Dimension Reduction Methods of Text Documents by Neural Networks

Lenka Skovajsová, Igor Mokriš,

[1] Slovak Academy of Science, Bratislava
upsylesk@savba.sk, igor.mokris@savba.sk

Abstract. The paper is oriented to introduce different dimension reduction methods in the text document retrieval area. First, the mostly used text document retrieval models are described, and then in second part the analytical approach and neural network approaches to dimension reduction of keyword space are described. Dimension reduction methods reduce keyword space into much smaller size together with retaining similarity on the highest possible level. The result of dimension reduction of text documents is saving memory space used for document representation.

Keywords: Information retrieval, Neural networks, Dimension reduction, Performance evaluation

1 Introduction

The Internet is full of electronic documents in different formats. We consider here documents in the HTML and/or XML format. We use them from the reason that they can be shared by different applications and they can be easily processed to obtain collection of documents for retrieval.

The documents are first downloaded from the internet. Then they are preprocessed to obtain retrievable collection of documents from the downloaded documents. When they are preprocessed, the inner representation of them is obtained to be able to use them for retrieval process. The system, which is able to return relevant document to the user query, is called information retrieval system.

The information retrieval process resides in the ability of system to return relevant documents when user enters query to this system. Query can be in different formats. It can be a keyword, a phrase in natural language, a pattern, etc. Information retrieval system processes the user query, creates the inner representation of it and on the base of comparison of query representation and document representation obtains the similarity of the document to the query. When the similarity is greater than some threshold value, the document is returned as relevant.

The paper is divided in the following parts. In the second part, the preprocessing methods are reviewed, in the third part the basic types of information retrieval systems are introduced and in the fourth part the dimension reduction methods are described.

2 Document preprocessing

Document preprocessing can be realized in two steps. First, the stop word removal is performed and second, the keywords are obtained by the Porter's algorithm.

2.1 Stop Word Removal

The documents obtained in the HTML of XML format must be firstly preprocessed to obtain representation, which is processable by the information retrieval system. First step in preprocessing is removing stop words. In this step the stop words, which have no semantic meaning and all non processable symbols are removed from documents. After stop word removal only indexable words are retained in the documents.

2.2 Porter's Algorithm

Porter's algorithm comes out of the document representation, in which only keywords are retained. These keywords have different shape and it is needed to retain only root base of the words. This is the role of Porter's algorithm. Porter's algorithm modifies the keywords by removing the suffixes to obtain the root base of keywords suitable for inner document representation.

3 The Information Retrieval Models

For information retrieval three parts of information retrieval model are needed. They are the inner document representation, inner query representation, and the manner of obtaining similarity of document to the query.

The three most used information retrieval models are described below, the Boolean Model, Vector Space Model, and Language model.

3.1 Boolean Model

In the Boolean model [11], all elements of the term-document matrix are either 1, to indicate presence of the term in the document or 0, to indicate absence of the term in the document. A query q is a conventional Boolean expression of index terms. Similarity of query q and document d is obtained by the formula:

$$sim(d_j, q) = \begin{cases} 1 \; if \; \exists c(q) | c(q) = c(d_j) \\ 0 \qquad\qquad otherwise \end{cases}$$

Where $c(q)$ is any of the query conjunctive components, and $c(d_j)$ is corresponding document conjunctive component.

3.2 Vector Space Model

In the Vector Space Model [8, 11], each document from the document collection is represented as a vector of keywords. The whole document collection is represented as a matrix called the Vector Space Model matrix. Vector Space Model Matrix has the form

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \cdots & \cdots & \ddots & \cdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}$$

where each column represents the document keyword vector and each row represents vector of frequencies of particular keyword in each of documents.

Query in Vector Space Model is represented as a vector of keywords

$$q = (q_1 \quad \cdots \quad q_m)$$

The similarity between the document and the query is computed by

$$sim(d_j, q) = q \times d_j$$

3.3 Language Model

A document in language model is represented as a sequence of words [11]

$$d_j = (x_1 x_2 \dots x_t)$$

where t is number of words in document d_j.

The similarity of document to the query can be expressed by

$$P_{uni}(t_1 t_2 t_3) = P(t_1)P(t_2)P(t_3)$$

for unigram model and

$$P_{bi}(t_1 t_2 t_3) = P(t_1)P(t_2|t_1)P(t_3|t_2)$$

for bigram model where $P(.)$ is the probability of terms in the judged document.

In rest of the paper we'll come out from Vector Space Model on which we'll apply different dimension reduction methods.

4 Dimension reduction methods

The size of document collections is very often huge and in most cases it is efficient to reduce this size. In this paper we focus on the dimension reduction of the keyword space where we replace keywords with much less dimensional features, which will retain the most of similarity and occupy much less space when they are stored on the disk.

The described methods are divided into two parts. First part is analytical method to dimension reduction of the document collection. The analytical method uses Singular Value Decomposition to reduce the keyword space. Other two described methods use neural networks for dimension reduction of document space. One is based on using Hebbian neural network with Oja learning rule. The second method uses autoassociative neural network.

4.1 The Analytical Approach – Latent Semantic Indexing Model

Latent semantic indexing [1, 2, 4, 6] model uses singular value decomposition to divide VSM matrix into three matrices:

$$X = USV^T$$

Where U is the matrix of left singular vectors, S is the diagonal matrix of positive singular values and V is the matrix of right singular vectors.

The next step is dimension reduction. Matrices are reduced to

$$X_R = U_r S_r V_r^T$$

and the reduced document and query representation is obtained as

$$d_r = dU_r S_r^{-1}$$
$$q_r = qU_r S_r^{-1}$$

The similarity of the reduced document d_r and reduced query q_r is computed by product

$$sim(d_r, q_r) = d_r q_r$$

4.2 Hebbian Neural Network with Oja Learning Rule

The first neural network approach to dimension reduction is by Hebbian neural network with Oja learning rule [3, 5, 7, 9]. It is unsupervised two-layer neural network with m neurons on the input layer and r neurons on the output layer. m represents number of keywords in the VSM matrix and r represents number of features after dimension reduction.

Training is given by these formulas

$$y_j = \sum_i w_{ki} x_{ij}$$

$$w_{ki}(t+1) = w_{ki}(t) + \gamma(t) y_{kj} - \sum_p y_{kp} w_{pi}$$

where y_j is the jth output neuron w_{ki} is the weight between k-th input neuron and i-th output neuron, and $\gamma(t)$ is learning parameter.

Reduced representations of documents and query are computed as

$$d_{red} = \sum_{i=1}^{m} w_{ki} d_i$$

$$q_{red} = \sum_{i=1}^{m} w_{ki} q_i$$

The similarity between document and query is computed as in LSI model by equation (X).

After training Hebbian neural network wit Oja learning rule, the documents are passed on the input layer and reduced document representations appear on the output layer.

4.3 Autoassociative Neural Network

Autoassociative neural network is unsupervised three-layer neural network that uses backpropagation learning rule [3, 5, 7, 10]. Its input and output layer have *n* neurons and hidden layer has *m* neurons. *m* is number of keywords in the VSM matrix and *r* is number of features after dimension reduction.

Training is performed by backpropagation learning rule where on the input is passed the document or query vector and is compared with the same document or query vector on the output layer. So, as the expected output which is required by backpropagation serves the input vector.

After training the reduced representations of documents are obtained by multiplying the input vector with weights W between input and middle layer

$$d_{red} = dW$$
$$q_{red} = qW$$

and the similarity between d and q is computed by (x).

5 Experiments

Methods presented in this paper were tested on different document collections. For verifying the correctness of these methods, experiments were made on the artificial

document collections and for verifying applicability in practice, experiments were made on the real collections of documents acquired from Reuters collection.

It was also shown, that F-measure is high even for small dimensions to which the VSM matrix is reduced. On artificial document collection, by dimension reduction from dimension 80 to dimension 6, by using the neural networks described above, the precision, Recall and F-measure were equal 1.

One disadvantage of using neural network approaches is that they are not precise after each training, because the weights are on the beginning of training chosen randomly. It is recommended to train the neural network more times and choose the networks with highest F-measure for that collection.

6 Conclusions

All experiments show that used dimension reduction methods are applicable in information retrieval area, and that with growing dimension the precision, recall and F-measure grows to value 1.

Acknowledgments.
This work was supported by the Slovak Science Agency VEGA No. 2/0054/12.

References

1. Kolda, Tamara G. and Bader, Brett W.: Tensor Decompositions and Applications. JSIAM Review, vol. 50, no. 3, 455-500 (2009)
2. Harrag, F. and El-Qawasmah, E.: Neural Network for Arabic text classification. In: Second International Conference on the ICADIWT '09, pp. 778-783, (2009)
3. Cord, Matthieu and Cunningham, Pádraig: Dimension Reduction. Springer Berlin Heidelberg, pp. 91-112, (2008)
4. Muflikhah, L. and Baharudin, B.: Document Clustering Using Concept Space and Cosine Similarity Measurement. In: International Conference on Computer Technology and Development, 2009. ICCTD '09, pp. 58-62. (2009)
5. Hujun Yin and Weilin Huang: Adaptive nonlinear manifolds and their applications to pattern recognition. Information Sciences, vol.180, no. 14, pp. 2649 - 2662 (2002)
6. Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, pp. 391-407 (1990)
7. Fiori S.:An Experimental Comparison of Three PCA Neural Networks. Neural Processing Letters, pp. 209-218 (2000)
8. Salton, G.: The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc. (1971)
9. Oja, E.: The Nonlinear PCA Learning Rule and Signal Separation – Mathematical Analysis., (1995)
10. Baldi, P., Hornik, K.: Neural Networks and Principal Component Analysis: Learning from Examples without Local Minima. Neural Networks, Vol. 2, No. 1, pp. 53-58 (1989)
11. Baeza-Yates, Ribeiro-Neto B., Modern Information Retrieval, Addison-Wesley (2011)