Significant Sentence Extraction by Euclidean Distance Based on Singular Value Decomposition

Changbeom Lee[1], Hyukro Park[2], and Cheolyoung Ock[1]

[1] School of Computer Engineering & Information Technology, University of Ulsan, Ulsan 680-749, South Korea
{chblee1225, okcy}@mail.ulsan.ac.kr
[2] Department of Computer Science, Chonnam National University, 300, Youngbong-dong, Buk-gu, Kwangju 500-757, South Korea
hyukro@chonnam.ac.kr

Abstract. This paper describes an automatic summarization approach that constructs a summary by extracting the significant sentences. The approach takes advantage of the cooccurrence relationships between terms only in the document. The techniques used are principal component analysis (PCA) to extract the significant terms and singular value decompostion (SVD) to find out the significant sentences. The PCA can quantify both the term frequency and term-term relationship in the document by the eigenvalue-eigenvector pairs. And the sentence-term matrix can be decomposed into the proper dimensional sentence-concentrated and term-concentrated marices which are used for the Euclidean distances between the sentence and term vectors and also removed the noise of variability in term usage by the SVD. Experimental results on Korean newspaper articles show that the proposed method is to be preferred over random selection of sentences or only PCA when summarization is the goal.

keywords: Text summarization; Principal component analysis; Singular value decomposition.

1 Introduction

Automatic text summarization is the process of reducing the length of text documents, while retaining the essential qualities of the orginal. Many search engines have tried to solve the problem of information overflowing by showing either the title and beginning of of a document. However, such the title and beginning are insufficient to decide the relevance of the documents which user wants to search, and this is the reason that the text summarization is required to resolve this problem.

The process of text summarization could consist of two phases: a document interpretation phase and a summary generation phase. The primary goal of a document interpretation phase is to find the main theme of a document and its

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 636–645, 2005.

corresponding significant words. Since the significant words collectively represent the main theme of a document, it is important to find them more reasonably. For the purpose of this doing , the word frequency of a document might be utilized[4,8]. But this approach is limited in that cooccurrence relationships among words are not considered at all. In contrast to word frequency, the other method by using WordNet or a thesaurus[2] makes good use of word relationships such as NT(narrow term), RT(related term), and so on. However such resources require a large cost to compile, and often represent too general relationships to fit a specific domain.

In this paper, we propose a new summarization approach by both principal component analysis (PCA) and singular value decomposition (SVD) that are called quantification methods or statistical analysis methods. PCA is utilized to find significant words or terms in the document by term-relationships. Since the necessary term-relationships can be acquired only from the given document by linear transformation of PCA, the proposed method need not exploit the additional information such as WordNet or a thesaurus. And the SVD is used to extract the significant sentences. After performing SVD, a sentence-term matrix is decomposed into three matrices; that is, a sentence-concentrated matrix, a term-concentrated matrix, and a singular value matrix. The distances between significant term vectors and sentence vectors can be calculated by using a sentence-concentrated matrix and a term-concentrated matrix. The shorter the distance is, the more important the sentence is. In a word, to produce the summary of a document, we first identify significant terms by the term-term relationships of being generated by PCA, and second extract the significant sentences by the distances betweeen significant term vectors and all sentence vectors.

This paper is organized as follows. Section 2 and 3 describe the way to identify the significant terms by PCA, and extract the significant sentences by SVD, respectively. Section 4 reports experimental results. A brief conclusion is given in Section 5. And this paper enlarges [7] whose main content is to find out the significant terms by PCA.

2 Significant Term Extraction by PCA

2.1 PCA Overview and Its Application

In this subsection, we will outline PCA which is adapted from [6] and which is used to extract the significant terms in the document.

PCA is concerned with explaining the variance–covariance structure through a few linear combinations of the original variables. Its general objective is data reduction and interpretation. Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \cdots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the orginal system with $X_1, X_2, \cdots, X_p$ as the coordinate axes.

PCA uses the covariance matrix (i.e., term-term correlation or cooccurrence matrix) instead of the obsevation-variable matrix (i.e., sentence-term matrix) such as Table 2. Let Σ be the covariance matrix associated with the random vector $X^T = [X_1, X_2, \cdots, X_p]$. Let Σ have the eigenvalue-eigenvector pairs (λ_1, e_1), (λ_2, e_2), ..., (λ_p, e_p) where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. The ith principal component(PC) is given by

$$Y_i = e_i^T X = e_{1i}X_1 + e_{2i}X_2 + \cdots + e_{pi}X_p, \qquad i = 1, 2, \ldots, p \tag{1}$$

The first PC is the linear combination with maximum variance, and has the widest spread in a new coordinate system geometrically. Consequently, we can say that it can cover the distribution of term frequency of a document as wide as possible, and also say that it has the power of explanation of the distribution as large as possible (but, not considering the meaning).

The PCs are uncorrelated and have variances equal to the eigenvalues of Σ, and the proportion of total population variance due to the ith PC, ρ_i, is

$$\rho_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad , \qquad i = 1, 2, \ldots, p \tag{2}$$

If most ($80 \sim 90\%$) of the total population variance, for large p, can be attributed to the first one, two, or three components, then these components can "replace" the original p variables without much loss of information. The first i PCs have also maximal mutual information with respect to the inputs among projections onto all possible i directions, and mutual information is given by $I(X, Y) = H(X) - H(X|Y)$ [5].

As we expected, all terms of the document are not necessary to represent the content (i.e., term frequency distribution) of the document by using a few first i PCs, because they have most of the total population variance and maximal mutual information as noted earlier. In addition, since PCA exploits a covariance matrix, we can use the term-term relationships by eigenvalues and eigenvectors without additional information resources. In the next subsection, we will describe how to extract the significant terms by eigenvalue-eigenvector pairs of a few first i PCs.

2.2 Extracting Significant Terms by Eigenvalue-Eigenvector Pairs

We assume that the candidates for the significant terms are confined only to nouns occurred more than 2 times in a document. We also regard the sentences as observations, the extracted nouns (terms) as variables, and the value of variables as the number of occurrence of terms in each sentence (cf. cumulative frequency of the document in [7]) .

Table 1 shows the term list extracted from one of the Korean newspaper articles composed of 12 sentences, and all of these terms have the occurrences more than twice in the document. Since the terms occurred just once are not reasonable to be representative nouns, we do not consider such terms. In our sample article, 9 terms are extracted for the cadidates for the significant ones as shown in Table 1. The sample article has 12 sentences (observations) originally, but there

Table 1. Variable (term) list

variable	notation
dae-tong-ryeong (president)	X_1
mun-je (problem)	X_2
guk-ga (nation)	X_3
sa-ram (person)	X_4
bu-jeong-bu-pae (illegality and corruption)	X_5
gyeong-je (economy)	X_6
guk-min (people)	X_7
bang-beop (method)	X_8
dan-che-jang (administrator)	X_9

Table 2. Observation-variable (sentence-term) matrix

obs. \ var.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	1	0	0	0	0	0	0	0	0
2	2	1	0	0	0	0	0	0	0
3	0	0	1	1	0	0	0	0	0
4	0	0	1	1	1	0	0	0	0
5	0	0	0	0	1	2	0	0	0
6	0	0	1	0	0	0	1	0	0
7	0	1	0	0	0	0	1	1	0
8	1	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	2

are 9 ones in sentence-term matrix as shown in Table 2. The only reason for this difference was that the sentences, which did not include the 9 extracted terms at all, were omitted. In Table 2, the column under the head X_1 shows the freqeuncy of X_1, that is, X_1 occurred once in first sentence, twice in second sentence, once in eighth sentence. In Table 3, the 9 PCs are obtained after performning PCA with the 9 variables. The column under the head PC_1 shows the eigenvector of the first PC, for instance, $\overrightarrow{PC_1} = (-0.713, -0.417, \cdots, 0.025, 0.111)$. Its eigenvalue is 0.871, and its proportion of total population is 32.34% computed by Eq. (2).

There are two steps to extract the significant terms by eigenvales and eigenvectors as shown in Table 3. First we need to decide how many PCs are selected, and second to find out how to express the each selected PC. In order to select a few most salient PCs, we can make good use of cumulative ratio of eigenvalues. For example, the first four PCs can justify more than 90% (91.28%) of the total sample variance, we can choose them without much loss information. In other words, sample variance is summarized very well by these four PCs and the data from 9 observations on 9 variables can be reasonably reduced to 9 observations on 4 PCs. Until now, we could not know what the selected PCs represent exactly, but we could describe them by their coefficients approximately. A PC can be represented by linear combination of variables multiplied by their respective coefficient. For instance,

Table 3. Eigenvector and corresponding eigenvalue of each PC

var. \ PC	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9
X_1	-0.713	0.220	0.068	-0.334	0.131	0.300	0.035	-0.467	0.000
X_2	-0.417	-0.003	0.010	0.337	-0.596	0.044	0.518	0.295	0.000
X_3	**0.303**	0.151	-0.490	-0.083	0.144	0.289	0.529	-0.139	-0.485
X_4	0.234	0.149	-0.333	-0.292	-0.498	-0.248	0.074	-0.421	0.485
X_5	0.278	0.268	0.212	-0.096	-0.396	0.726	-0.309	0.134	0.000
X_6	0.281	**0.337**	**0.710**	0.135	0.076	-0.151	0.402	-0.310	0.000
X_7	0.038	-0.111	-0.181	**0.653**	0.258	0.375	0.038	-0.288	0.485
X_8	0.025	-0.464	0.092	0.236	-0.358	-0.028	-0.236	-0.548	-0.485
X_9	0.111	-0.703	0.234	-0.416	0.050	0.267	0.362	0.043	0.243
eigenvalue(λ_i)	0.871	0.676	0.563	0.349	0.134	0.049	0.035	0.017	0.000
cumulative ratio(%)	32.34	57.42	78.31	**91.28**	96.25	98.07	99.38	100.00	100.00

$$PC_1 = -0.713 \times X_1 - 0.417 \times X_2 + 0.303 \times X_3 + \cdots + 0.111 \times X_9 \quad (3)$$

Since the coefficients represent the degree of relationship between variables and a PC, the variables with coefficient higher than 0.5 can be reasonably used to express the PC. When the PC has not such coefficient higher than 0.5, the variable with the highest coefficient can be also used to represent the PC. For example, in table 3, the correlation coefficient between PC_1 and X_3 is 0.303, so X_3 can be selected for the description of PC_1. As the most part variance of a document justified by some of the PCs, we selected variables which have a strong correlation (≥ 0.5 *or highest*) with one of these PCs as significant terms. In our example, the extracted significant terms from PC_1 to PC_4 are X_3, X_6 and X_7.

3 Significant Sentence Extraction by SVD

3.1 SVD Overview and Its Application

We will give an outline of SVD adapted from [3,9] and how to make good use of extracting the significant sentences.

Let A be any rectangular matrix, for instance an $S \times T$ matrix of sentences and terms, such as Table 2. The matrix A can be written as the product of an $S \times R$ column-orthogonal matrix U, an $R \times R$ daigonal matrix W with positive or zero elements (i.e., the singular values), and the transpose of a $T \times R$ orthogonal matrix V. Here, R is the rank of the matrix A ($R \leq min(S,T)$). The SVD decompositon is shown in Eq. (4).

$$A = U \cdot W \cdot V^T \quad (4)$$

where $U^T U = I$, $V^T V = I$, and W is the diagonal matrix of singualr values. In contrast to our usage of SVD, [3] used term-document matrix: our

sentence-term matrix can be regarded as the transpose of term-document matrix, since the documents can be thought of the sentences in the summarization fields.

In this regard, the matrix A can be regarded as the sentence-term matrix like Table 2, U as the sentence-concentrated matrix whose number of rows is equal to the number of rows of the matrix A, and V as the term-concentrated matrix whose number of rows is equal to the number of columns of the matrix A. Then, the sentence vector $\boldsymbol{s}_i$ is defined as $\boldsymbol{s}_i = (u_{i1}, u_{i2}, \cdots, u_{iR})$ where R is the rank of the matrix A. As before, the vector for a term $\boldsymbol{t}_j$ is represented by $\boldsymbol{t}_j = (v_{j1}, v_{j2}, \cdots, v_{jR})$. Consequently, both the sentence and term vectors can be used to calculate the distances between them.

Actually the reduced dimensionality can be used instead of the full rank, R, by the cumulative ratio of the singular values. The cumulative ratio, σ_k, can be calculated by Eq. (5). When the σ_k is more than 90%, k can be selected for the reduced dimensionality. And this is large enough to capture most of the important underlying structure in association of sentences and terms, and also small enough to remove the noise of variability in term usage.

$$\sigma_k = \frac{\sum_{i=1}^{k} w_i}{w_1 + w_2 + \ldots + w_R}, \quad k = 1, 2, \ldots, R \tag{5}$$

To extract the significant sentences, in the first step, the Euclidean distances can be computed between all the sentence vectors and the significant term vectors (not all the term vectors). In this regard, the shorter the distance is, the more important the sentence is, since the significant terms can be described as representative words of a document. In the second step, the sentences are extracted by means of these Euclidean distances, and then these are included in the summary in the order of their sequences. And the number of the included sentences is depend on the compression rate of user's need.

3.2 Extracting Significant Sentences by the Decomposed Matrices

In this subsection, we will illustrate how to extract the significant sentences by examples. In [7], the importance of each sentence is computed by repeatedly summing 1 for each occurrence of significant terms in the sentence. However, the proposed method can be regarded as more formal or reasonable, since the Euclidean distance between vectors is used to calculate the degree of importance of each sentence.

Computing the SVD of the sentence-term matrix as shown in Table 2 results in the following three matrices for U', W', V'. The matrices are reduced by the cumulative ratio of the singular values computed by Eq. (5). Since the first six singular values can justify 92.31% of the total, the 9-dimension can be reduced to 6-dimension. Thus, the sentence-concentrated matrix, U', and the term-concentrated matrix, V', can be represented only by 6-dimensional vectors. The U' and V' are the vectors for the 9 sentences and 9 terms respectively. The diagonal matrix W' shows the first six values (originally, nine values).

$$\mathbf{U}' = \begin{pmatrix} -0.289 & 0.032 & -0.086 & 0.022 & -0.195 & 0.293 \\ -0.771 & 0.063 & -0.153 & 0.033 & -0.182 & 0.094 \\ -0.015 & -0.367 & -0.019 & -0.441 & -0.209 & -0.077 \\ -0.017 & -0.577 & -0.067 & -0.366 & -0.260 & -0.344 \\ -0.006 & -0.657 & -0.189 & 0.704 & 0.118 & 0.085 \\ -0.052 & -0.269 & 0.070 & -0.363 & 0.354 & 0.767 \\ -0.280 & -0.101 & 0.320 & -0.094 & 0.750 & -0.371 \\ -0.481 & 0.031 & -0.067 & 0.011 & 0.013 & -0.198 \\ -0.094 & -0.115 & 0.904 & 0.181 & -0.340 & 0.092 \end{pmatrix}$$

$$\mathbf{W}' = \begin{pmatrix} 2.826 & 2.424 & 2.314 & 2.117 & 1.672 & 0.984 \end{pmatrix}$$

$$\mathbf{V}' = \begin{pmatrix} -0.818 & 0.078 & -0.199 & 0.047 & -0.326 & 0.288 \\ -0.542 & -0.003 & 0.043 & -0.024 & 0.348 & -0.483 \\ \mathbf{-0.030} & \mathbf{-0.500} & \mathbf{-0.007} & \mathbf{-0.553} & \mathbf{-0.069} & \mathbf{0.352} \\ -0.011 & -0.389 & -0.037 & -0.381 & -0.281 & -0.427 \\ -0.008 & -0.509 & -0.111 & 0.160 & -0.085 & -0.263 \\ \mathbf{-0.004} & \mathbf{-0.542} & \mathbf{-0.163} & \mathbf{0.666} & \mathbf{0.141} & \mathbf{0.173} \\ \mathbf{-0.118} & \mathbf{-0.153} & \mathbf{0.168} & \mathbf{-0.216} & \mathbf{0.660} & \mathbf{0.402} \\ -0.132 & -0.089 & 0.529 & 0.041 & 0.245 & -0.284 \\ -0.066 & -0.095 & 0.781 & 0.171 & -0.407 & 0.187 \end{pmatrix}$$

The Euclidean distances between the significant term vectors and the sentence vectors can be computed by above two matrices, V' and U', to extract the significant sentences. The significant term vectors are the third, sixth and seventh rows in the V'. The significant sentence by X_3, for instance, is the third sentence of the document, since the distance between them is the shortest. All the distances from X_3, X_6 and X_7 vectors are shown in Table 4. Consequently, the three significant sentences (third, fifth and sixth) can be included in the summary of our sample article. When the number of the selected sentences are less than that of user's need, the summary can be supplemented with the other sentences by their distances.

4 Experiments

We compared the proposed method with both only PCA[7] and random selection of sentences.

The [7] selected the significant sentences by the appearance of the significant terms. The [7] also exploited from one to three consecutive sentences for the observation of PCA; however, the performance of extracting the significant sentences was similar. In this paper, we used each sentence within a document for the observation of PCA.

To extract sentences randomly, first, random numbers amounting to 30% of the total number of sentences in a document were created, and then the sentences were extracted by these random numbers.

We tried out the proposed method with two ways. First, the sentences were extracted by the distances of each significant term as described in subsection

Table 4. Euclidean distances between all the sentence vectors and the significant term vectors (X_3, X_6 and X_7)

	distance		
num. of the sentence	X_3	X_6	X_7
1	0.841	0.979	0.904
2	1.144	1.210	1.201
3	**0.484**	1.209	1.062
4	0.752	1.226	1.293
5	1.321	**0.154**	1.279
6	0.668	1.260	**0.526**
7	1.317	1.322	0.821
8	1.057	1.071	1.026
9	1.289	1.342	1.341

Table 5. Evaluation result

	Method			
			PCA & SVD	
Measure	Random	PCA	All	Each
Average Precision	0.256	0.386	0.395	0.407
Average Recall	0.413	0.451	0.486	0.500
F-Measure	0.316	0.416	0.436	0.449

3.2. Second, the sentences were selected by all the distances of all the significant terms. In Table 5, "All" and "Each" denote the latter and the former, respectively.

We used 127 documents of Korean newspaper articles for the evaluation, which were compiled by KISTI(Korea Institute of Science & Technology Information). Each document consists of orginal article and manual summary amounting to 30% of the source text. We regarded the sentences within this manual summary as correct ones.

We use three measures to evaluate the methods: precision, recall, and F-measure. Let

- Count(SystemCorrect) denote the number of correct sentences that the system extracts.
- Count(SystemExtract) denote the number of sentences that the system extracts.
- Count(Correct) denote the number of correct sentences provided in the test collection.

The measures are defined respectively as follows.

$$Precision = \frac{Count(SystemCorrect)}{Count(SystemExtract)}$$

$$Recall = \frac{Count(SystemCorrect)}{Count(Correct)}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 5 shows that, by means of F-measure, the proposed method has improved the performance by about 2% $\sim$ 3.3% over only PCA, and by about 12% $\sim$ 13.3% over random selection. Table 5 also shows that the "Each" method is superior to "All". Furthermore, the performance of using PCA is better than that of using term frequency or a thesaurus [7].

5 Conclusion

In this paper, we have proposed a summarization approach that constructs a summary by extracting sentences in a single document. The particular techniques used are PCA and SVD for extracting the significant terms and sentences respectively.

PCA can quantify the information on both the term frequency and the term-term cooccurrence in the document by the eigenvalue-eigenvector pairs. These pairs were used to find out the significant terms among the nouns in the document. In addition, these terms can be regarded as those extracted by relationships between terms in the document, since PCA exploits the variance-covariance structure.

In contrast to PCA, SVD has the inforamtion on the sentences and terms after computing it of sentence-term matrix. In this regard, we can use the decomposed matrices to calculate the distances between the sentence and term vectors, and to make an effective removal of the noise of variability in term usage.

Experimental results on Korean newspaper articles show that the proposed method is superior to the methods of both random selection and only using PCA, and that extracting sentences by the distances per each term is better performance than extracting by all the distances of all the terms.

In conclusion, the information on the cooccurrence relationships between the terms by the PCA and the vector expressions of the sentences and terms by the SVD can be helpful for the text summarization. Furthermore, the proposed methods only exploited the pattern of the statistical occurrences within a document without the additional resources like a thesaurus to find out the relationships between the terms, and the proper dimension of the vectors.

Acknowledgements

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment)

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. New York: ACM Press (1999)
2. Barzilay, R., Elhadad, M. : Using Lexical chains for Text Summarization. In: I. Mani, & M. T. Maybury (eds.): Advances in automatic text summarization. Cambridge, MA: The MIT Press (1999) 111–121.
3. Deerwester, S., Dumais, S. T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41 (6). (1990) 381–407
4. Edmundson, H. P.: New Methods in Automatic Extracting. In: I. Mani, & M. T. Maybury (eds.): Advances in automatic text summarization. Cambridge, MA: The MIT Press (1999) 23–42.
5. Haykin, S. S.: Neural networks: A comprehensive foundation. 2nd edn. Paramus, NJ: Prentice Hall PTR (1998)
6. Johnson, R. A., Wichern, D. W.: Applied Multivariate Statistical Analysis. 3rd edn. NJ: Prentice Hall (1992)
7. Lee, C., Kim, M., Park, H.: Automatic Summarization Based on Principal Component Analysis. In: Pires, F.M., Abreu, S. (eds.): Progress in Artificial Intelligence. Lecture Notes in Artificial Intelligence, Vol. 2902. Springer-Verlag, Berlin Heidelberg New York (2003) 409–413
8. Luhn, H. P.: The Automatic Creation of Literature Abstracts. In: I. Mani, & M. T. Maybury (eds.): Advances in automatic text summarization. Cambridge, MA: The MIT Press (1999) 15–21.
9. Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P.: Numerical recipes in C++. 2nd edn. New York: Cambridge University Press (1992)