	4484	591	3893	230	3813	361	80	0,389	0,742	0,979	0,914
NN- Run 6	4484	404	4080	183	3953	221	127	0,453	0,590	0,969	0,947
NN- Run 7	4484	403	4081	183	3954	220	127	0,454	0,590	0,969	0,947
NN- Run 8	4484	373	4111	172	3973	201	138	0,461	0,555	0,966	0,952
NN- Run 9	4484	591	3893	230	3813	361	80	0,389	0,742	0,979	0,914
NN- Run 10	4484	370	4114	171	3975	199	139	0,462	0,552	0,966	0,952

Table 9: Results of training the neural network on FAST and testing it on Excite - Case 4

Origin of results	Total no. of queries included in analysis	Number of topic shifts	Number of topic continuations	Correctly estimated no. of shifts	Correctly estimated no. of contin.s	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}

Human expert	3667	*Ntrue shift* = 152	*Ntrue contin* = 3515				—		—		
Neural Network	3667	*Nshift*	*Ncontin*	*Nshift & correct*	*Ncontin & correct*	Type A error	Type B error	P_{shift}	R_{shift}	P_{contin}	R_{contin}
NN-Run 1	3667	319	3348	92	3288	227	60	0,288	0,605	0,982	0,935
NN- Run 2	3667	306	3361	92	3301	214	60	0,301	0,605	0,982	0,939
NN- Run 3	3667	347	3320	103	3271	244	49	0,297	0,678	0,985	0,931
NN- Run 4	3667	346	3321	92	3261	254	60	0,266	0,605	0,982	0,928
NN- Run 5	3667	320	3347	102	3297	218	50	0,319	0,671	0,985	0,938
NN- Run 6	3667	302	3365	92	3305	210	60	0,305	0,605	0,982	0,940
NN- Run 7	3667	311	3356	92	3296	219	60	0,296	0,605	0,982	0,938
NN- Run 8	3667	314	3353	92	3293	222	60	0,293	0,605	0,982	0,937
NN- Run 9	3667	347	3320	103	3271	244	49	0,297	0,678	0,985	0,931
NN- Run 10	3667	299	3368	92	3308	207	60	0,308	0,605	0,982	0,941

An additional result that the figures support is that the choice of training dataset did not affect the performance of automatic new topic identification. In all the figures, had the training dataset been effective on the performance of the neural network, the line corresponding to the training network would have shown superior results. Even though the training dataset is different in Figures 3a and 3b, FAST did better in terms of P_{shift}. Had the training dataset been effective, Excite should have done better, when Excite was the training dataset, and FAST should have done better when FAST was the training dataset. However, this is not the case. The same comment applies to all the performance measures covered in Figures 3 through 6.

Conclusion

This study shows that neural networks can be successfully applied for automatic new topic identification. 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. However, the performance of the neural network changed with respect to the performance measure and the test dataset that is used. Shift-based performance measures tend to have better values with datasets having more shifts and continuation-based performance measures tend to acquire better values with datasets having more continuations. To have a more consistent performance of automatic new topic identification with neural network, enhancing and refinement of the neural network structure and parameters could be necessary, such as changing threshold values of the neural network.

The findings of this study also indicate that the estimation power of the neural network is independent of the training dataset for the neural network. Conclusively, no matter which training dataset is used, the application results of the neural network were successful. Based on these indications, further studies should be performed with more datasets to validate these findings.

Notes

[1] 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". *Back*

References

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

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

Cooley, R., Mobasher, B., & Srivastava, J. (1999) Data preparation for mining world wide web browsing patterns *Knowledge and Information Systems* 1, 5-32

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

Jansen, B.J., Spink, A. & 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, 207-227

Muresan, G. & 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), pp. 892-910

Ozmultu, H.C. & Spink, A., (2002) Characteristics of question format web queries: an exploratory study *Information Processing & Management* 38, 453-471

Ozmutlu, S. (2006) Automatic new topic identification using multiple linear regression *Information Processing and Management* 42, 934-950

Ozmutlu, H.C. & Cavdur, F. (2005a) Application of automatic topic identification on excite web search engine data logs *Information Processing and Management* 41(5), 1243-1262

Ozmutlu, H.C. & Cavdur, F. (2005b) Neural network applications for automatic new topic identification *Online Information Review* 29, 35-53

Ozmutlu, H.C., Cavdur, F., Ozmutlu, S. & Spink, A. (2004a) Neural Network Applications for Automatic New Topic Identification on Excite Web search engine datalogs *Proceedings of ASIST 2004, Providence, RI* 310-316

Ozmutlu, S., Ozmutlu, H. C. & Spink, (2002b) Multimedia Web searching *ASIST 2002: Proceedings of the 65th American Society of Information Science and Technology Annual Meeting, Philadephia* 403-408

Ozmutlu, S., Ozmutlu, H.C. & Spink, A. (2003a) Multitasking Web searching and implications for design *Proceedings of ASIST 2003, Long Beach, CA* 416-421

Ozmutlu, S., Ozmutlu, H. C., & Spink, A., (2003b) Are people asking questions of general web search engines *Online Information Review* 27, 396-406

Ozmutlu, S., Spink, A., & Ozmutlu, H. C. (2003c) Trends in multimedia web searching: 1997-2001 *Information Processing and Management* 39, 611-621

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

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

Pu, H.T., Chuang, S-L. & 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 University Press, Princeton, NJ, 1976

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., Jansen, B.J. & Ozmultu, H.C. (2000) Use of query reformulation and relevance feedback by Excite users *Internet Research: Electronic Networking Applications and Policy* 10, 317-328.

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

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

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