Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN=2723-1747

p-ISSN=2722-1016

Vol. 1, No. 3, pp. 103-108, 2020

Detecting Similarities in Posts Using Vector Space and Matrix

Al Fataa Waliyyul Haq[1*], Ema Carinia[1], Sudradjat Supian[1], and Subiyanto[2]

[1]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia*
[2]*Department of Marine Sciences, Faculty of Fishery and Marine Sciences, Universitas Padjadjaran, Indonesia*

Coresponding Author(s) email: fataa19001@mail.unpad.ac.id

Abstract

This study discusses the application of two linear algebraic materials, namely vector and matrix spaces. The application of the two materials is related to an article, the writing can be in the form of an article, book, and so on. The writings examined in this study use example sentences made by the author. Two materials of linear algebra, namely the vector space and the matrix are used to analyze whether there is a similarity between the writing made with other writing. As a result, vector space and matrix can be used to detect similarities in a text.

Keywords : Linear algebra, vectors, vector spaces, matrices

1. Introduction

Linear algebra is widely used in several fields, such as programming, optimization, industry, graphics, libraries (Psarras et al., 2019; Luo et al., 2018; Kirby and Mitchell, 2018). In programming, linear algebra is usually used to initialize something with algebraic variables (Phothilimthana et al., 2019). On the chart, the study of graphs related to various algebraic structures begins by introducing the idea of graphing the zero divisor of the commutative ring of unity (Das, 2017). In addition, graph theory also helps to characterize various algebraic structures by means of studying certain graphs associated to them (Das, 2016b; Dörfler et al., 2018; Sanderson et al., 2019; Zhang and Chen, 2018). Meanwhile, for libraries, linear algebra computations into efficient sequences of library calls (Barthels et al., 2019; Solomonik et al., 2017; Bousse et al., 2018). In this study, linear algebra is used to detect similarities in writing, the writing can be a paper, thesis, and others. The application of linear algebra can be useful for research purposes and can help to make it easier to write.

In general, linear algebra can be used to detect similarities in writing. This detection uses the same linear algebra concept as the working principle of search engines, namely vector and matrix spaces. One of the methods used is the vector space model. The way it works is by implementing a document or writing as a matrix, and the similarity between two documents or two matrices is expressed in terms of the angle between the two vectors. First looking for the frequency of occurrence of words in the document, then calculating the similarity with the document being compared (Sentosa, 2016).

A vector is a geometric object that has both a magnitude and a direction (Sentosa, 2016). Vector space are finite dimensional over a field $\mathbb{F}$ and $n = \dim_{\mathbb{F}}(v)$ (Das, 2016a). In addition to vector space, this study uses cosine similarity. Cosine similarity is used to measure the similarity between two vectors. As a result, in this study it can be seen that vector and matrix spaces can be used to detect similarities in a document or writing. This is the same as the previous research conducted by Sentosa (2016) and can be useful to be developed in further research.

2. Materials and Methods

2.1. Materials

The writings examined in this study use example sentences made by the author. To compare them, an example sentence in a paper is also used. The following is an example of the first sentence taken by the author in a paper:

- There are several papers both on interval matrices and on partial matrices (Rubei, 2020).
- The sentence will be compared with the second sentence made by the following author:
- There are many papers on interval matrices and other matrices to study
- There is one papers both on interval matrices and on partial matrices

2.2. Methods

Vector space, matrix, and cosine similarity are used to detect similarities in a text. The following is an explanation of the method to be used.

2.2.1. Vector

A vector is a geometric object that has a quantity and a direction. Each vector can be represented geometrically as a directed line segment in a plane or space. If drawn, the vector is denoted by an arrow $(\rightarrow)$. The magnitude of the vector is proportional to the length of the arrow and its direction coincides with the direction of the arrow. Vectors are often marked as $(\overrightarrow{AB})$. While the vector elements are written sequentially or like a one-column matrix or use the unit vector notation $\vec{\imath}$, $\vec{j}$, $\vec{k}$ (Sentosa, 2016).

A vector that has a unit length is called a unit vector. Usually the unit vector is used to define direction. To form a unit vector, a vector is divided by the length of the vector.

2.2.2. Vector Space

A vector space is a mathematical structure formed by a set of vectors, namely objects that can be added and multiplied by a number, which are called scalars. An example of a vector space is the Euclidean vector which is often used to represent physical quantities such as forces. The vector space model is a basic technique in obtaining information that can be used for research on the relevance of documents against search keywords (query) on search engines, document clarification, document grouping, information retrieval systems, and others (Sentosa, 2016).

2.2.3. Cosine Similarity

Cosine similarity is used to measure the similarity between two vectors. Cosine similarity is the result of the cosine of the angle between the two vectors. Can be formulated as follows (Sentosa, 2016).

$$sim(Q,D) = \cos\theta = \frac{Q.D}{\|Q\|\|D\|} = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n} Q_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

with,

Q = Document query

D = Test document

3. Result and Discussion

The sample sentence taken from the paper is counted the total number of words as well as the same words can be seen in Table 1.

Table 1. Same number of words in sentences taken from paper

Word	Frequency
There	1
are	1
several	1
papers	1
both	1
on	2
interval	1
matrices	2
and	1
partial	1

After that, Table 2 and Table 3 are also made for other examples as follows

Table 2. Same number of words in the second sentence

Word	Frequency
There	1
are	1
several	0
papers	1
both	0
on	1
interval	1
matrices	2
and	1
partial	0

Table 3. Same number of words in the third sentence

Word	Frequency
There	1
are	1
several	0
papers	1
both	0
on	1
interval	1
matrices	2
and	1
partial	0

Then, the three tables are converted into vector form, for the first sentence it becomes,

$$Q = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$

Meanwhile, for the second and third sentences it becomes,

$$D_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 2 \\ 1 \\ 0 \end{pmatrix}, D_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$

After that put into equation (1),

$$Q.D_1 = (1.1)+(1.1)+(1.0)+(1.1)+(1.0)+(2.1)+(1.1)+(2.2)+(1.1)+(1.0)$$
$$= 1+1+0+1+0+2+1+4+1+0$$
$$= 11$$

$$Q.D_2 = (1.1)+(1.0)+(1.0)+(1.1)+(1.1)+(2.2)+(1.1)+(2.2)+(1.1)+(1.1)$$
$$= 1+0+0+1+1+4+1+4+1+1$$
$$= 14$$

$$\|Q\| = \sqrt{1^2+1^2+1^2+1^2+1^2+2^2+1^2+2^2+1^2+1^2} = \sqrt{16}$$

$$\|D_1\| = \sqrt{1^2+1^2+0^2+1^2+0^2+1^2+1^2+2^2+1^2+0^2} = \sqrt{10}$$

$$\|D_2\| = \sqrt{1^2+0^2+0^2+1^2+1^2+2^2+1^2+2^2+1^2+1^2} = \sqrt{14}$$

$$sim = \cos\theta_1 = \frac{Q.D_1}{\|Q\|\|D_1\|} = \frac{11}{\sqrt{16}\sqrt{10}} = \frac{11}{\sqrt{160}} = \frac{11}{12.6491} = 0.8696$$

$$\theta_1 = 29.5919°$$

$$sim = \cos\theta_2 = \frac{Q.D_2}{\|Q\|\|D_2\|} = \frac{14}{\sqrt{16}\sqrt{14}} = \frac{14}{\sqrt{224}} = \frac{14}{14.9666} = 0.9354$$

$$\theta_2 = 20.7056°$$

From the results of these calculations, the two sentences that are compared with the sentences taken from the paper, each have a different angle. The angle formed between the first sentence and the second sentence is $29.5919°$. Meanwhile, the angle formed between the first sentence and the third sentence is $20.7056°$. If the angle formed between the subspaces has a small value, it means that the two sentences have a high similarity. On the other hand, if the angle formed between subspaces has a large value, it means that the two sentences have a low similarity (Sentosa 2016).

4. Conclusion

In this study, a calculation trial was carried out on sentences taken from a paper, then compared with two sentences made by the author. The calculation involves linear algebra using vectors and vector spaces. As a result, two sentences are compared with the sentences taken from the paper, each with a different angle. The angle formed between the first sentence and the second sentence is $29.5919°$. Meanwhile, the angle formed between the first sentence and the third sentence is $20.7056°$. Thus, the angle produced by the first sentence and the second sentence has a larger angle than the angle formed by the first and third sentences. This means that the first sentence and the third sentence have a higher similarity than the first sentence and the second sentence. This is in line with manual calculations that you do yourself by seeing which words are more in the sentence being compared. The first sentence and the second sentence have only the same eight words while the first sentence and the third sentence have the same nine words. So the vector space and matrix can be used to detect similarities in a sentence compared to other sentences.

References

Barthels, H., Psarras, C., & Bientinesi, P. (2020, June). Automatic generation of efficient linear algebra programs. In *Proceedings of the Platform for Advanced Scientific Computing Conference* (pp. 1-11).

Boussé, M., Vervliet, N., Domanov, I., Debals, O., & De Lathauwer, L. (2018). Linear systems with a canonical polyadic decomposition constrained solution: Algorithms and applications. *Numerical Linear Algebra with Applications*, *25*(6), e2190.

Das, A. (2017). On nonzero component graph of vector spaces over finite fields. *Journal of Algebra and Its Applications*, *16*(1), 1750007.

Das, A. (2016a). Nonzero component graph of a finite dimensional vector space. *Communications in Algebra*, *44*(9), 3918-3926.

Das, A. (2016b). Subspace inclusion graph of a vector space. *Communications in Algebra*, *44*(11), 4724-4731.

Dörfler, F., Simpson-Porco, J. W., & Bullo, F. (2018). Electrical networks and algebraic graph theory: Models, properties, and applications. *Proceedings of the IEEE*, *106*(5), 977-1005.

Kirby, R. C., & Mitchell, L. (2018). Solver composition across the PDE/linear algebra barrier. *SIAM Journal on Scientific Computing*, *40*(1), C76-C98.

Luo, S., Gao, Z. J., Gubanov, M., Perez, L. L., & Jermaine, C. (2018). Scalable linear algebra on a relational database system. *IEEE Transactions on Knowledge and Data Engineering*, *31*(7), 1224-1238.

Phothilimthana, P. M., Elliott, A. S., Wang, A., Jangda, A., Hagedorn, B., Barthels, H., & Bodik, R. (2019, April). Swizzle inventor: data movement synthesis for GPU kernels. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems* (pp. 65-78).

Psarras, C., Barthels, H., & Bientinesi, P. (2019). The Linear Algebra Mapping Problem. *arXiv preprint arXiv:1911.09421*.

Rubei, E. (2018). A generalization of Rohn's theorem on full-rank interval matrices. *Linear and Multilinear Algebra*, 1-9.

Sanderson, D. J., Peacock, D. C., Nixon, C. W., & Rotevatn, A. (2019). Graph theory and the analysis of fracture networks. *Journal of Structural Geology*, *125*, 155-165.

Sentosa, J. (2016). Aplikasi Model Ruang Vektor Dan Matriks Untuk Mendeteksi Adanya Plagiarisme." *Jurnal Aljabar Geometri*, *2*(2), 1-12.

Solomonik, E., Carson, E., Knight, N., & Demmel, J. (2017). Trade-offs between synchronization, communication, and computation in parallel linear algebra computations. *ACM Transactions on Parallel Computing (TOPC)*, *3*(1), 1-47.

Zhang, C., & Chen, T. (2018). Exponential stability of stochastic complex networks with multi-weights based on graph theory. *Physica A: Statistical Mechanics and its Applications*, *496*, 602-611.