Econo-ESA reduction scheme and the impact of its index matrix density

Faisal Rahutomo\textsuperscript{a,b}, Yuki Manabe\textsuperscript{b}, Teruaki Kitasuka\textsuperscript{b}, Masayoshi Aritsugi\textsuperscript{b,*}

\textsuperscript{a}State Polytechnic of Malang, Soekarno Hatta 9, Malang 65141, Indonesia
\textsuperscript{b}Kumamoto University, 2-39-1 Karokami, Chuo-ku, Kumamoto 860-8555, Japan

Abstract

Econo-ESA is an economic scheme of the explicit semantic analysis (ESA). The scheme properly decreases the ESA index matrix dimensions to achieve faster process with similar results. This paper discusses index matrix dimensional reduction schemes of econo-ESA. We did experiments with several schemes: random selection, k-means clustering, norm-based clustering, densest, and sparsest schemes. Each resulted matrix had different element values and density. Our experimental results showed that the random selection scheme, which had the nearest density to the original index matrix, gave the best results. We thus conclude that the index matrix density is an additional feature which has to be considered in econo-ESA.

Keywords: econo-ESA, matrix density.

1. Introduction

The ESA\textsuperscript{1} uses Wikipedia documents in its interpretation part. The method preprocesses Wikipedia documents to produce a term-document matrix. The ESA named it as the index matrix, which is used to transform a term vector into a concept vector. The method assumes each Wikipedia document preserves a specific concept.

The method is simple but costly by two reasons. First, ESA requires many multiplications to produce a concept vector; the method multiplies the overall index matrix by a term vector. Second, the method measures the similarity between two concept vectors, which have high dimensions. Thus, the number of documents affects the size of the index matrix and, as a result, the runtime. Anderka and Stein\textsuperscript{2} prove that the size of the matrix influences the results. The larger the size of index matrix is, the more stable results can be obtained but the longer processing time ESA spends. Meanwhile, a small index matrix is more effective than a large one, but it may give worse outcomes. We proposed econo-ESA\textsuperscript{3} to achieve faster procedure with fit performance to the ESA. Econo-ESA is a novel research direction related to the ESA, which reduces the index matrix dimensions into 50\% or 60\% of the original matrix.

\textsuperscript{*}Corresponding author. Tel.: +81-96-342-3641; fax: +81-96-342-3630.
E-mail address: faisal.polinema@gmail.com,\{y-manabe, kitasuka, aritsugi\}@cs.kumamoto-u.ac.jp

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).
Peer-review under responsibility of KES International.
dimensions. In spite of reporting good results, econo-ESA does not specify how to reduce the index matrix of the ESA.

This paper proposes several possible approaches to reduce the ESA index matrix dimensions with regard to econo-ESA requirement. The approaches are random selection, k-means clustering, norm-based clustering, densest, and sparsest schemes. Each scheme has two different features: element values and matrix density. We guess the features influence the results. In this paper, we examine the impact of the matrix density. We do not study the association strength effect of the matrix element values, because previous researchers already investigated it. We suggest the density as a new feature that should be considered when reducing the index matrix dimensions.

Our contributions in this paper are (1) we sharpen the econo-ESA proposal by considering the density of the index matrix during the reduction, and (2) we find random selection of the documents is the best approach, because it preserves the original density. Our empirical results showed the evidence of the contributions.

The remainder of this paper is organized as follows. Section 2 mentions the related work. Section 3 describes our reduction schemes proposal. Section 4 reports the empirical results of our experiments, and Section 5 concludes this paper.

2. Related work

2.1. ESA

In the beginning, ESA preprocesses Wikipedia documents into a term-document matrix $I$ with size $n \times m$, where $n$ is the number of terms and $m$ is the number of Wikipedia documents or concepts. Then, ESA transposes the matrix as $I^\top$ with size $m \times n$. The method defines it as the index matrix of the ESA, being understood as concept vectors of terms.

After preprocessing completed, ESA computes the similarity of two texts through two steps: interpretation and similarity measurement steps. In the first step, ESA transforms a term vector $x$ (with $n$ dimensions) into a Wikipedia-based concept vector $v$ (with $m$ dimensions) by multiplying the index matrix $I^\top$ with the term vector $x$. This process interprets a vector of term space into a vector of concept space. Equation (1) defines the interpretation step as:

$$v = I^\top x$$

with the following details:

$$
\begin{pmatrix}
  v_1 \\
  \vdots \\
  v_j \\
  \vdots \\
  v_m
\end{pmatrix} =
\begin{pmatrix}
  i_{1,1} & \cdots & i_{1,k} & \cdots & i_{1,n} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  i_{j,1} & \cdots & i_{j,k} & \cdots & i_{j,n} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  i_{m,1} & \cdots & i_{m,k} & \cdots & i_{m,n}
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  \vdots \\
  x_k \\
  \vdots \\
  x_n
\end{pmatrix}
$$

Each weight $v_j$ of concept $j$ in vector $v$ is defined as:

$$v_j = \sum_{k=1}^{n} i_{jk} \times x_k$$

where $x_k$ is the dimension weight of term $k$ in vector $x$, and $i_{jk}$ is the weight of concept $j$ for term $k$.

In the second step, the method measures the similarity between two concept vectors $u$ and $v$ by cosine similarity metric:

$$Sim(u, v) = \frac{u \cdot v}{||u|| ||v||}$$

2.2. Econo-ESA

Econo-ESA proposes to reduce the dimensions of the index matrix $I^\top$ by reducing the number of documents, thus reducing the dimensions of concept space. Benefit of this proposal is less multiplication, and thus it achieves faster
ESA uses TFIDF as its association strength scheme. It does not use new association strength scheme such as 6, 7, 4.

percentage (40%, 60%, 50%, and 70%) as the independent variable. In this paper, we set the reduction percentage to results the similarity measurement steps. Precision, recall, and F-score results between ESA and econo-ESA are similar; when the rows represent the documents or concepts, econo-ESA reduces the index matrix dimensions with three require-

Stein2 by a logarithmic model as shown in Figure 13. Then, econo-ESA proposes 50% reduction of the concepts still processing time. To maintain the results of the original scheme, econo-ESA models the data provided by Anderka and Stein2 by a logarithmic model as shown in Figure 13. Then, econo-ESA proposes 50% reduction of the concepts still gives the similar results and twice as fast3.

Econo-ESA gains good runtime and performance results3. The scheme is twice as fast in both interpretation and similarity measurement steps. Precision, recall, and F-score results between ESA and econo-ESA are similar; when the results are different, the gaps are small. The experiments show good Pearson’s correlation coefficient (PCC) results between ESA and econo-ESA. The scheme also prefers 60% reduction of the index matrix dimensions. It recommends the use of 50% econo-ESA (econo50) for long texts and 60% econo-ESA (econo60) for short texts.

Econo-ESA uses several ESA features to highlight the improvement of econo-ESA to the original ESA3. The features are similarity measurement, association strength, semantic interpretation process, and index matrix. Econo-ESA uses cosine measurement for similarity measurement. It does not use new similarity scheme such as 5, 6, 7. Econo-ESA uses TFIDF as its association strength scheme. It does not use new association strength scheme such as 6, 7, 4.

Econo-ESA uses ESA semantic interpretation process. It does not use new process such as 7. Econo-ESA uses Wikipedia corpus. It does not propose a new index matrix such as 5, 6, 7, 4, 8.

Econo-ESA investigation is related to the dimensions of the matrix. It decreases the dimensions of ESA index matrix appropriately. Econo-ESA reduces the cost of the processes by limiting the dimensions at the translation process. The scheme decreases the overall processing cost in both interpretation and relatedness measurement stages. Previous proposals2, 4 decrease the computational cost in the relatedness measurement step, but do not decrease the cost in interpretation step. In 7, 4, additional sorting step of concept vectors before the reduction adds the cost.

In our previous paper3, we set the reduction scheme (random selection) as the control variable, while the reduction percentage (40%, 60%, 50%, and 70%) as the independent variable. In this paper, we set the reduction percentage to 50% as the control variable while the reduction scheme (random selection, k-means clustering, norm-based clustering, densest, and sparsest) as the independent variable. Indeed, we want to investigate the impact of the reduction scheme on the results. Actually, we can choose econo50 or econo60, but this paper chose to use econo50 because the percentage is the initial econo-ESA proposal. Econo50 running time beats econo60 in most processes. Econo50 PCC is better than econo60 in 0.5 similarity threshold. Even though econo60 beats econo50 in 0.6 similarity threshold, the average difference is small. This paper uses certain reduction schemes (random selection, k-means clustering, norm-based clustering, densest, and sparsest) for econo-ESA based on the specific reduction requirements of the index matrix dimensions. Section 3 describes the requirements.

3. Index matrix reduction schemes

This section describes the reduction schemes examined in this paper for econo-ESA. We choose the following schemes based on the specific requirements of the econo-ESA. If the columns of matrix $\Gamma$ represent the terms while the rows represent the documents or concepts, econo-ESA reduces the index matrix dimensions with three require-
ments as follows. First, econo-ESA reduces the number of matrix rows and preserves the columns. This requirement cannot be fulfilled by general dimensional reduction techniques, such as principal component analysis (PCA). The techniques reduce the both numbers of rows and columns of the matrix. Second, econo-ESA sets a target amount of the desired rows. In this paper, we set it to 50% of the original matrix rows. This requirement shows the original and the target concept numbers are equally high. In the experiments of this paper, the initial concepts are 60,233 and the target concepts are 30,116. Third, the resulting matrix element values should be proportional to the original matrix. We make this factor as a control variable with equivalent values. We want to examine the matrix density effect that, to our knowledge, has never been studied before.

Based on the requirements, we propose to use these possible schemes: random selection, k-means clustering, norm-based clustering, densest, and sparsest schemes. Figure 2, Figure 3, Figure 4, and Figure 5 show schematic situations of the schemes.

- Random selection scheme selects the rows of the matrix 50% randomly, as shown in Figure 2. This scheme preserves the entire values of the elements in the selected rows.

- K-means clustering scheme uses k-means clustering algorithm with a special approach, as shown in Figure 3. This scheme only uses one step of k-means process to fulfill the econo-ESA requirements by two reasons. First, k-means algorithm is very difficult to be converged in our case. The case is high dimensional clustering
with high clusters target. Second, k-means clustering may generate clusters less than the targeted amount. In the econo-ESA requirements, the target amount is set at 50% of the original concept; the resulting number of clusters must be equal to the value of \( k \). Therefore, we only use one step of the k-means process to prevent the above two undesirable characteristics. We choose the seeds of the \( k \) clusters randomly during the initialization of the algorithm.

![Fig. 4. Norm-based clustering scheme.](image)

- Norm-based clustering scheme generates the new matrix based on the norm value of each term-vector \((i_{j1}, i_{j2}, ..., i_{jn})\), as shown in Figure 4. First, this scheme calculates and then sorts the norms of the term vectors. Then, the scheme forms the clusters based on the sorted norm values. Based on the norm sequence, document 1 is clustered with document 2, document 3 is clustered with document 4, and so on. Finally, the scheme generates the new concept elements from the average values of the original concept elements in the same cluster.

![Fig. 5. Sparsest and densest schemes.](image)

- Densest scheme chooses 50% of the original matrix rows by the row density. We define row density as the number of non-zero elements in a term vector \((i_{j1}, i_{j2}, ..., i_{jn})\). First, the scheme sorts the row based on each row density. Then, the scheme selects 50% of the rows with the highest density as the densest matrix. Figure 5 illustrates this scheme and its opposite scheme, the sparsest matrix.
- Sparsest scheme chooses 50% of the sparsest rows of the original matrix.

All the above schemes produce the same matrix form, 50% of the original concepts with the same amount of the terms. Two features vary between the schemes: the matrix element values and the matrix density. We guess both
features influence the results. Sorg and Cimiano\(^4\) have investigated the matrix element values. This paper investigates the density. We describe our conjecture as follows.

In the sparse matrix case, the possibility that variable \(v_j\) is 0 will increase if many values of \(i_{jk}\) are 0 when referring to Equation (2). Thus, the resulting concept vector has a few dimensions, despite the large dimensions of the space. This may be happened because of too many zero values. This condition reduces the possibility that two concept vectors overlap in the same dimensions. Therefore, the condition increases the possibility that both concept vectors become orthogonal with no similarity.

In contrast, a growing number of non-zero elements in the matrix will lead to more diverse values of \(v_j\). This condition increases the possibility that two concept vectors overlap in the same dimensions. Thus, the condition increases the similarity of two concept vectors, which are not equal to zero.

Based on the consideration, we expect the density of econo-ESA matrix also affects its performance. Selection of an appropriate scheme will give the best results. We assume in our previous work\(^3\) (see conclusion section, future work paragraph) the clustering scheme may deliver the better results, because it involves more sensible approach than just selecting the rows randomly.

4. Empirical evaluation

4.1. Experimental setup

In this section we examine the possibility that the density matrix of econo-ESA affects the results. Figure 6 shows our experimental setup diagram. We used Microsoft Wikipedia corpus (MSWik)\(^{10}\) as semantic interpreter part of ESA and econo-ESA. We merged train, dev, and test parts of the corpus into a single unit corpus. This corpus consists of 60,233 samples of Wikipedia documents in 2009 with 20,000 terms. ESA and econo-ESA implementations were in Perl 5.12.3 and MySQL 5.5.16. Experiments ran on a 3.4GHz Intel COREi7 PC with 8GB RAM. We built econo-ESA index matrix directly from ESA index matrix by several reduction schemes. Thus, we did not recalculate the IDF values in the process. Indeed, econo-ESA modifies the existing term-document matrix of ESA, which consists of TFIDF values. Econo-ESA reduces the number of rows of the index matrix which relate to the documents. Econo-ESA does not reduce the documents itself.

Fig. 6. Experimental setup diagram.
We used recycling test collections of Glasgow proposed in our previous work\(^\text{11}\). All test collections can be accessed from http://ir.dcs.gla.ac.uk/resources/test_collections/. We reported these test sets were fitted for semantic text similarity task\(^\text{11}\). We preprocessed all Glasgow test collections’ documents and queries into term vectors by referring the same 20,000 term library of MSWik. We did not apply stemming nor folding to the terms, then calculated their TFIDF values. ESA and econo-ESA interpreted all the test collections’ documents and queries term vectors into concept vectors. We performed five econo-ESA experiments based on the investigated schemes. Thus, we interpreted each test collection term vectors into econo-ESA concept vectors five times. Cosine similarity metric measured semantic similarity of all possible pairs between documents and queries concept vectors. We then applied two similarity threshold, 0.5 and 0.6, to the results; the results are “0” or “1”. Finally, we compared the econo-ESA and ESA results with \(PCC\) metric with Equation (4):

\[
PCC = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}
\]

where \(x\) and \(y\) are the ESA and the econo-ESA similarity values, respectively, while \(n\) is the number of data.

4.2. Results and discussion

Table 1 shows our density measurements. The matrix density is a comparison between non-zero elements with the overall matrix elements. We realized that the densities of the matrices were different from one and another as shown in Table 1. Table 1 shows the density of 50% random selection scheme is closest to the original scheme. For comparison purpose, we also tested 40% and 60% of the random selection scheme. Surprisingly, all the three random selection schemes had the similar density values. Clustering and densest schemes provided denser matrices than the original scheme, while sparsest scheme’s density is much smaller. In the following discussion, we will show that these differences affect the results. Thus, we should choose the proper scheme when reducing the index matrix dimensions.

Table 1. Index matrix density.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Matrix elements</th>
<th>Non-zero elements</th>
<th>Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% index matrix</td>
<td>1,204,660,000</td>
<td>14,202,329</td>
<td>0.0118</td>
</tr>
<tr>
<td>40% random</td>
<td>481,620,000</td>
<td>5,768,866</td>
<td>0.0120</td>
</tr>
<tr>
<td>50% random</td>
<td>602,320,000</td>
<td>7,249,574</td>
<td>0.0120</td>
</tr>
<tr>
<td>60% random</td>
<td>722,780,000</td>
<td>8,689,408</td>
<td>0.0120</td>
</tr>
<tr>
<td>50% k-means</td>
<td>602,320,000</td>
<td>12,042,467</td>
<td>0.0200</td>
</tr>
<tr>
<td>50% norm-based</td>
<td>602,320,000</td>
<td>12,836,806</td>
<td>0.0213</td>
</tr>
<tr>
<td>50% densest</td>
<td>602,320,000</td>
<td>12,738,708</td>
<td>0.0211</td>
</tr>
<tr>
<td>50% sparsest</td>
<td>602,320,000</td>
<td>1,463,621</td>
<td>0.0024</td>
</tr>
</tbody>
</table>

We tested the whole schemes of existing matrices to prove our previous conjecture described in the end of Section 3. Table 2 and Table 3 list our experimental results of 0.5 and 0.6 similarity thresholds, respectively. Econo-ESA with 50% random selection scheme is superior to the other schemes. Meanwhile, the sparsest index matrix is the worst.

We thought previously in Section 3 that dimensional reduction by the clustering schemes may provide better results. We thought the reduction of the dimensions by clustering is more reasonable than random selection. To examine the conjecture, we chose the two clustering schemes: k-means clustering and norm-based clustering. We were surprised when the results were not better than the random selection scheme (see Table 2 and Table 3). Then, to examine the conjecture that the matrix density affects the results, we tested the two extreme schemes: the densest and the sparsest matrix of the original scheme. The schemes were not better than the random selection scheme (see Table 2 and Table 3). The random selection scheme gave the best results for the entire test set, except ADI with 0.6 similarity threshold. The results also show the k-means clustering scheme was the nearest rival for the random selection scheme.

In this experiment, the association strength of the index matrix did not affect the results; the overall schemes used TFIDF. The experiments were free from the influence of matrix element values, because the values range among the schemes are comparable. Random selection, sparsest, and densest schemes preserved the original matrix element values. The clustering schemes produced the new concept vectors from the average values of two or more original
concept vectors inside the same cluster; the average of several TFIDF values was comparable to TFIDF. Therefore, the only factor affecting in the experiments was the matrix density. We did not recalculate the IDF values. In this experiment, econo-ESA directly modified the existing term-document matrix of ESA, which consisted of MSWik TFIDF values.

Why did the random selection scheme become the best among the five schemes? Because it had the nearest density to the original scheme. We calculated the matrix density differences between the schemes and the original scheme. The results were 0.0002, 0.0082, 0.0095, 0.0093, and -0.0094 for random selection, k-means clustering, norm-based clustering, densest, and sparsest schemes, respectively. This results were surprisingly similar to the average PCC results (see Figure 7). The first rank was random selection scheme, then followed by k-means clustering scheme, densest scheme, and norm-based clustering scheme.

ADI and Time results showed the different trend to the average value in 0.6 similarity threshold (see Table 3). For both test collections, k-means clustering scheme results were better than random selection scheme results. Both are tiny test collections; ADI contains 82 documents and 35 queries, and Time contains 423 documents and 83 queries. Small differences between ESA and econo-ESA results yielded a big PCC decrement. ADI was the most affected test collection by the similarity threshold changes (see Table 2 and Table 3).

The results showed a denser matrix to the original would probably be preferred to a sparser matrix. K-means clustering, norm-based clustering, and densest schemes results were much better than the sparsest scheme. The differences among norm-based clustering, densest, and sparsest schemes were similar, but the signs were different. Norm-based clustering and densest schemes differences were positive, while the sparsest scheme difference was negative. Thus, the sparsest scheme was the most severe scheme, because of the big and minus density difference. As discussed in Section 3, the zero values of \(i_{jk}\) in Equation (2) increase the possibility that two concept vectors become orthogonal with no similarity.
5. Conclusion

This paper examined several schemes for econo-ESA proposal. The schemes were random selection, k-means clustering, norm-based clustering, densest, and sparsest schemes. This paper also discussed an additional factor to be considered in the econo-ESA proposal: the index matrix density. The experimental results showed the random selection of concepts in econo-ESA proposal was the best approach in 50% decremental percentage. Random selection achieved the best results because it preserved the original matrix density. Our previous conjecture that the reduction of the matrix by the clustering scheme will deliver the better results was not proven.

Our experiment complemented previous reports on how the ESA method actually works. The influence factors previously reported are: the document numbers, the conformity category between the documents in the index matrix with the test set, and the selection of association strength schemes. We added in our previous report econo-ESA scheme can be faster than ESA with the similar results. Then, this paper showed the index matrix density also had the influence. Thus, with these results we complemented our previous report that in 50% decremental percentage of econo-ESA results were similar as the original scheme by reducing the index matrix concepts randomly.

For the future work, we will investigate the decremental percentages other than 50%. We consider the experiments can reveal the other factors affecting econo-ESA results. We will also investigate further the relation between matrix density and the PCC performance in future. We want to generate several artificial matrices with different density characteristics with a matrix as a base in the center. If the results are consistent with the results in this paper, we may be able to build a mathematical model on how the ESA actually works, based on the matrix density perspective.

References


