ESANN'2003 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 23-25 April 2003, d-side publi., ISBN 2-930307-03-X, pp. 9-14

Subject Categorization for Web Educational Resources using MLP

Minoru NAKAYAMA* , Yasutaka SHIMIZU†

* CRADLE, Tokyo Institute of Technology
O-okayama, Meguro-ku, Tokyo 152-8552 Japan

† National Institute of Educational Policy Research
Shimo-meguro, Meguro-ku, Tokyo 153-8681 Japan

Abstract. The purpose of this study is to develop subject categorization methods for educational resources using multilayer perceptron (MLP) and to examine the performance of the test documents as an application system. To examine the performance two methods are examined: Latent Semantic Indexing method (LSI) and a three layer feedforward network as a simple MLP. The document vectors were estimated by the term feature vectors which were extracted from the teaching guidelines based on the singular value decomposition method (SVD). Comparing recall and precision rates and F1 measure for the subject categorization, the categorization performance using MLP showed better than using LSI.

1. Introduction

The Japanese governmental project, "e-Japan" promotes the use of the World Wide Web (WWW) in the classrooms. Presenting the difficulty with Web retrieval is well known. Thus appropriate support is necessary to enable for school teachers to create lessons using the WWW. Most school lessons in Japan are normally designed according to the national curriculum guidelines defined by the Japanese Ministry of Education, Sports, Culture, Science and Technology (MEXT)[1]. The guidelines with respect to teaching describe the required content based on school subjects and grade level.

Various methods for information retrieval have been developed and implemented in many systems such as learning support system. Generally, those implementations are achieved by using specific methods and a dictionary to specify domain content. For example, some machine training methods which used artificial neural networks(ANN) or support vector machines (SVM)[2] have produced good performances in some instances. For most of these however training data are prepared from a huge collection of document characteristics and expert's categorization label. This preparation often poses a serious constraint to developing a categorization model because the definition of the web for educational use is not clear and there is no educational collection for the machine training. Therefore, it is hard to apply the machine training for the educational resources.

In this paper, we propose developing a categorization model using multilayer perceptron (MLP) as a simple ANN for information retrieval and evaluating it's performance for educational resources. We argue that if the categorization

This research is partially supported by the Japan Society for the Promotion of Science (JSPS) , Grant-in-Aid for Scientific research (B)(2)–135558012, 2001-2003

correspondence author; email: nakayama@cradle.titech.ac.jp model is able to be trained by the guidelines as an essential data, it can then be applied to various data sets including Web documents.

The purposes of this paper are:

1. To develop a training procedure using MLP, based on the generated training data which extract characteristics of typical documents.
2. To examine the performance of this procedure for educational resources, in comparison with a traditional method.

2. The categorization model

2.1. Training data preparations

In order to extract terms and the terms' histogram from the guidelines of primary school teaching [1], the morphological analysis was conducted on all description text using the Japanese text processor 'ChaSen'[4]. The noun and verb terms were selected, but some terms were deleted as stop words. This set of extracted terms is the corpus of keywords for educational resource. The terms' frequencies were summarized for all the content in all school subjects. The total number of corpus-terms was 1919 and the total number of categories for subjects was 11. Therefore, the dimensionality of the term document matrix was 1919×11.

Deerwester et al.[3] reported a document retrieval method using feature vectors of term and document. This is called the latent semantic indexing (LSI), which is a kind of vector space model. In this method, the feature vectors for each term and document category can be extracted from the terms' histograms for the categories as a term document matrix by singular value decomposition (SVD). The matrix X of the term document matrix was divided into three matrices ($X = TSD'$) [1]. The decomposed T and D' can be referred to as the 11-dimensional feature vectors of term and subject respectively.

The curriculum guidelines for teaching can also be categorized as 53 items across 11 subjects and 6 school years [2]. The 53 dimensional feature vectors of term and subject-year were extracted from a 1919×53 term document matrix by the SVD. In this paper, all feature vectors did not reduce the dimensionality. According to the LSI method, document categorization can be conducted by comparing the dot product of the subject feature and the document feature vector based on the corpus-terms' frequency. This estimation procedure is shown later below.

2.2. The training procedure

In order to conduct supervised training, on an appropriate system for subject categorization, the subjects information must be available to the feature vectors. However, while the subjects' feature had the necessary subject information for such categorization, this was not the case for the terms' features.

The subjects' features were originally extracted from the terms' frequency. In this research, the normalized frequency across subjects was defined as the supervised signal for the terms' feature vectors. Training was conducted with the above data pattern on a MLP network with one hidden layer, as shown in Figure 1. The input layer consists of 11 nodes for 11-dimensional feature vectors

[1] T: 11-dimensional left-singular vectors for 1919 terms, S: diagonal matrix of 11 singular values, D: 11-dimensional right-singular vectors for 11 subjects.

[2] 53 categories: 7 categories of language (Lng), arithmetic (Ari), drawing and craft (Drw), music (Mus); summary and school year 1-6, 5 categories of science (Sci) and social study (Soc); summary and year 3-6, 4 categories of sports (Spo) and moral (Mor); summary and 3 steps, 3 categories of living study (Liv) and domestic science (Dom); summary and 2 school years, 1 category of special activity (Spe).

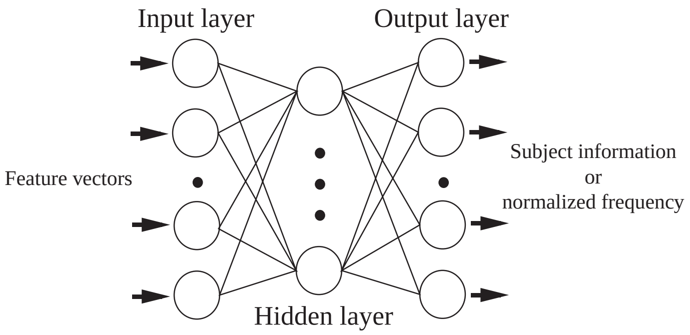

Figure 1: a MLP network with one hidden layer.

$y_j^{(0)}$ and the output layer consists of 11 nodes for the subject information $\hat{y}_j^{(2)}$. These training data, consisting of input and output activations pairs $\{Y_i^{(0)}, \hat{Y}_i^{(2)}\}$ was given a MLP to simulate the mapping. The following equation represents $y_j^{(2)}$:

$$y_j^{(k)} = f_j^{(k)}\left(\sum_{i=1}^{N_{k-1}} w_{ij}^{(k-1)} y_i^{(k-1)} - \theta_j^{(k)}\right)$$

Here $f_j^{(k)}(\cdot)$ is a sigmoidal function and $\theta_j^{(k)}$ is the threshold of the j'th unit in the k'th layer, w_{ij} is the connection weight value between the unit of the two layers. In order to avoid conflict in the training, features suggesting more than one subject were removed, thus producing 1437 (1426+11) pattern data. Using backpropagation for adjustment, the network was trained to simulate the mapping of features and subject information with the hidden layer units: $min\sum_{i=1}^{I}||Y_i^{(2)} - \hat{Y}_i^{(2)}||^2$. The initial state of the network is determined by the weight values assigned to the connections of nodes between two layers before training. Those weights can be assigned according to a random seed. Therefore training was conducted five times for each architecture, evaluating average root mean square for the network as a performance index.

For 53-dimensional feature vectors and the subject information, training was conducted as well as 11-dimensional features. The training data were 1479 (1426+53) pattern data.

2.3. The test document for performance evaluation

A set of teaching scheme for a lesson was selected as test documents. This selection is appropriate because each scheme has a required subject-grade label, including educational contents. An estimation of the feature vector for the teaching scheme was based on the feature vectors that occurred. The extracted terms from the guidelines, however, did not cover all terms which can occur in teaching schemes.

The feature of a document y was predicted by the following equation, $\hat{y}$. The term $\hat{y}$ is the sum of the frequency weighted term vectors; f_i is the frequency for term i; a squared histogram vector $Q = \{q_1, .., q_{1919}\}$; $f_i = {q_i}^2$ [6], ; and $\hat{y} = QT$.

Table 1: The global contingency table

		document category	
		YES	NO
Model categorization	YES	w	x
	NO	y	z

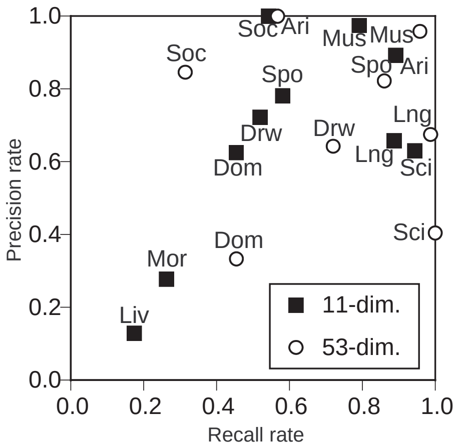

Figure 2: Recall-Precision rate of categorization for MLP.

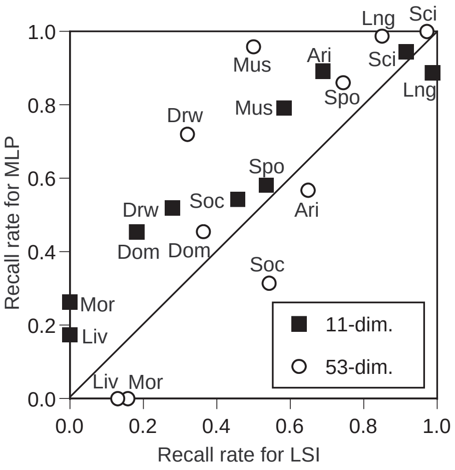

Figure 3: Recall rate of categorization between MLP and LSI.

To estimate the schemes' feature vectors, morphological analysis was conducted on 403 schemes on the Web for specific subjects such as 11 subjects; language (Lng), arithmetic (Ari) and etc.

2.4. The categorization by the vector space model

Both training data and test document features refer to the LSI model, following which the categorization performance for LSI was examined as a reference. To categorize a document for a subject based on the LSI method, every similarity between vectors of the scheme and target subject was tested. The dot product between two vectors gave a degree of cosine similarity when both vectors were normalized. The subject categorization was conducted by the degree of the similarity in the following equation [5]: $(D_i, \hat{y}) = \max_i(D_i, \hat{y})$.

3. The categorization performance

In order to examine the performance of the trained network, three evaluation measures were introduced for the teaching schemes: recall rate, precision rate and F1 measurement.

The document categorization result is often summarized as a contingency table, such as Table 1[7, 2]. The recall rate is defined as the conditional probability that, if a document ought to be classified under the category, this deci-

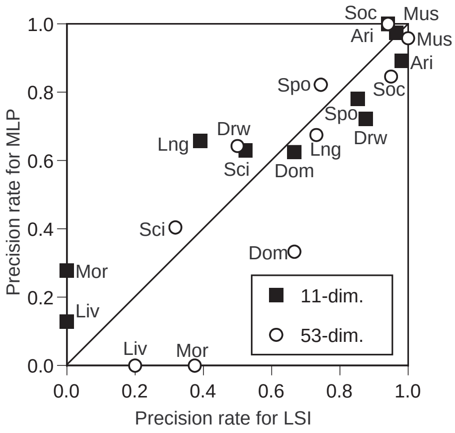

Figure 4: Precision rate of categorization between MLP and LSI.

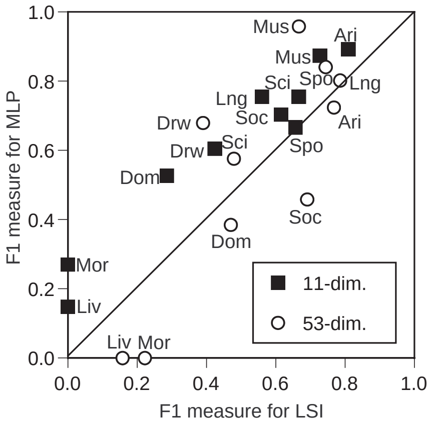

Figure 5: Comparison of F1 measure between MLP and LSI.

sion is taken. The precision rate is defined as the conditional probability that if a document is classified under the category, this decision is correct. According to the table, three measures are defined as the following equations, recall rate: $R = w/(w + x)$, precision rate: $P = w/(w + y)$, F1 measure: $F1(R, P) = 2RP/(R + P)$. The F1 measure shows a total performance based on both of recall and precision rates [8].

The optimum model was achieved respectively for two dimensional features. The model performance was often evaluated by the recall rate for training data, after which the recall rates are compared across the training conditions. All training conditions obtained the perfect recall for 11-dimensional features. Therefore the optimum condition was achieved by the recall rates for the test document, of which there were 18 hidden layer units. In addition, there were 60 hidden layer units for training data of 53-dimensional features.

The relationship between recall rate and precision rate is displayed in Figure 2. The solid cube symbol indicates each subject performance of categorization for 11-dimensional features, the open circle symbol indicates for 53-dimensional features. There were some differences of performance between the two dimensional features. However, the measure of performance between the two features depends on the subject; some subjects require higher dimensionality features. The figure suggests that both rates are higher for most subjects. In particular, Mus, Ari and Spo show that both rates of over 80%. This result suggests the effectiveness of this training procedure for MLP in document categorization.

In order to examine the categorization performance, both recall and precision rates were compared with the rates for the LSI method in Figure 3 and 4. The diagonal line showed equal performance between the two methods respectively. Recall rates for MLP were higher compared with LSI in most subjects (Figure 3). However the precision rates between them were equivalent (Figure 4). In order to clarify the performance difference between them, F1 measure as a total performance was compared, as shown in Figure 5. The figure also shows most major subjects locating in area of MLP dominant. As F1 measure was a combination of recall rate and precision rate, this result reflects the performance of recall rate. Therefore, categorization performance of this method using the MLP is better than LSI.

4. Conclusion

In order to categorize web documents for appropriate educational resources, a document categorization method for school subjects was developed based on the national curriculum guideline for teaching using MLP. First, feature vectors for terms and document categories were extracted by the singular value decomposition (SVD) according to the latent semantic indexing method (LSI). Training data sets were prepared based on the feature vectors for three layer feedforward network (MLP). Second, document feature vectors were estimated by the sum of the terms' feature vectors, then document categorization was conducted on two methods: LSI and MLP. Comparing recall rate, precision rate and F1 measure for the subject categorization, the categorization performance using MLP showed better results compared to using LSI.

The better performance for each subject depends on the dimensionality of feature and categorization method. Also, machine learning methodologies provide various document categorization algorithms such as using SVM (support vector machines), Naive Bayes, etc. Therefore, appropriate system ensemble methods should be considered in order to achieve the best performance. Furthermore, appropriate application system for Web categorization should be developed based on these categorization methods. Those will be our future work.

References

[1] Monbukagakusho(MEXT) *"The guidelines for teaching[primary school edition]"*. Tokyo, 1989.

[2] Sebastiani F. *"Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol.34, No.1, pp.1–47.* 2002.

[3] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. *"Indexing by Latent Semantic Analysis", J. of the Am. Soc. for Info. Sci., Vol.41, No.6, pp.391–407.* 1990.

[4] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., & Asahara , M. *"Japanese Morphological Analysis System "ChaSen""*. URL http://cl.aist-nara.ac.jp/lab/nlt/chasen, 1992.

[5] Luo F.-L. Unbehauen R. *"Applied Neural Networks for Signal Processing", pp.23–73.* Cambridge University Press, New York, USA, 1997.

[6] Takayama, Y., Flournoy, R., Kaufman, S., & Peters, S. *" An Information Retrieval System based on Word Associations - InfoMap", IPSJ Technical Report of Fundamental Infology, Vol.53, No.1, pp.1–8.* 1999.

[7] Tokunaga, T. *"Information Retrieval and Language Processing"*. Univ. of Tokyo Press, Tokyo, 1999.

[8] Pierre, J. M. *"On the Automated Classification of Web Sites", Computer and Information Science, Vol.6, nr0, pp.1–12.* 2001.