Bull. Nov. Comp. Center, Comp. Science, 39 (2016), 61–67

Regression analysis of text ranking algorithms by neural networks

M.S. Tarkov, O.A. Kozhushko

Abstract. Neural network models for the analysis of the document ranking algorithms are proposed. The models use the Kohonen neural network and a multilayer perceptron. These models were verified using test data, and their application features were revealed depending on the input data.

1. Introduction

The efficient document search at the World Wide Web becomes a challenging problem. This is caused by a rapid increase in the information volume and the appearance of new features of the web document collections. Nowadays, the search engines are constantly optimized with primary attention to the ranking of retrieved documents by their relevance to a text query. To solve this problem, the search engines machine learning from assessor knowledge bases is widely used [5]. This approach allows one to achieve a high efficiency, but it turns the ranking algorithm to a "black box" system. This leads us to the search system identification problem and its key elements retrieval. This problem can be partially solved using the neural network approach [1–5].

2. System identification problem

Analysis of the hidden ranking algorithm is a Data Mining problem [6]. To solve this problem we must identify the factors having a significant influence on the ranking result and the relation of documents positions to these factors values. The data investigated are presented in the vector form, where the vector components correspond to the factors characterizing documents and queries. Examples of such factors examples are: document length, the number of words in a document, the query length, etc.

The ranking algorithm constructs a relevance function that maps a pair of vectors (d, q), describing a document and a text query, respectively, as related to the relevance numerical value: $\text{rel} = f(d, q)$.

In this paper, the way of distinguishing significant factors and their values providing the top document ranking results for a given text query is considered.

3. Data preprocessing

The factors describing a document and a text query may be both quantitative (such as the document size) and qualitative (for example, a document subject). The purpose of the data pre-processing for the subsequent neural network analysis is to convert data to the uniform form. The algorithm has the three stages [1, 2]:

1. A priori exclusion of insignificant components. At this stage, the correlations between the characteristics both of input and output data are separately evaluated, and then insignificant factors are excluded.
2. Representation of inputs and outputs in the numerical form for nominal factors using the binary encoding. Each nominal factor value is mapped onto a vector with components corresponding to digits of a class number binary representation.
3. The data normalization using a bipolar sigmoid activation function [2]: $f(x) = \tanh(\beta x)$, where β is a given coefficient, x is a factor value.

4. Neural network models

The search engine analyzes all known factors but uses only the most important ones in the relevance function. Therefore it is necessary to implement the data factor analysis to identify significant factors. For different pairs (d, q) the various factors are often used. In this paper, important factors are determined as factors taking similar output values for similar input vectors. Characteristics vectors of the query q are the input vectors, and characteristics vectors of the document d are the output vectors y. Also, a model, where some components of the vector d are included in the input vector x and other components are included in the output vector y, can be considered.

The proposed algorithm of the factor analysis uses the Kohonen neural network [1, 2, 4, 7] and the k-means clustering method [1, 9]. The Kohonen network usage is caused by the retraining time reduction and the ability to use a trained network at the next analysis stage.

4.1. The algorithm of factor analysis consists of the two steps:

Carrying out the multi-dimensional input vectors clustering using the Kohonen network. The number of output neurons (the number of clusters) is determined experimentally.

Carrying out the one-dimensional data clustering into two clusters by the k-means algorithm for each data cluster and each factor. Use the clustering result and given parameters ε and p to evaluate the factors importance by the following rule: if the vectors percentage in the smallest cluster exceeds $p \in (0, 1)$ and

$$\sum_{i,k} L(x_i, w_k) > \varepsilon,$$

the factor is determined as insignificant. Here w_k, $k = 1, 2$, are cluster centers, x_i $i = 1, \ldots, N$, are factor values, N is the number of samples, $L(x, w)$ is the distance between x and w. Using the parameters ε and p allows us to distinguish the situations of one wide cluster or two close clusters.

When the factor analysis finishes, we have the following results:

- The partition of input vectors into clusters and their clusters centers;
- The trained Kohonen neural network;
- A set of significant factors for every cluster.

4.2. Regression analysis. In the course of the regression analysis it is required to obtain optimal values of significant factors. Two systems based on a multilayer perceptron [1, 2, 4] are considered as the regression analysis models.

A model based on a hybrid neural network. A hybrid neural network is a cascade union of the Kohonen network and the three-layer perceptron [2, 7]. The Kohonen self-organizing network identifies the significant factors of input vectors. These factors correspond to the input neurons of the multilayer perceptron, so the input number of vectors is equal to the number of neurons in the perceptron input layer. The number of output neurons is equal to that of significant factors, where each factor is significant at least for one cluster. The number of hidden neurons is determined experimentally.

The hybrid network handles all of the clusters. It requires an additional modification of a training set because some factors are insignificant for some clusters. The training set modification lies in zeroing the output vector components which are insignificant for the cluster corresponding to the input vector.

The hybrid neural network disadvantage is its difficulty in training. It was shown in [8] that a large number of zero components in the training set intails the gradient methods training complexity because the neural network error surface has vast flat areas. The use of hybrid network is justified when sets of the significant factors have large intersections.

A model based on a set of multilayer neural networks. In this model we propose to use a single three-layer perceptron for each cluster allocated by the Kohonen network. The perceptron training set consists of input and output vector pairs where each input vector belongs to the corresponding cluster.

The number of input neurons is equal to that of input vector components number; the number of hidden neurons is determined experimentally; the number of output neurons is determined by the number of significant factors in the cluster in question.

This model allows us to avoid the training set zero values problem arising in the hybrid neural networks, but it is more cumbersome because each cluster corresponds to a single approximating neural network. This model is justified if the number of the important factors or the number of the clusters is small.

The neural network models learning process has the following features:

- The result quality of every stage (cluster analysis, factor analysis, and regression analysis) has the key influence on the subsequent stage.
- The training quality is not impaired if the number of Kohonen network neurons exceeds the actual cluster count. This is explained by the fact that a partitioning of a set of similar vectors into smaller clusters does not adversely affects the factor analysis stage.
- If a certain cluster has a large percentage of incorrectly recognized vectors, the hybrid neural network error value decreases because the number of the significant factors also decreases.

The described features allow us to choose the regression analysis model based on the factor analysis results:

- In the case of allocating a number of large clusters, the model based on the hybrid neural network is preferable.
- In the case of weak intersections between sets of the significant factors in different clusters, the multilayer perceptron model is preferable.
- In the case of allocating a large number of clusters with weak intersections between sets of the significant factors, it is possible to use the model based on a complex of the hybrid neural networks where every hybrid network handles clusters with strong intersections between significant characteristic sets.

5. Experiments

The classical algorithm Okapi BM25 [9] as an investigated ranking algorithm is chosen. The retrieval function BM25 ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). The relevance function value is calculated as follows:

$$\mathtt{BM25}(d,q) = \sum_{t\in q} w_{t,d},$$

where

$$w_{t,d} = I_t \frac{(k_1+1)F_{t,d}}{k_1(1-b+bL/\bar{L})+F_{t,d}},$$

d is a text document, q is a text query, t is a query term, L is the document length, $\bar{L}$ is an average document length, $F_{t,d}$ is the frequency of the term t in the document d, I_t is the inverse frequency of the term t, $k_1 = 2$ and $b = 0.75$ are verified experimentally.

The training set is based on the text collection ROMIP-2003, and the queries are taken from the task ROMIP-2006 [10]. We selected the queries with length between 2 and 5. They do not have digits, misspelling, unknown words, and, at the same time, the collection contains at least five documents including all the query terms. The vector q consists of 11 components, some components may be zero. The first 5 component pairs include the values $F_{t,q}$ and I_t, $t \in q$, which are query factors. The last component determines the query length.

Input data are the query vectors q, output data are the documents vectors d with the first rank in the ranking by the Okapi `BM25` algorithm. The vector d consists of 5 values of $F_{t,d}$ and the value of L. The training set consists of 141 queries of length 2, 158 queries of length 3, 157 queries of length 4, and 133 queries of length 5; 80 % of the vectors are used as the training vectors and 20 % of them are used for testing.

We use the mean square error function

$$\frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - d_{ij})^2,$$

where N is the number of the examples, M is the output vectors dimension, y_{ij} are the neural network output vector components, and d_{ij} are expected output vector components.

At the data preprocessing stage, the component corresponding to the number of query words is excluded because it is a linear combination of other five components. After the factor analysis the input vectors were partitioned into clusters by the number of query terms. An optimal separation was obtained using 8 neurons in the Kohonen network — per 2 clusters corresponding to the queries of lengths 2, 3, 4, and 5.

For each cluster, the significant factors are identified as $F_{t,d}$ and L, where $t \in q$ for a large portion of documents. The portion size is determined by the parameters ε and p. In the experiment, we use the values $\varepsilon = 0.01$ and $p = 0.25$, that corresponds to the portion of 75 %. In the perceptron hidden layer of the hybrid neural network, 16 neurons are used. In the hidden layers of the perceptron complex, 8 neurons are used. The trained neural network response to the input query vector is the vector d ensuring high document rankings of the query. The neural network response is considered as incorrect if the document relevance is lower than the relevance of documents from the training set. Tables 1 and 2 show the error values obtained on training and testing data (the portion of incorrect answers) in 8 different

Table 1. Hybrid network training results

S_c	51	63	58	61	64	50	60	63
$\epsilon \cdot 10^4$	1	1	0	0	533	3	2	1
p	0	0	0	0	0.0312	0	0.0167	0

Table 2. Multilayer perceptron complex training results

S_c	51	63	58	61	64	50	60	63
$\epsilon \cdot 10^4$	2	1	6	2	2	1	1	1
p	0	0	0	0	0.0312	0	0.0167	0

clusters allocated by the Kohonen network. Here S_c is a cluster size, ϵ is a training error, p is a portion of incorrect answers. The results of testing show successful training and the low values of training and testing errors.

6. Conclusion

An algorithm for identifying the text search systems, including the factor and the regression analysis, is proposed. The factor analysis includes the data clustering based on the Kohonen network use. At the regression analysis stage, two neural network models are used: a model based on the hybrid neural network, and the model based on a complex of multilayer perceptrons. The algorithm testing indicates to the successful models training and to unsuccessful training and testing error values. The hybrid neural network model has some training difficulties if the number of the significant factors is large.

The algorithm bottleneck is the factor analysis method allowing one to distinguish the significant factors. Except the proposed statistical recognition method distinguishing the most important factors, the Bayesian networks [11] or fuzzy logic methods can be used [3]. The next investigation step can be the analysis of the complex factors impact that are nonlinear combinations of the measured characteristics. The modifications are also possible by introducing the characteristics weights that influence the neural network models training and the results interpretation.

References

[1] Ezhov A.A., Shumsky S.A. Neurocomputing and its Applications in Economics and Business. — INTUIT, BINOM, Laboratory of Knowledge, 2007 (In Russian).

[2] Osowski S. Neural Networks for Information Processing. — Moscow: Finance and Statistics, 2002 (In Russian).

[3] Rutkowski D., Pilinsky M. Rutkowski L. Neural Networks, Genetic Algorithms and Fuzzy Systems. — Moscow: Goryachaya Liniya — Telecom, 2004 (In Russian).

[4] Staroverov B.A., Mormylev M.A. Complex usage of neural networks for prediction automation of electro consumption on regional level // Bulletin ISPU. — 2009. — No. 4. — P. 78–81 (In Russian).

[5] Tarkov M.S. Neurocomputer Systems. — INTUIT, BINOM, Knowledge Laboratory, 2006 (In Russian).

[6] Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. — 3rd ed. — Elsevier, 2012.

[7] Gulin A., Karpovich P., Raskovalov D., Segalovich I. Optimization of ranking algorithms by machine learning methods // Trudy ROMIP'2009. — P. 163–168. — http://romip.ru/romip2009/15_yandex.pdf (Usage date: 16.12.2015) (In Russian).

[8] Kordos M., Duch W. A survey of factors influencing MLP error surface // Control and Cybernetics. — 2004. — Vol. 33, No. 4. — P. 611–631.

[9] Upstill T.G. Document ranking using web evidence: Ph.D. thesis. — The Australian National University, 2005.

[10] Russian seminar on estimation of information research. — http://romip.ru/ (Usage date: 16.12.2015) (In Russian).

[11] Terekhov S.A. Introduction to Bayesian Networks // 5th All-Russian Scientific Conference "Neuroinformatics 2003" / Lectures on Neuroinformatics. Part 1. — Moscow: MIFI, 2003. — P. 149–187 (In Russian).