A Study on the Useful Number of Dimension Reduction in Retrieving Similar Documents using Random Projection

Kayano Umezawa
Kobayashi Lab., Advanced Sciences,
Graduate School of Humanities and Sciences,
Ochanomizu University
2-1-1 Ohtsuka Bunkyo-ku, Tokyo 112-8610
Email: umezawa.kayano@is.ocha.ac.jp

Ichiro Kobayashi
Advanced Sciences,
Graduate School of Humanities and Sciences,
Ochanomizu University
2-1-1 Ohtsuka Bunkyo-ku, Tokyo 112-8610
Email: koba@is.ocha.ac.jp

Abstract—Recently, it has become able to access to the huge amount of documents through the Internet, fine document retrieval technologies have been required. In this study, we especially discussed a retrieval technology that can retrieve similar documents. In similar document retrieval technology, each document is firstly analyzed by morphological analysis and is expressed with a document vector consisting of the words of the document with the values of tfidf index which expresses the important degree of the words in the document. The similarity between documents is estimated with cosine similarity, however, the number of the dimension of a document vector corresponds to that of the words included in the document, therefore, in general the number of the dimension would become considerably large, as the number of the target documents increases. Since it is quite hard to obtain cosine similarity between high dimensional vectors because of the expensive cost of calculation. Therefore, in this study, we attempt to raise the speed of retrieving similar documents keeping its accuracy as much as possible by introducing random projection in vector calculus. Furthermore, when reducing the dimension of a document vector, we have conducted an experiment to verify the useful number of dimensions reduction for raising retrieval speed keeping its accuracy as much as possible. We show the result of the experiment and discuss it.

I. INTRODUCTION

Recently, it has been able to access to the huge amount of documents through the Internet, as the number of documents has increased, fine document retrieval technologies have been required. As one of the retrieval methods, a similar document retrieval method has become an important technique. For example, researchers need to find papers similar to their study and companies need to find patent documents similar to their own patents. As for document retrieval techniques, documents are usually expressed with document vectors and they can be retrieved by calculating the similarity between the document vectors. However, since the number of the dimension of a document vector corresponds to that of the words included in the document, therefore, in general the number of the dimension would be considerably large as the number of the target documents increases.

Since it is quite hard to calculate high dimensional vectors as they are, therefore, we need a method to reduce their dimensions. Several methods to reduce the dimension of a vector have so far been studied: for example, LSI (Latent Semantic Indexing) [2][3][4], PCA (Principal Component Analysis) and pLSI (Probabilistic LSI) which can reduce the number of dimensions based on the latent cooccurrence relationship between words and documents [5]. In this study, we adopted the random projection method [6] for the reduction of dimensions. The ability of reducing dimensions by the random projection method was verified as better than LSI and pLSI [7]. Furthermore, we conducted experiments to reduce the number of dimension of document vectors under various numbers of dimension and then verified the effective number of reduction for raising retrieval speed keeping its accuracy as much as possible. We show the result of the experiment and discuss it. The rest of this paper is organized as follows: section 2 presents related studies to this study. In section 3 we show a similar document retrieval method. Section 4 explains the method to reduce the dimension of document vectors. In section 5 we show a result of experiments with actual document data sets, and conclude in section 6.

II. RELATED STUDIES

As studies to reduce the dimension of document vectors using random projection method, Sasaki et al. [7] verified the case where the dimension reduction is applied to the document vector consisting of 4329 words extracted from 1033 documents, and reported that the ability of dimension reduction by the random projection method is better than LSI and pLSI. Watanabe et al. [8] reported that an efficient algorithm for the dimension reduction is realized by combining the random projection method and linear programming.

As a study on document retrieval using random projection for the dimension reduction, Ouchi et al. [9], [10] applied it to efficiently retrieve time-series texts such as news stream [9][10]. They compared the ability of dimension reduction between random projection and LSI and reported that the
random projection is better in retrieving time-series text. Furthermore, as for dimension reduction, random projection is not only applied to document retrieval but also applied to the speech feature transformation and is reported as useful for that [11].

In this study, in terms of reducing the dimension of a document vector, we consider the useful number of dimension reduction in retrieving similar documents using the random projection method realizing high-speed information retrieval keeping its accuracy as much as possible.

III. SIMILAR DOCUMENT RETRIEVAL METHOD

Figure 1 shows the outline of the similar document retrieval process. First of all, feature elements, i.e., important words, are extracted from documents and they are compiled as a document vector. Next, the cosine similarity between the document vector and a document vector provided as a query is measured for retrieving similar documents. The result is ranked in high similarity order and shown to a user.

A. Feature elements extraction

As a pre-processing for similar document retrieval, we explain a method to extract feature elements of a document.

We define the words that represent the content of the document as feature words, and the value that expresses the importance of the word as feature value. The feature words become elements of a document vector and the feature value is often expressed with tf-idf value. A document vector consists of feature words, so it usually becomes a high-dimensional vector if it is composed of quite a few documents.

B. Morphological analysis

For extracting feature values, at first, morphological analysis is applied to the target documents. Morphological analysis is the process to split a sentence into morphemes which are regarded as a minimum semantic unit of a text. In our study, we use a morphological analyzer, Mecab [13] for morphological analysis.

Among the words obtained by the morphological analysis, we choose nouns, verbs, and adjectives as feature words. Nouns, verbs, and adjectives usually express the contents of a document, and the other parts of speech such as postpositional particles are defined as stop words [14]. In our study, we use only nouns and verbs as feature words because the target documents we deal with are Japanese research articles whose contents tend to appear in both nouns and verbs. In addition, several types of nouns, e.g., suffix, number, are excluded from the candidates of feature words because they do not directly express the contents of a document.

C. tf-idf

It is necessary to decide feature values of a document vector taking account of the importance of feature words in documents, therefore, in this study, we use tf-idf method [15]. The tf-idf is the index obtained by multiplying the index of the term frequency, i.e., tf, and the index of inverse document frequency, i.e., idf and shows the degree of importance of a feature word in a document.

The tf-idf value of a feature word $t$ in a document $d$ is estimated in the following equations.

$$ tf(d, t) = \frac{n_t}{W_d} $$  

$$ idf(t) = \log \frac{N}{w_t} + 1 $$  

$$ tfidf(d, t) = tf(d, t) \times idf(t) $$

$n_t$ is the frequency of the word $t$ appeared in a document $d$ and $W_d$ is the total number of all feature words included in a document $d$. $N$ is the total number of all documents, $w_t$ is the number of documents including the word $t$, and 2 is the base of logarithm. $tfidf(d, t)$, the feature value of $t$ in documents $d$, is estimated by multiplying $tf(d, t)$ and $idf(t)$.

Furthermore, in order to take account of the influence caused by the length of a document, it is necessary to normalize the values in the document vector of a document $d$. As a method to normalize the value, we use cosine normalization. In the process of cosine normalization, at first, the norm of each document vector is calculated, and then the frequency of each term of the document is divided by the norm. The norm $\|x_d\|$ of the document vector $x_d = (x_1, x_2, \ldots, x_n)$ of a document $d$ is obtained in the following equation.

$$ \|x_d\| = \sqrt{\sum x_i} $$

Each feature value of a document vector $x_d$ of a document $d$ is divided by the value of norm $\|x_d\|$. Through this calculation, the feature values of each document are expressed in a document vector.

D. Cosine similarity

To measure a similar degree of each document, we calculate cosine similarity of the feature vector of each documents. Cosine similarity is an index of the degree of similarity often used by the text manipulation. Cosine similarity between the
feature vector \( x \) of the document \( d_1 \) and the feature vector \( y \) of the document \( d_2 \) is estimated in the following equation.

\[
s_{\text{cos}(x,y)} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sum y_i^2} \tag{5}
\]

It is judged that the larger the value of cosine similarity level is, the higher a similar degree of two document is.

IV. RANDOM PROJECTION

As a method to reduce the dimension of a document vector, we use the random projection method [6] [16][17] in our study.

The random projection is a method to reduce the number of dimensions of a vector in which a document matrix \( X \) is projected to a low-dimensional subspace of the matrix \( X \) by means of a matrix \( R \) whose elements are randomly decided. A document matrix \( X_{d \times N} \) which has \( d \) feature words and \( N \) documents is expressed with the matrix whose size is \( d \) columns and \( N \) rows. Each row of the matrix corresponds to one document. The value of the \( i \)-th column and the \( j \)-th row element \( x_{ij} \) is the normalized \( tf-idf \) value of the word \( i \) in the document \( j \). The random projection matrix \( R_{k \times d} \) is the matrix whose elements are randomly decided and is made so as its matrix size will be \( k \times d \) in order for the \( d \times N \) size matrix \( X \) to be projected to the \( k \times N (k \ll d) \) size matrix \( R \). The value of the \( i \)-th column and the \( j \)-th row element \( r_{ij} \) of a random projection matrix is decided based on the following constraint.

\[
r_{ij} = \begin{cases} 
+1 & \text{Probability } 1/6 \\
0 & \text{Probability } 2/3 \\
-1 & \text{Probability } 1/6
\end{cases} \tag{6}
\]

The dimension reduction of a \( X_{d \times N} \) matrix by means of random projection is calculated in the following equation.

\[
X_{k \times N}^{RP} = R_{k \times d} \times X_{d \times N} \tag{7}
\]

The calculation cost of this process is \( O(dkN) \) [4]. This means that as the number of dimensions is reduced, the calculation time will be shorter. The calculation cost to make a matrix by following the above constraint is estimated as \( O(kd) \), and also the actual process time does not take so long because of the relation \( k << d \).

In case of retrieving, the query vector is also projected to the low-dimension subspace of the original matrix (see, equation (8)) and then the similarity between the query and the target document is calculated.

\[
g_{k \times 1}^{RP} = R_{k \times d} \times x_{d \times 1} \tag{8}
\]

As a result of the retrieval, the ranking of similarity is output.

The error of the dimension reduction by random projection is defined by the Euclidean distance between vectors. For example, let us assume that there are two vectors \( x_1 \) and \( x_2 \) extracted from the \( d \times N \) matrix \( X \). The Euclidean distance of \( d \) dimension between \( x_1 \) and \( x_2 \) is defined as \( |x_1 - x_2| \). The Euclidean distance between \( x_1 \) and \( x_2 \) in the space where the dimension of the original matrix is reduced to \( k \) dimension by a random projection matrix \( R \) is defined in the following equation [16].

\[
\sqrt{\frac{d}{k}} |R_{1 \times d} x_1 - R_{1 \times d} x_2| \tag{9}
\]

In order to make the above equation established, the matrix \( R \) has to be an orthogonal matrix. As \( R^{T} R \) approximates to an identity matrix, the matrix \( R \) becomes close to an orthogonal matrix. The \( d \times d \) matrix \( \epsilon \), which indicates the error to the orthogonality of the matrix \( R \), is defined in the following equation.

\[
\epsilon = R^{T} R - I \tag{10}
\]

Here, the elements of \( \epsilon \) take a Gaussian distribution with the mean of 0 and the deviation of \( 1/k \). Therefore, as the number of reduction dimension \( k \) increases, the error of the Euclidean distance decreases [17].

V. EXPERIMENTS

Here, we explain our experimental environment: the detail about the data we use, and the evaluation index of the experiment, and further discuss about the effective number of dimensions reduction showing the experiment results and considering the result of similar document retrieval.

A. Experimental environment

Table I shows the constitution of a PC used in the experiment.

<table>
<thead>
<tr>
<th>Table I</th>
<th>EXPERIMENTAL ENVIRONMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>Mac OS X 10.6.4</td>
</tr>
<tr>
<td>CPU</td>
<td>2.53 GHz Intel Core 2 Duo</td>
</tr>
<tr>
<td>Memory</td>
<td>4 GB 1067 MHz DDR3</td>
</tr>
</tbody>
</table>

B. Evaluation method

The precision ratio and recall ratio are two widely used statistical indices for evaluating information retrieval performance. The precision ratio is an index showing the retrieval accuracy of how many objective documents are included in the retrieved documents. It is expressed in the equation (11).

\[
\text{precision} = \frac{\text{Number of objective documents among retrieved documents}}{\text{Number of retrieved documents}} \tag{11}
\]

The recall ratio is an index showing the retrieval completeness of how many objective documents are retrieved in all objective documents. It is expressed in the equation (12).

\[
\text{recall} = \frac{\text{Number of objective documents in retrieved documents}}{\text{Number of all objective documents in all documents}} \tag{12}
\]

There is a trade-off relationship between recall and precision.

To take account of these two indices to evaluate the performance of information retrieval, the 11-point interpolated
average precision or F-measure are usually used. In particular, as an index to evaluate a retrieval result, we adopt 11-point interpolated average precision in this study.

The 11-point interpolated average precision is measured at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0. For each recall level, we then calculate the arithmetic mean of the interpolated precision for the evaluation of a retrieval result.

We examine the ability of similar document information retrieval between the cases with and without random projection by means of 11-point interpolated average precision.

C. Data for the experiment

As for the documents used in the experiment, we used the abstract of randomly selected 3,134 documents among 339,501 Japanese research articles included in the document data set NTCIR-1 provided by a Japanese competition workshop for information retrieval, NTCIR [18]. By using NTCIR-1, we can use the information about paper ID, title, author names, abstract, keywords, and the name of academic society where the paper was published. Among these kinds of information, we use the abstract of the articles as the information which represents the contents of the articles.

In our experiment, as the result of extracting feature words from the abstract of the 3,134 documents, the number of the feature words is 9,861.

As the objective documents for similar document retrieval, we chose the documents which were published at the same academic society to the query document. This is because the documents for similar document retrieval consist of the articles which were published at a variety of academic societies; the contents of articles should vary differently in each society; and therefore if the retrieved article belongs to the same society as the query does, the article can be regarded as an objective article.

Each article is published at one of 43 academic societies and the different number of articles are published in each society – the minimum number of articles published in an academic society is 1, on the other hand, the maximum number of articles published in another society is 8523. As a query document, we used one of articles published by a particular academic societies and regarded the objective documents as the articles published at the same academic society to the query.

With five academic society document sets shown in Table II, we conducted five experiments of similar document retrieval using random projection.

<table>
<thead>
<tr>
<th>Society</th>
<th>Research subject</th>
<th>Number of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Informatics</td>
<td>17</td>
</tr>
<tr>
<td>(2)</td>
<td>Lighting</td>
<td>55</td>
</tr>
<tr>
<td>(3)</td>
<td>Visual and auditory senses technology</td>
<td>95</td>
</tr>
<tr>
<td>(4)</td>
<td>Television</td>
<td>220</td>
</tr>
<tr>
<td>(5)</td>
<td>Bacteriology and microbiology</td>
<td>511</td>
</tr>
</tbody>
</table>

D. Result

As results, each retrieval result examined by means of 11 point interpolated average precision is shown in Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, respectively. The horizontal axis of the graph indicates the number of dimensions after random projection. The number of dimensions are reduced at the level of 100, 200, 300, 400, and 500. The vertical axis indicates the value of 11 point interpolated average precision. In the graph, both retrieval results with or without random projection are displayed to compare each other.

For each experiment, we measured the retrieval time and calculated the arithmetic mean of the retrieval speed for each reduced dimension (see, Table III).

E. Discussions

Through the experiment, in general, we see that as dimension size increases, the precision also increases. This is because the error of dimension reduction by random projection decreases as the number of document vector dimensions increases. Whereas, at the same time we also see in each
result that there is a case where the precision decreases even though the number of dimensions increases: for example, in the case of 500 dimensions in Figure 2, in the case of 400 dimensions in Figure 5, etc. The reason for this is because the value of the elements in a random projection matrix are randomly selected, therefore, the result will be different in each time of experiments even though using the same data set. So, we can say that this is a theoretical characteristic of random projection. However, we think that the result will be stabilized as the precision increases as the number of dimensions increases, if we normalize the retrieval results by calculating the arithmetic mean of the precision obtained by more experiments. Furthermore, we also think the reason why the result was not stable is because the size of reduction dimensions in the experiments was big, i.e., 100, 200, 300, 400, 500, compared to the original dimension size, i.e., 9861. Therefore, there is possibility that the result will be stable if the number of dimensions reduction is changed to small, although it of course happens that retrieval speed will be slow.

As for the experiment result of each data set, we also see that the difference between the precision values with and without random projection varies in each data set, we think the reason for this is because depending on the query we chose, similarity between the query and the objective documents would be originally different. So, considering this kind of situation, we also have to consider to choose a proper query as it represents the object documents group for a precise experiment result.

Moreover, as a result the experiment, we also see that the retrieval speed with random projection is faster than that without the projection. As the number of dimensions reduced by random projection decreases, retrieval time decreases.

**VI. CONCLUSION**

In this study, we have conducted an experiment to reduce the number of document vector's dimensions by using the random projection method in similar document retrieval, and considered the effective number of dimensions reduction by the method.

As a result, in general we see that as the number of dimensions increases, precision also increases, however, there are also several cases where precision decreases although the number of dimensions increases. We concluded this experiment result as the one being influenced by the characteristic of random projection. But we expect that the result would be stabilized as the precision increases as the number of dimensions increases, if we normalize the retrieval results by calculating the arithmetic mean of the precision obtained by more experiments.

**TABLE III**

<table>
<thead>
<tr>
<th>Number of dimensions</th>
<th>Process time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9861(before dimension reduction)</td>
<td>23.374957</td>
</tr>
<tr>
<td>500</td>
<td>6.442756</td>
</tr>
<tr>
<td>400</td>
<td>5.033891</td>
</tr>
<tr>
<td>300</td>
<td>3.941183</td>
</tr>
<tr>
<td>200</td>
<td>2.473330</td>
</tr>
<tr>
<td>100</td>
<td>1.394341</td>
</tr>
</tbody>
</table>
As future works, we will increase the number of documents used in the experiment for discussing about a more general relationship between the dimension reduction and the retrieval accuracy.

REFERENCES


