Improvement of Information Filtering Using Topic Selection

Takeru Yokoi
Computer and Systems Science
Osaka Prefecture University
1-1 Gakuencho Sakai-City
Osaka 599-8531
takeru@sig.cs.osakafu-u.ac.jp

Hidekazu Yanagimoto
Computer and Systems Science
Osaka Prefecture University
1-1 Gakuencho Sakai-City
Osaka 599-8531
hidekazu@cs.osakafu-u.ac.jp

Sigeru Omatu
Computer and Systems Science
Osaka Prefecture University
1-1 Gakuencho Sakai-City
Osaka 599-8531
omatu@cs.osakafu-u.ac.jp

Abstract

We propose an information filtering system using Independent Component Analysis (ICA). A document-word matrix is generally sparse and has ambiguity of synonyms. To solve this problem, we propose a method to use document vectors represented by independent components generated by ICA. The independent component is considered as a topic. Concretely speaking, we map the document vectors into topic space. Since some independent components are useless for recommendation, we select necessary components from all independent components by Maximum Distance Algorithm. We create a user profile from transformed documents with Relevance Feedback. Finally, we recommend documents by the user profile and evaluate accuracy of the user profile by 11-point average precision. We carry out an experiment to confirm the advantage of the poposed method.

Keywords: Independent Component Analysis, Information Filtering System, User Profile, Maximum Distance Algorithm

1 Introduction

As information technologies have been advanced, a plenty of information are served in the Internet. It has been difficult to find what we demand from large amount of information by existing retrieval systems since we cannot express our guery correctly. A lot of researchers pay attention to developing information retrieval systems which automatically select the information depending on our interests.

The information retrieval systems with user's interests have been studied. For example, there are ranking methods that sort information depending on the user's interests and filtering systems which select the information depending on the user's interests. Since documents usually have noise which worsens accuracy of information filtering, it is a promising method to cut off the noise from documents. It is reported that the method such as LSA[1], which transforms document space, is effective in denoising. The LSA focuses on variance of documents and cuts off the components of low variance for denoising.

We take account of independence of topics included in documents and introduce Independent Component Analysis (ICA) to obtain the topics. ICA is recently used for signal processing, image processing and so on. It has been already reported that the independent components mean topics included in the documents in applying ICA to documents[2][3]. We use the independent components for transformation of document vectors and improve recommendation accuracy of documents.

However, some components obtained by ICA are unnecessary components(noise) for the object of information filtering. Though it is important to remove the noise, there is no criterion to select the independent components. Hence, to focus on the similarity of topics, we use Maximum Distance Algorithm (MDA)[4] which is often used in the field of pattern recognition. This algorithm is applied to classifying the independent components and extracting the useful components for document recommendation. Then we map the document vectors into the space which consists of the selected topics and construct a user profile with the transformed document vectors by relevance feedback(RFB)[5]. Finally, to confirm the proposed method, we carry out an experiment on test collection (NTCIR2[6]).

2 Our proposed method

In this chapter, we explain a user profile, ICA and MDA.

2.1 User Profile

A document vector is a row vector whose elements are weights of words in a document. When the number of words is n and the weight for the ith word is w_i, the document vector d is denoted as

$$d = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_n]^T \tag{1}$$

where $[\cdot]^T$ means transportation.

A user profile is also denoted by a row vector whose elements are weights of words like a document vector. Interesting words have large weights and uninteresting ones have small weights. We construct the user profile using RFB. The update fomula of RFB is denoted as

$$U = a\sum_i D_i - b\sum_j D_j \tag{2}$$

where U means a user profile, D_i means an interesting document, D_j means an uninteresting document. Both a and b are arbitary positive numbers. When we construct the user profile with documents, the weight of the word included in the interesting document increases.

2.2 Abstract of ICA and Space Transformation

In signal processing, ICA extracts independent signals from some mixed signals. When ICA is applied to speech processing, observed variables are time series data recorded by microphones and independent variables are source signals. On the other hand, when ICA is applied to documents, the inputs of microphones correspond to document vectors and the independent components are equivalent to independent topics included in the documents.

Now, we assume that m document vectors denoted as $x_1, x_2, \cdots, x_m$ are described with the combination of n unknown topics denoted as $s_1, s_2, \cdots, s_n$. Each topic vector is statistically independent and its mean is 0.

A document vector matrix X and a topic vector matrix S are denoted by equation(3).

$$\begin{aligned} X &= [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_m]^T \\ S &= [\mathbf{s}_1 \quad \mathbf{s}_2 \quad \cdots \quad \mathbf{s}_n]^T \end{aligned} \tag{3}$$

At that time, we assume X is linear combination of topic vectors.

$$X = AS \tag{4}$$

Here, A is an m n full rank mixing matrix. In addition, if the number of document is larger than that of topic, the solutions are unspecified. Thus, we assume $m \geq n$. If A is known, we can obtain the generalized inverse matrix $A^\dagger$ of A easily. However, $A^\dagger$ cannot be generally found because the mixing matrix A is unknown.

The purpose of ICA is to estimate a topic matrix S with only observed variables X under the condition where each topic is independent. In other words, ICA finds a restored signal matrix Y which is statistically independent using the restored matrix W in the following equation.

$$Y = WX \tag{5}$$

In addition, by the property of evaluation criteria, a magnitude and an order of the restored signals have not been determined uniquely.

Fast ICA[7] is one of ICA solution algorithms. This paper uses Fast ICA to find the independent components. The update criteria by hyperbolic tangent to find the independent components is equation(6).

$$\begin{aligned} w^+ &= E[Yg(w^TY)] - E[g'(w^TY)]w \\ g(u) &= tanh(u) \end{aligned} \tag{6}$$

The independent components Y obtained by ICA mean topics included in the documents. In this paper, document vectors $\mathbf{x_i}$ is mapped to the space constructed by the topics and represented with the topics. Here, we construct a user profile with $\hat{X}$ represented by topics in equation(7).

$$\hat{X} = YX^T \tag{7}$$

2.3 Abstract of MDA

Some topics obtained by ICA are unnecessary topics for information filtering which worsen accuracy of information filtering. For instance, a document about GIS system includes topics such as city plan and information retrieval. Considering with land-use plan, the topic of information retrieval is regarded as noise. In this paper, to remove those topics, we apply MDA, which does not have to decide the number of classes, to categorize similar topics and select topics.

MDA is stated in the follow steps.

Step1 Set the threshold ratio r which denotes the distance between the farthest clusters.

Step2 Set topic y_1 as cluster center Z_1.

Step3 Calculate $D_i = \min_j(y_i, Z_j)$ for $y_2, y_3, \cdots, y_n$. $D(y_i, Z_j)$ is defined as equation(8).

$$D(y_i, Zj) = \sqrt{(y_i - \bar{Z}_j)^T(y_i - \bar{Z}_j)} \tag{8}$$

Step4 Calculate $l = \max_i D_i$ and set the element of l with y_k.

Step5 If $l/MAX > r$, where $MAX = max(Z_i, Z_j)$ means the distance between the farthest classes, make new class whose center Z_{j+1} is y_k and return Step3. Otherwise, go to Step6.

Step6 Output all classes.

After classification, we extract all components of a class which has the most components. The components in the class are considered as specific topics on a theme. In other words, this selection method means to extract the detail components for a main topic.

3 Experiment and result

3.1 Experiment environment and procedure

The data for an experiment are the 625 documents concerning with information retrieval from test collection NTCIR2. These documents have already been evaluated whether each document is relevant or not. In the documents, there are 34 relevant documents.

Each document is represented as a vector with vector space model[8]. As a methodology to represent a document with a vector, at first, we apply morphological analysis tool ChaSen[9] to documents and extract nouns. After that, we remove stop words and high frequency words thorough all documents. We set the threshold of frequency with 20 documents. With the above process, we get 5,948 words and the dimension of document vector is 5,948. Using tf-idf, these words are weighted.

We apply ICA to 625 document vectors and obtain 623 independent components. The number of topics is less than the numbers of documents since some documents are dependent. Next we normalize the 623 independent components and remove the unnecessary components with MDA. It has been already mentioned in Section2.3 how to select the useful components. In consequence, we select 324 topics. After that, we converted input documents with the components selected by MDA.

We use RFB for construction of a user profile in cross validation. We provide 625 documents into 5 subsets which include 125 documents. The number of relevant documents and non-relevant ones included in each subset is showed in Table 1. We put 3 subsets together as training data for the construction of the user profile and set the others with evaluate data. Hence, we carry out experiments on 10 patterns of training data.

In RFB, to make the ratio of relevant documents and non-relevant ones set to 1:1, the coefficients, a and b, in the equation(2) are defined as

$$\begin{aligned} a &= +1 \\ b &= -N_i/N_j \end{aligned} \tag{9}$$

where N_i means the number of relevant documents and N_j means the number of non-relevant documents.

Finally, we recommend documents depending on the user profile and evaluate accuracy of recommendation with 11-point average precision ratio. The recommended documents are determined depending on the similarity S_i between the user profile and the ith document vector D_i. Similarity S_i is defined as

$$S_i = U^T D_i. \tag{10}$$

We summarize the experiment in the following steps.

Step1 Make document vectors with vector space model.

Step2 Apply ICA to document vectors

Step3 Classify the independent components with MDA and remove unnecessary components.

Step4 Transform the input documents.

Step5 Construct the user profile with RFB.

Step6 Recommend documents and evaluate the user profile with 11-point average precision ratio.

Moreover, we construct the user profile with other two methods to confirm the advantage of the proposed method, which are construction of the user profile using only RFB with original documents and ICA and RFB without topic selection.

3.2 Results

In this section, we show the result of the experiment. Figure 1 shows recall precision curves and Table 2 shows 11-point average precision ratios.

Table 1: The Input Data

@@@@	all	set1	set2	set3	set4	set5
relevant	34	7	13	5	3	6
non-relevant	591	118	112	120	122	119

Figure 1: Imputation Precision with RFB.

Table 2: 11-points Average Precision Ratio.

	Original	ICA	Euclid
0	0.458	0.463	0.638
0.1	0.420	0.432	0.535
0.2	0.325	0.368	0.525
0.3	0.301	0.346	0.459
0.4	0.270	0.320	0.381
0.5	0.257	0.288	0.325
0.6	0.238	0.237	0.250
0.7	0.152	0.198	0.165
0.8	0.095	0.160	0.122
0.9	0.080	0.103	0.106
1	0.248	0.085	0.092
Average	0.248	0.273	0.327

4 Discussion

From the Figure 1, the precision ratio of ICA version becomes better than one of original version. This is the reason why the concretization of topics contributes the improvement of precision ratio. The proposed method which selects topics with MDA also left much better result than only ICA. Especially, the improvement is clear at low recall ratio points. We cannot clearly recognize some precision ratios at high recall ratio points to become better in Figure 1. Therefore, comparing the result from the viewpoints of 11-points average ratio in Table 2, it is clearly found that the proposed method can improve the precision ratio. This is because the unnecessary components for information filtering are removed with selecting components.

5 Conclusion

In this paper, we proposed a new method to select necessary topics for information filtering using the MDA and confirmed the advantage of the proposed method. This gives that the accuracy of information filtering can be improved with removing unncecessary topics.

Since we deal with the one topic data in this experiment, we will have to extend the data including multiple topic in future.

References

[1] Deerwester.S, Dumais.T, Landauer.T, Furnas.W, Harshman.A, "Indexing by Latent Semantic Analysis", *Journal of the Society for Information Science* Vol.41, No.6, pp.391–497.

[2] Ata Kabán and Mark Girolami, "Topic Separation and Keyword Identification in Document Collections : A Projection Approach", Technical Report available in http://cis.paisley.au.ck/reseach/reports/index.html.

[3] T.Kolenda and L.K.Hansen, "Independent Components in Text", Advances in Independent Component Analysis, Springer-Verlag, 2000.

[4] Tou,J.T. and Gonzalez,R.C., "Pattern Recognition Principles", Addison-Wesley, Reading, MA. 1974.

[5] J.Rocchio, "Relevance feedback in information retrieval" , The SMART Retrieval System Experiments in Automatic Document Processing, pp. 313–323,1971

[6] "NTCIR2", NII-NACSIS Test Collection for IR System, http://research.nii.ac.jp/ntcir/index-en.html.

[7] Aapo Hyvārinen, Erkki Oja, "Independent component analysis: A tutorial", Neural Network, Vol. 13, pp. 411–430, 2000.

[8] G.Salton, M.J.McGill, "Introduction to Modern Information Retrieval", McGraw-Hill Book Company, 1983.

[9] Y.Matsumoto, "Japanese Morphological Analysis System:CHASEN", Information Science Technical Report NAIST-IS-TR97007, Nara Institute of Science Technology,1997.