Spectral properties of a matrix of correspondences between terms

Dmitry V. Bondarchuk[1]
dvbondarchuk@gmail.com

Alexander V. Martynenko[1,2]
amartynenko@rambler.ru

1 – Ural State University of Railway Transport (Yekaterinburg, Russia)
2 – Institute of Economics, The Ural Branch of Russian Academy of Sciences (Yekaterinburg, Russia)

Abstract

The semantic core is widely used for a compact representation of documents in problems of classification and intelligent search. In order to create the semantic core, various methods are available, including the singular value decomposition of a term-document matrix and the eigenvalue decomposition of a matrix of correspondences between terms. In the present paper, we investigate how lengths of documents influence the semantic core created by these methods.

Keywords: intelligent search, semantic core, term-document matrix, matrix of correspondences between terms.

1 Introduction

Vector models are a simple way to represent a document. Such models are based on the frequency of occurrence of words and do not take into account grammar, semantics or other features of texts in a natural language [1]. The vector models are often used to analyze relationships between a set of documents and terms that are contained in these documents. The relationships are presented as a large-dimensional matrix. In order to reduce the dimension of the space, some methods were proposed, such as latent semantic analysis (LSA) [2, 3]. This method operates with a term-document matrix. A slightly different method based on matrix of correspondences between terms (MCT) was used in papers [4, 5]. In the present paper, we consider some mathematical aspects of this method.

2 Latent semantic analysis and MCT

Suppose we have a collection of text documents $D = \{d_1, d_2, ..., d_m\}$ with terms from the collection $W = \{w_1, w_2, ..., w_n\}$. We present the data as the $n \times m$ matrix $X = (x_{ij})$ with $x_{ij} = tf(d_i, w_j)$, where the term frequency $tf(d, w) \in N$ shows how many times the term w occurs in the document d. The matrix X is called a term-document matrix. Rows of X correspond to the documents and columns of X correspond to the terms.

The main idea of the latent semantic analysis is as follows [3]. In view of the singular value decomposition, the matrix X can be presented in the form

In: A.A. Makhnev, S.F. Pravdin (eds.): Proceedings of the 47th International Youth School-conference "Modern Problems in Mathematics and its Applications", Yekaterinburg, Russia, 02-Feb-2016, published at http://ceur-ws.org

$$X = USV^T,$$

where U is a $m \times m$ real unitary matrix, S is a $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal and V is a $n \times n$ real unitary matrix. The diagonal elements σ_i of S are known as the singular values of X. The matrices U and V can be written such that the singular values are in descending order:

$$\sigma_1 \geq \sigma_2 \geq \ \geq \sigma_r > \sigma_{r+1} = \ = \sigma_n = 0.$$

In order to reduce the dimension we choose the critical value $\sigma^* > 0$. In the matrices U, V and S we drop the rows and columns corresponding to the singular values $\sigma_i < \sigma^*$. As a result we obtain the matrices $\widetilde{U}$, $\widetilde{V}$ and $\widetilde{S}$ respectively, which specify the matrix

$$\widetilde{X} = \widetilde{U}\widetilde{S}\widetilde{V}^T. \tag{1}$$

The matrix $\widetilde{X}$ is a low-rank approximation of the matrix X [3]. Decomposition (1) allows to build the semantic core of the documents collection D.

Another way to identify the semantic core is to employ the matrix $G = \{g_{ij}\}$, where g_{ij} is some measure of proximity between the terms w_i and w_j. The matrix G is called a matrix of correspondences between terms. If we put

$$g_{ij} = g_{ji} = \sum_{k=1}^{m} tf(d_k, w_i)tf(d_k, w_j),$$

then

$$G = X^T X.$$

Since the matrix G is a symmetric non-negative definite matrix, it can be presented in the form

$$G = TZT^T,$$

where T is a real unitary matrix and Z is a diagonal matrix with non-negative real numbers on the diagonal. The diagonal elements λ_i of Z are known as eigenvalues of G. If we truncate the matrices T and Z according to the condition $\lambda_i < \lambda^*$, we obtain

$$\widetilde{G} = \widetilde{T}\widetilde{Z}\widetilde{T}^T. \tag{2}$$

Note that decompositions (1) and (2) are equivalent, because $\lambda_i = \sigma_i^2$ and $\widetilde{T} = \widetilde{V}$ if $\lambda^* = (\sigma^*)^2$ [5].

Let us consider the matrix $Y = \{y_{ij}\}$, where

$$y_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}}.$$

The matrix Y is called a normalized term-document matrix. This matrix can be used instead of X for building of the semantic core. When all documents have the same length, the matrices X and Y yield the same result. In the opposite case, these results are significantly different from each other [5]. The natural question arises: How do the lengths of documents influence the result of LSA? We answer to this question in the next section.

3 Main results

Let us consider the collection of text documents $D = \{d_1, d_2, ..., d_m\}$ with terms from the collection $W = \{w_1, w_2, ..., w_n\}$. And let us suppose that lengths of the first k documents are Φ and lengths of the rest of the documents are ϕ. It is easily seen that the term-document matrix $X = (x_{ij})_{i=1,j=1}^{m,n}$ has the elements

$$x_{ij} = \begin{cases} \Phi a_{ij} & \text{for} \quad i = \overline{1,k}, \quad j = \overline{1,n}, \\ \phi b_{ij} & \text{for} \quad i = \overline{k+1,m}, \quad j = \overline{1,n}, \end{cases}$$

where the real numbers $\Phi, \phi, a_{ij}, b_{ij}$ satisfy the following conditions

$$a_{ij} \geq 0 \qquad \forall i = \overline{1,k}, \quad j = \overline{1,n}, \tag{3}$$

$$b_{ij} \geq 0 \qquad \forall i = \overline{k+1,m}, \quad j = \overline{1,n}, \tag{4}$$

$$\sum_{j=1}^{n} a_{ij} = 1 \qquad \forall i = \overline{1,k}, \tag{5}$$

$$\sum_{j=1}^{n} b_{ij} = 1 \qquad \forall i = \overline{k+1,m}, \tag{6}$$

$$\Phi \geq \phi \geq 1. \tag{7}$$

It is convenient to write the matrix X in the form

$$X = \Phi A + \phi B,$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kn} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \\ b_{k+1,1} & b_{k+1,2} & \dots & b_{k+1,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix}.$$

For the matrix $G = X^T X$ we obtain

$$G = \big(\Phi A + \phi B\big)^T \big(\Phi A + \phi B\big) =$$

$$\Phi^2 A^T A + \Phi\phi A^T B + \phi\Phi B^T A + \phi^2 B^T B.$$

It is obvious that $A^T B = B^T A = 0$, thus

$$G = \Phi^2 G_A + \phi^2 G_B, \tag{8}$$

where $G_A = A^T A$, $G_B = B^T B$.

The matrices G, G_A and G_B are symmetric non-negative definite matrices. We also note that

$$rank\ G_A \leq \min\big\{rank\ A^T, rank\ A\big\} = rank\ A \leq k.$$

These properties mean that all eigenvalues of the matrix G_A are non-negative real numbers and at least the smallest $n-k$ eigenvalues are equal zero.

We are interested in the influence of the numbers Φ and ϕ on eigenvalues of the matrix G under the condition Φ is sufficiently greater than ϕ. In order to investigate this influence, it is necessary to introduce additional notation. In particular, for any matrix $P = (p_{ij})_{i,j=1}^n$ we denote

$$\|P\|_E = \sqrt{\sum_{i,j=1}^{n} p_{ij}^2}.$$

The value $\|P\|_E$ is called the Euclidian norm of the matrix P. Note, that for any matrices P and Q, we have

$$\|PQ\|_E \leq \|P\|_E \|Q\|_E, \tag{9}$$

$$\|P^T\|_E = \|P\|_E. \tag{10}$$

All eigenvalues of a symmetric matrix are real, so they can be ordered. Let $\lambda_1(P) \geq \lambda_2(P) \geq ... \geq \lambda_n(P)$ be eigenvalues of the symmetric matrix P sorted in decreasing order. It is obvious that $\lambda_s(\mu P) = \mu\lambda_s(P)$ for any real μ. For our purposes we need the following theorems ([6], [7]).

Theorem 1. *Let P and Q be symmetric matrices of the same size $n \times n$, then*

$$|\lambda_s(P) - \lambda_s(Q)| \leq \|P - Q\|_E \qquad \forall s = \overline{1,n}.$$

Theorem 2. *Let P be a symmetric matrix, Q be a non-negative definite matrix and let them have the same size $n \times n$, then*

$$\lambda_s(P+Q) \geq \lambda_s(P) \quad \forall s = \overline{1,n}.$$

Using these theorems, we can estimate the influence of the numbers Φ and ϕ on eigenvalues of the matrix G.

Theorem 3. *Let (6) hold, then*

$$\Phi^2\lambda_s(G_A) \leq \lambda_s(G) \leq \Phi^2\lambda_s(G_A) + \phi^2(m-k) \qquad \forall s = \overline{1,n}. \tag{11}$$

Proof. Since the matrices G and G_A are symmetric, it follows from Theorem 1 that for any $s = \overline{1,n}$

$$|\lambda_s(G) - \lambda_s(\Phi^2 G_A)| \leq \|G - \Phi^2 G_A\|_E =$$

$$= \|\phi^2 G_B\|_E = \phi^2\|G_B\|_E = \phi^2\|B^T B\|_E. \tag{12}$$

From condition (6) we obtain

$$\|B\|_E = \sqrt{\sum_{i=k+1}^{m}\sum_{j=1}^{n} a_{ij}^2} \leq \sqrt{\sum_{i=k+1}^{m} 1} = \sqrt{m-k}.$$

This inequality and properties (9) and (10) give

$$|\lambda_s(G) - \lambda_s(\Phi^2 G_A)| \leq \phi^2\|B^T\|_E\|B\|_E = \phi^2\|B\|_E^2 = \phi^2(m-k). \tag{13}$$

Since $\phi^2 G_B$ is non-negative definite matrix, then from Theorem 2 we get for any $s = \overline{1,n}$

$$\lambda_s(G) \geq \lambda_s(\Phi^2 G_A).$$

Using estimate (13) and the last inequality we arrive at the desired estimate (11). The proof is completed.

If $\lambda_s(G_A) \neq 0$, then we can rewrite (11) in the form

$$1 \leq \frac{\lambda_s(G)}{\lambda_s(\Phi^2 G_A)} \leq 1 + \frac{\phi^2(m-k)}{\Phi^2\lambda_s(G_A)}.$$

From this it follows that for any $\varepsilon > 0$ there exists $\Phi^* = \Phi^*(m,k,\phi,A)$, such that for any $\Phi > \Phi^*$ the following estimate holds

$$1 \leq \frac{\lambda_s(G)}{\lambda_s(\Phi^2 G_A)} \leq 1 + \varepsilon. \tag{14}$$

If $\lambda_s(G_A) = 0$ (it was shown above that this is true, in particular, for $s > n-k$), then estimate (11) can be rewritten in the form

$$0 \leq \lambda_s(G) \leq \phi^2(m-k). \tag{15}$$

4 Conclusion

From estimates (14) and (15) it follows that the influence of long documents on the semantic core obtained by LSA is much stronger than the influence of short documents. In particular, if the differences in lengths are large enough, then the semantic core does not contain information from short documents. Hence it is necessary to use the normalized term-document matrix in the case when all documents must be taken into account.

References

[1] G. Salton, A. Wong, C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[2] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1999.

[3] T. Landauer, P.W. Foltz, D. Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284, 1998.

[4] D.V. Bondarchuk. Intelligent method of selection of personal recommendations, guarantees a non-empty result. *Control Systems and Information Technology*, 2(92):130–138, 2015.

[5] D.V. Bondarchuk, G.A. Timofeeva. Isolation of the semantic kernel based on the matrix of terms correspondence. *Control Systems and Information Technology*, 3.1(61):134–139, 2015.

[6] G.H. Golub, C.F. van Loan. *Matrix Computations*. Baltimore: Johns Hopkins University Press, 1996.

[7] E.F. Beckenbach, R. Bellman. *Inequalities*. New York: Springer-Verlag, 1961.