Neural Network Applications for Automatic New Topic Identification on Excite Web Search Engine Data Logs

H. Cenk Özmutlu[*], Fatih Çavdur, and Seda Özmutlu
Department of Industrial Engineering, Uludag University, Gorukle Kampusu, Bursa, Turkey.
Tel: (++90-224) 442-8176; Fax: (++90-224) 442-8021. Email: {hco, fcavdur, seda}@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. Email: aspink@sis.pitt.edu

The analysis of contextual information in search engine query logs is an important, yet difficult task. Users submit few queries, and search multiple topics sometimes with closely related context. Identification of topic changes within a search session is an important branch of contextual information analysis. The purpose of this study is to propose a topic identification algorithm using neural networks. A sample from the Excite data log is selected to train the neural network and then the neural network is used to identify topic changes in the data log. As a result, 76% of topic shifts and 92% of topic continuations are identified correctly.

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

Introduction and Related Research

The popularity of search engines has attracted the attention of many researchers to investigate Web search engine user behavior. The investigation of user interactions with information retrieval systems are performed within the broader context of their information-seeking behaviors and the context of their search for information (Ingwersen, 1996; Spink, 1998; Vakkari, 2001). In general, a search context contains information about a user's search task and the topic areas the user is interested in (He, *et al.*, 2002). Users interested in different context could benefit more from custom-tailored Web retrieval systems according to their contextual needs, however, currently, Web search engines are not designed to differentiate according to the user's profile. Exploiting the user 's context has the potential to improve Web retrieval systems (Goker, 1997; Talja, *et al.*, 1999).

Obtaining contextual information on Web search engine logs is difficult. Users submit few numbers of queries during their searches (Spink, *et al.*, 2001) and it is difficult to make inferences about the individual search engine user's judgments based on few queries. Recent information retrieval studies suggest that users' searches may have multiple goals or topics (Miwa, 2001) and there is a growing trend among Web search engine users to submit several queries on multiple subsequent topics. (Spink, *et al.*, 2002). Spink, *et al.* (1999) found that 3.8% of Excite users responding to a Web-based survey reported multitasking searches. In subsequent years, 11.4% of Excite 1999 users and 31.8% of FAST 2001 search engine users performed multitasking searches. (Ozmutlu, *et al.*, 2003). Each multitasking user searched an average of 3.2 different topics during each search session for Excite 1999 and Fast 2001 datasets. (Ozmutlu, *et al.*, 2003). The context behind a series of successive searches is probably either closely related to each other or the same (He, *et al.*, 2002). Beyond these studies limited knowledge exists on the characteristics and patterns of multitasking searches. (Ozmutlu, *et al.*, 2003), which makes obtaining contextual information on Web users all the more difficult. He and Goker (2002) mention that there has not been a consistent definition of topics and no clear topic delimiter in Web searches.

There are many implications of contextual information on Web search engine design. Some of these are listed in Ozmutlu, *et al.* (2003). The main element in developing a retrieval system that exploits user context is new topic identification. For example, custom-tailored graphical user interfaces can be offered to the Web search engine user, if topic changes can be estimated correctly by the search engine. In addition, a search engine could provide windowing facilities to allow Web users to generate and track separate topic or related topic queries and facilitate task switching.

There is a limited number of studies focusing on topic identification. He, *et al.* (2002) developed a topic identification algorithm using Dempster Shafer Theory (Shafer, 1976) to automatically identify topic changes using statistical data from Web search logs. Dempster Shafer Theory enables the combination of two separate probabilistic events related to a single property (such as

[*] To whom all correspondence should be directed.

the topic change). He et al. (2002) used the search pattern and duration of a query as the probabilistic events for Dempster Shafer theory. Dempster Shafer Theory requires the probabilities of each event, the weights (importance) of these separate probabilistic events and a threshold value used to identify a topic shift. Probabilities are easily obtained through the analysis of the data log, as the probability for a specific search pattern and query duration. To set the weights and the threshold, He, *et al.* (2002) applied a genetic algorithm. The performance of the topic identification algorithm is determined by the measures of precision and recall, and their combination, the fitness function.

He, *et al.*'s (2002) approach was replicated on Excite search engine data (Ozmutlu and Cavdur, in press). Main finding of Ozmutlu and Cavdur was that the application of genetic algorithms for topic identification could be somewhat problematic. (Ozmutlu and Cavdur (in press)).

At this point, there is a need for further research, which can enable fast and successful automatic topic identification (with a simpler fitness function), and therefore result in effective exploitation of contextual information for development of new search engine algorithms. In this study, we propose a neural network to automatically identify topic changes. The advantage of neural networks is that the training is done by the actual data, not by the help of a fitness function as in genetic algorithms. The training procedures of the neural network does not depend on any complex performance measure, just the information whether its estimation is correct or not. In this study, the neural network is trained with a training sample from the datalog and used to identify topic changes on a test dataset. The success of the network is determined comparing its results to that of the human expert.

A neural network is an algorithm, which imitates the human brain, in terms of learning a specific concept and functioning with respect to what it has learnt. Haykin (1994) defines a neural network as "a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use." He mentions that a neural network " resembles the brain in two respects: 1) Knowledge is acquired by the network through a learning process, 2) Interneuron connection strengths known as synaptic weights are used to store the knowledge." The learning process is established through a learning algorithm. During the learning process, the input and the output of the problem to be solved are provided to the neural network. Knowing the inputs and the answers, the neural network adapts its synaptic weights. Then, only the inputs are provided to the neural network and the network provides the answers or output using the values for synaptic weights. Each neural network has neurons or computing cells, which process the information given to the neural network. The way that the neurons are organized form the structure of the neural network, such as single-layer feeedforward networks, multilayer-feedforward networks, recurrent networks and lattice structures (Haykin, 1994).

Methodology

Research Design

The 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. In the Excite data log structure, 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.

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.

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.

Precision (P), Recall (R) and a fitness function (F_{β}) used in Ozmutlu and Cavdur (in press) are adapted from He *et al.* (2002). These performance measures are only used to compare the performance of the proposed neural network to that of the previous studies. The formulation of these values are as follows:

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

$$R = \frac{N_{shift\,\&\,correct}}{N_{true\,shift}} \quad (2)$$

$$F_{\beta} = \frac{(1+\beta^{2})PR}{\beta^{2}P+R} \quad (3)$$

Proposed Algorithm

In this study, we propose a neural network to identify topic changes in the Excite search engine query log. The search engine query log consists of 10,003 search queries. The general steps of the methodology applied in this paper are as follows:

- Evaluation by human expert
- Dividing the data into two sets
- Identify search pattern and time interval of each query in the training dataset
- Forming the neural network
- Training the neural network
- Applying the neural network to the test data set
- Comparison of results from human expert and the neural network
- Evaluation of results

These steps are explained in detail in the following paragraphs.

Evaluation by human expert:

A human expert goes through the 10,003 query set 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:

Approximately, first half of the data (5014 queries) is used to train the data and the second half (4989) is used to test the performance of the neural network.

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) and Ozmutlu and Cavdur (in press) to avoid any bias during comparison.

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 1 for distribution of the queries with respect to time interval in the training dataset. It should be noted that not all of 5014 queries can be used for training, since the last query of each user session cannot be processed for pattern classification and time duration. Pattern classification and time duration cannot be determined for the last query of each session, since there are no subsequent queries after the last query of each session. In the training dataset, there were 1201 user sessions, so excluding the last query of each session, the test dataset is reduced to 3813 queries from 5014 queries. After the human expert identified the topic shifts and continuations, 3544 topic continuations and 269 topic shifts were identified within the 3813 queries. We also use seven categories of search patterns in this study, which are as follows (Ozmutlu and Cavdur, in press):

- 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 the choice of ''related pages ''.
- Others: If the second query does not fit any of the above categories, it is labeled as other.

The search patterns are automatically identified by a computer program. The 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 2 for distribution of queries with respect to search patterns in the training dataset.

Table 1: Distribution of queries with respect to time interval

Time Interval (min)	Intra-topic	Inter-topic
0-5	3001	77
5-10	218	18
10-15	85	14
15-20	47	7
20-25	22	13
25-30	20	5
30+	151	135
Total	3544	269

Table 2: Distribution of queries with respect to search pattern

Search Pattern	Intra-topic	Inter-topic
Browsing	2371	0
Generalization	58	0
Specilization	166	0
Reformulation	327	1
New	622	268
Relevance feedback	0	0
Other	0	0
Total	3544	269

```
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)
        $B = \{t \mid t \in Q_c$ and $t \in Q_n\}$, the set of terms (terms determined using "space" as a divider) that are
            common in both $Q_c$ and $Q_n$
        $C = \{t \mid t \in Q_c$ and $t \notin Q_n\}$, the set of terms, which appear in $Q_c$ only
        $D = \{t \mid t \notin Q_c$ 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

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.

Training the neural network :

The neural network is trained using the training data set (5014 queries or the first half of the 10,003 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. The training is performed on a Pentium III 800MHz computer in 85.863 seconds.

Applying the neural network to the test data set:

Using the information from training, the neural network is used to identify topic changes in the second half of the 10,003 query datalog (4989 queries). The 4989 queries are fed into the input layer of the neural network. The output layer of the neural network 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) gave a greater priority to Type B errors by increasing β value in their fitness functions. To create similar priority effect in the results of the neural network application, we use a threshold value of 1.3 (any value over 1.3 is considered as 2, and under 1.3 is considered as 1), thus lowering the risk of Type B errors. It should be noted that this threshold value is different from the β in the fitness function equation.

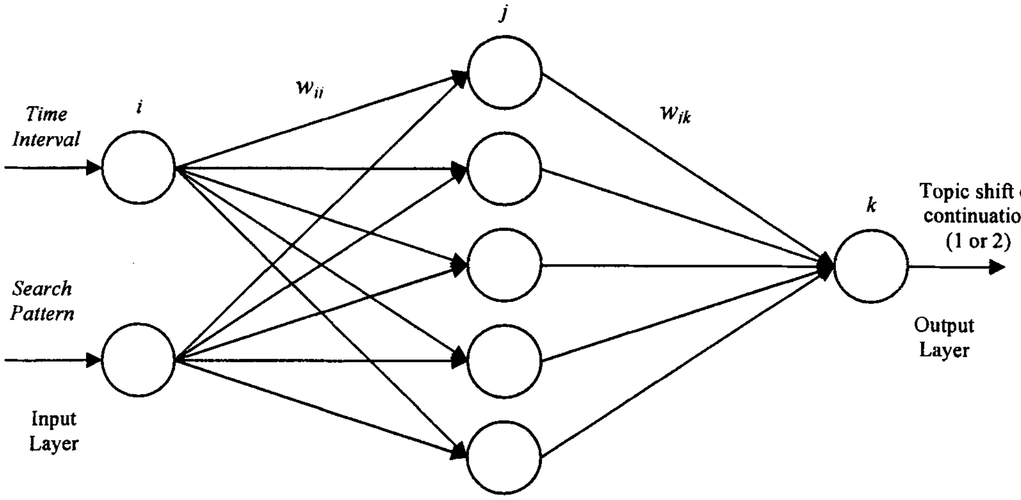

Figure 2: The structure of the proposed neural network

Comparison of results from human expert and the neural network:

The results of the neural network 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 training procedures of the neural network does not depend on any complex performance measure, just the information whether its estimation is correct or not. Still, in order to create a platform for comparison, the performance of the neural network is evaluated in terms of precision, recall and the fitness function. Precision (*P*) is the ratio of the correctly marked topic shifts to the total number of marked topic shifts. Recall (*R*) is the ratio of the correctly marked topic shifts to the actual number of topic shifts. The third and the main measure, fitness function F_{β}, combines precision and recall values into a single value, where β is a given parameter to prioritize different types of resulting errors of estimating topic changes. β is set to 1.5 in this study, as in Ozmutlu and Cavdur (in press) and He *et al.* (2002), giving a higher priority to recall. Higher P, R and F_{β} 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.

When the human expert evaluated the 10,003 query dataset, 7059 topic continuations (94%) and 421 topic shifts (6%) were found. In the subset used for training (first half of the dataset-5014 queries), there are 3544 topic continuations (93%) and 269 topic shifts (7%), and in the second half of the dataset (4989 queries), there are 3515 topic continuations (96%) and 152 topic shifts (4%). The values relevant to the second half of the dataset are given in Table 3.

After running the neural network, we obtain the results in Table 3. We observe that the neural network marked 3268 queries (89%) as topic continuation, whereas the human expert identified 3515 queries (96%) as topic continuation. Similarly, the neural network marked 399 queries (11%) as topic shifts, whereas the human expert identified 152 queries (4%) as topic shifts. During the topic identification process, we observed 283 Type A errors and 36 type B errors.

Using the results of the neural network, we have calculated the *P*, *R* and F_{β} values (also given in Table 3) as 0.291, 0.76 and 0.5088, respectively. These values are used to compare the performance of the neural network approach to previous approaches. Using the neural network approach, 116 (76%) out of 152 topic shifts are identified correctly and 3232 (92%) out of 3515 topic continuations are identified correctly. In terms of F_{β}, we have achieved approximately the same result as those of the genetic algorithm applied by Ozmutlu and Cavdur (in press), who obtained a fitness function value of 0.508 on the same dataset. The important finding of this study is that neural network performs almost as good as the proposed genetic algorithm approach and does not require the usage of any complex probability functions or fitness function.

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)) to avoid any bias in comparison of our methods with previous approaches. We have to point out that it is possible to obtain smaller values for the number of total errors, however such solutions do not provide a better fitness function. 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.

Table 3: Topic shifts and continuations in the entire dataset as result of neural network and as evaluated by the human expert

Origin of results	Number of topic shifts	Percentage of topic shifts	Number of topic continuations	Percentage of topic continuations	$N_{shift\ \&\ correct}$:	$N_{contin\ \&\ correct}$	Type A error	Type B error	P	R	F_{β}
Results from the neural network	N_{shift} = 399	11 %	N_{contin} = 3268	89 %	116	3232	283	36	0.291	0.76	0.51
Results from the human expert	$N_{true\ shift}$ = 152	4 %	$N_{true\ contin}$ = 3515	96 %	----	----	----	----	----	----	----

Conclusion

This study proposes a neural network to automatically identify topic changes in the search engine query logs. The search query log comes from the Excite search engine and a sample of 10,003 queries was selected. The neural network was trained with approximately half the data set. Then, the neural network was run on the second half of the data set to automatically identify topic shifts and continuations. The results were compared to those of a human expert. The performance of the neural network is compared with the previous topic identification algorithms according to their evaluation criteria (precision, recall and fitness function).

As a result of running the neural network, 76% of topic shifts and 92% of topic continuations are identified correctly. A fitness function of 0.5088 was calculated using the results of the neural network. Compared to the previous automatic topic identification algorithms, the neural network approach achieved a similar performance. Consequently, we have proved that it is possible to create successful topic identification algorithms without depending on complex evaluation measures used in other studies.

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. Future work includes developing more enhanced neural network structures and determining which network structure is the most successful in automatic topic identification

REFERENCES

Goker, A. (1997).Context learning in Okapi. *Journal of Documentation*, 53 (1), 80-83.

Haykin, S. (1994). Neural Networks, Englewood Cliffs, NJ, Macmillan College Publishing Company.

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

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

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 3–50.

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

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

Ozmutlu, S., Ozmutlu, H.C. and Spink, A. (2003). Multitasking Web searching and implications for design, *ASIST 2003, Annual Meeting of the American Society for Information Science and Technology*, October 19-22, Long Beach, CA, 416-421

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

Spink, A. (1998). Toward a theoretical framework for information retrieval (IR) within an information seeking context. *Proceedings of the 2nd international information seeking in context conference*, August 12–15, 1998. Sheffield, UK: University of Sheffield, Department of Information Studies.

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

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

Vakkari, P. (2001). A theory of the task-based information retrieval process: A summary and generalization of a longitudinal study. *Journal of Documentation*, 57(1),44–60.