텍스트 유사성을 위한 파라미터 및 비 파라미터 측정

존 믈랴히루, 김종남*
부경대학교 IT 융합응용공학과

Parametric and Non Parametric Measures for Text Similarity

John Mlyahilu, Jong-Nam Kim*
Department of IT Convergence and Application Engineering, Pukyong National University

요 약 인터넷상에서의 진짜 및 가짜 정보의 범람이 수많은 텍스트 분석에 대한 연구를 이끌었다. 문헌 표기 없이 타인의 저작물을 무단 복제 및 관련 없는 연구결과 조작 등이 한동안 세간의 주목을 이끌었다. 연구 분야에서 표절과 이의 대항 및 감소를 위해 다양한 도구들이 개발되었다. 본 연구에서는 코사인 유사성과 Pearson 및 Spearman 상관관계를 이용하는 파라미터 및 비 파라미터 방법을 이용하여 문장 유사성을 측정한다. 코사인 유사성과 Pearson 상관관계는 가장 높은 유사성 계수를 얻었으나 Spearman 상관관계는 낮은 유사성 계수를 보여주었다. 본 논문에서는 정상성 가정과 편향성에 의존하는 파라미터 방법들에 반하도록 비정상성 가정으로 인한 문장 유사도를 측정하는 데 있어 비 파라미터 방법들을 사용하는 것을 제안한다.

- 주제어 : 유사성 측정, 표절, 상관관계, 정상성, 파라미터 측정

Abstract The wide spread of genuine and fake information on internet has lead to various studies on text analysis. Copying and pasting others' work without acknowledgement, research results manipulation without proof has been trending for a while in the era of data science. Various tools have been developed to reduce, combat and possibly eradicate plagiarism in various research fields. Text similarity measurements can be manually done by using both parametric and non parametric methods of which this study implements cosine similarity and Pearson correlation as parametric while Spearman correlation as non parametric. Cosine similarity and Pearson correlation metrics have achieved highest coefficients of similarity while Spearman shown low similarity coefficients. We recommend the use of non parametric methods in measuring text similarity due to their non normality assumption as opposed to the parametric methods which relies on normality assumptions and biasness.

- Key Words : Similarity measure, Plagiarism, Correlation, Normality, Parametric measure

Received 03 December 2019, **Revised** 22 December 2019, **Accepted** 25 December 2019

* **Corresponding Author** Jong-Nam Kim, Dept. of IT Convergence & Application Engineering, Pukyong national University,45, Yongso-ro, Nam-gu, Busan, Korea. E-mail: jongnam@pknu.ac.kr

Ⅰ. Introduction

Text similarity measures are the fundamental tools for text analysis [1] that can be used according to the essence and need of a certain theme in a document to be examined [2]. They can purposely be used to determine word or sequence similarities for different documents. The text similarity algorithms are preferably used to determine a certain sequence and predict the next string of words expected to follow. Moreover, they play a vital role in information retrieval, text classification, topic detection, and a list is too long to mention individually.

Text similarity analysis is semantically or lexically categorized depending on the sequences they posses as detailed in [3]. Lexical similarity analysis is done with reference to a string that has common sequence from a large corpus while semantic similarity is based on strings from large networks as referred in [4]. There are several methods that are used to analyse texts similarity including cosine similarity, correlation coefficients, euclidean distance, dice coefficient and a list is too long to mention as described in [5] and [6].

Ⅱ. Related Work

Text similarity measurements have been studied for several decades due to their essence in academics, and other fields. This is influenced by the abundance and vast of either genuine or fake information in the world of technology. The existence, comparison, evaluation and evolution of similarity measures is thoroughly described in various literature including [7]. Among other scholars, [8] pointed out that Pearson correlation coefficient provides better results than cosine distance and euclidean distance measures for both English and Arabic words in a document. From the above literature review we define the parameters that are involved in text analysis as described in [3].

Let $T = t_1, t_2, t_3, \ldots, t_n$ be the terms collection fro n documents denoted by $D_1, D_2, D_3, \ldots, D_n$. For every distinct t_i, the total terms collection in the general document has $t_i \epsilon D_1 | t_i \epsilon D_2 | t_i \epsilon D_3 \ldots t_i \epsilon D_n$ terms. We represent the document as an n-dimensional vector v_D. Let $f(D, t_i)$ be the frequency of the term $t_i \epsilon T$ in the general document D. We define the vector of the document as

$$v_D = [f(D, t_1), f(D, t_2), f(D, t_3), \ldots, f(D, t_n)] \qquad (1)$$

Suppose we define strings $S_1, S_2, S_3, \ldots, S_n$ that represents document term matrix in a vector space v_D. Then we employ both parametric and non parametric methods to determine the relationship between the six documents.

The following figure describes the steps of analyzing text data with cosine similarity measure, Pearson correlation coefficient and Spearman correlation.

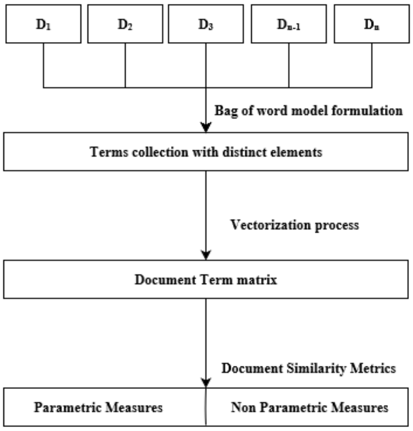
D1
D2
D3
Dn-1
Dn
Bag of word model formulation
Terms collection with distinct elements
Vectorization process
Document Term matrix
Document Similarity Metrics
Parametric Measures
Non Parametric Measures

Fig. 1. Procedure of the proposed workflow

There are so many measures of similarity for text analysis. In this study we employed three methods of which cosine similarity, pearson correlation coefficient and Spearman correlation coefficient are thoroughly discussed.

2.1 Cosine Similarity

Mathematically, cosine distance measure is an extension Euclidean distance as a basic similarity function defined as inner product. We define the inner product of two non-zero vectors D_1 and D_2 as follows:

$$inner(D_1, D_2) = \sum_{i=1}^{2} D_1 D_2 = \langle D_1, D_2 \rangle \qquad (2)$$

where each vector represents a document. Since expression (2) is unbounded, we divide the inner product by the norms to obtain the cosine similarity. Cosine similarity is bounded between zero and one for none negative D_1 and D_n.

$$\cos(D_1, D_n) = \frac{\langle D_1, D_n \rangle}{\| D_1 \| \ \| D_n \|} \qquad (3)$$

Cosine similarity is a non-invariant similarity measure that is easily affected by scale changes. We therefore introduce the scale invariant similarity measure, which is known as Pearson correlation coefficient.

2.2 Pearson Correlation

Pearson correlation (r) is used to measure the degree of the relationship between two variables. It has the assumption that the two variables are normally distributed. Let $\overline{D}_1$ and $\overline{D}_2$ be the means of each non zero vector documents D_1 and D_2 respectively, such that

$$r = cor(D_1, D_2) = \frac{\sum_{i=1}^{n} (D_1 - \overline{D_1})(D_2 - \overline{D_2})}{\sqrt{(D_1 - \overline{D_1})^2}\ \sqrt{(D_2 - \overline{D_2})^2}} \qquad (4)$$

Expression (4) can be reduced to the following expression (5) hereinafter referred to normalized cosine distance

$$r = \frac{\sum_{i=1}^{n} (D_1 - \overline{D_1})(D_n - \overline{D_n})}{\sqrt{(D_1 - \overline{D_1})^2}\ \sqrt{(D_n - \overline{D_n})^2}}$$

$$= \cos(D_1 - \overline{D_1}, D_2 - \overline{D_2}) \qquad (5)$$

Once the vectors in the Pearson correlation coefficient are normalized such that $\overline{D_1}$ and $\overline{D_2}$ are zero, then we obtain the cosine similarity as $\cos(D_1, D_2)$. Therefore Pearson correlation is also known as weighed or adjusted cosine similarity measure.

2.3 Spearman Correlation

This is a non-parametric test that determines the degree of relationship between two variables. Its measurement does not rely on the distribution of data and is an appropriate measure of relationships for scaled variables.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (d_i^2)}{n(n^2 - 1)} \qquad (6)$$

where ρ is the Spearman correlation, d_i is the difference between the ranks of the documents D_1 and D_2 and is the number of observation in the document and i is the index of the respective document.

Ⅲ. Data and Methodology

This section gives a clear description of the dataset, preprocessing methods and results visualization. For processing, we create a bag of words for each document in an independent term matrix and test for normality. We calculate the respective correlation coefficients using the parametric and non parametric measures. The data sets used are five abstracts from different international journal articles about sentiment analysis and a reference paper from the same field but not related to the selected abstracts. We used R program version 3.6 with NLP (Natural Language Processing) library for handling text and MVN library for multivariate normality testing.

3.1 Bag of word for Data Set

We define a bag of word as a multiset

representation of words without considering word order, grammar but keeping multiplicity. In this context, words of size 1 are referred to as unigrams representing a feature vector. We train our data in a bag of words model and use unigrams to determine word similarity in a corpus.

3.2 Multivariate Normality Testing

We employ Mardia’s score to test the univariate normal distribution for each independent dataset. From the test, we use the Shapiro-Wilk’s constant to identify the normality assumption for each dataset as referenced in [9]. For more justification Mardia’s and Henze-Zirker’s test is applied for determining multivariate normality test of the combined dataset. The following consecutive tables, Table 1 and Table 2 indicate that the dataset were not univariate or multivariate normally distributed, rather skewed to the right.

Table 1. Univariate distribution test results

Test	Variable	Std. Dev	p-value	Normality
Shapiro-Wilk	Jrn	0.5316	<0.001	NO
	Abs1	0.5969	<0.001	NO
	Abs2	0.5604	<0.001	NO
	Abs3	0.4567	<0.001	NO
	Abs4	0.2918	<0.001	NO
	Abs5	0.3874	<0.001	NO

Table 2. Multivariate distribution test results

Test	Statistic	p-value	Multivariate Normality
Mardia skewness	737.702	0	NO
Mardia Kurtosis	31.405	0	NO
Henze-Zirkler	13.950	0	NO

Using the tests above, the following figure is the QQ-plot for the dataset showing that the data sets were not normally distributed. The QQ-plot indicates that the majority of the data points do not lie on the percentile line.

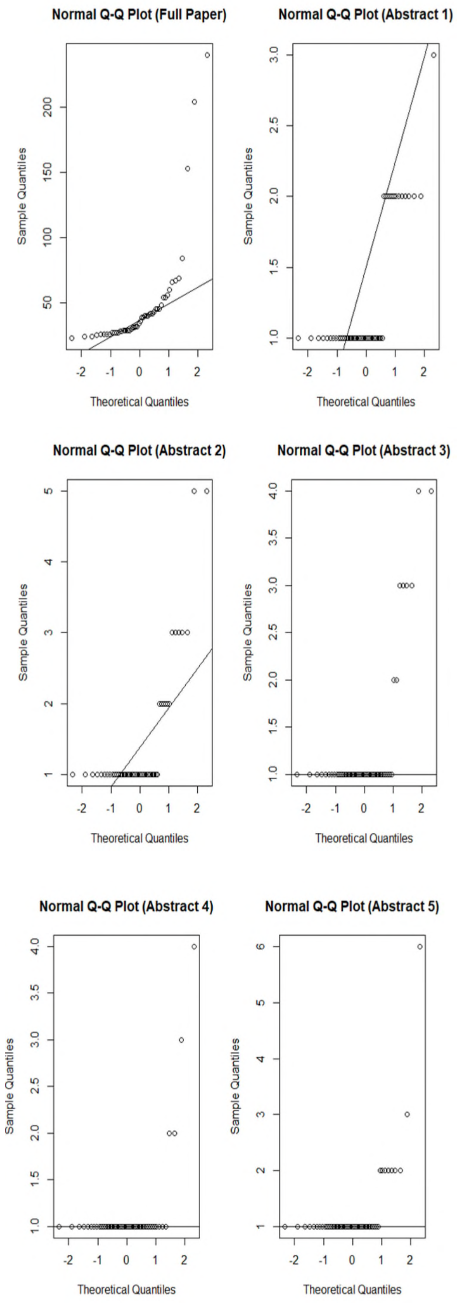

Fig. 2. QQ-Plot for non-normality distribution

As seen from figure 2, the data sets have right-skewed distribution which is not normal. We infer that, they exhibit non-normality behavior as most of the data points lie out the line for normality test. The following figure shows the distribution of the dataset justified to be not normally distributed.

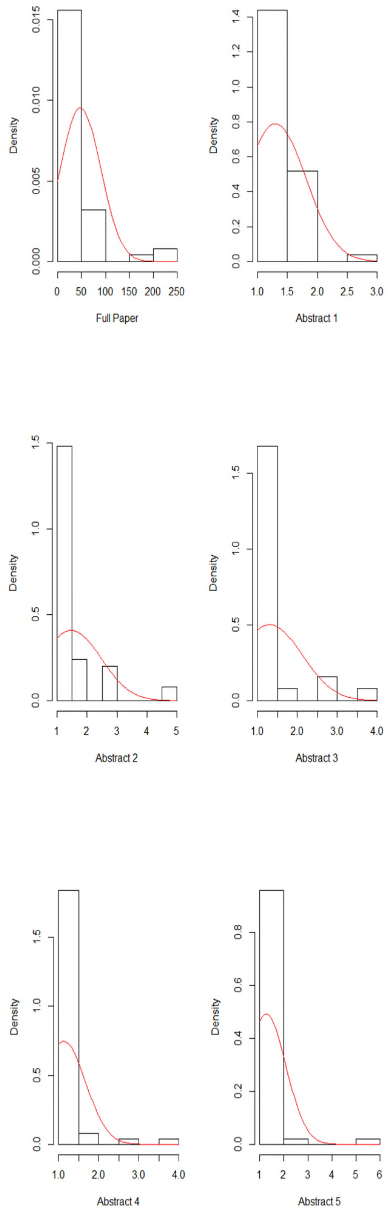

Fig. 3. Plot for non-normality distribution

Ⅳ. Experimental Results

The experimental results organized in Table 3, 4 and 5 represent cosine similarity measure, correlation coefficients for Pearson and Spearman respectively. For cosine similarity the data sets are highly related with a minimum and maximum coefficient of 0.88 and 0.97 respectively. The minimum similarity measure is between document 1 and the reference paper while the maximum similarity measure is between document 4 and document 5.

Table 3. Similarity values for cosine measurement

	P	A1	A2	A3	A4	A5
P	1.00	0.88	0.96	0.95	0.94	0.95
A1	0.88	1.00	0.96	0.93	0.94	0.93
A2	0.96	0.95	1.00	0.98	0.95	0.96
A3	0.95	0.93	0.98	1.00	0.95	0.96
A4	0.94	0.94	0.95	0.95	1.00	0.97
A5	0.95	0.94	0.96	0.95	0.97	1.00

For Pearson correlation coefficient, the documents were highly correlated with a minimum and maximum coefficient of 0.74 and 0.94 respectively. The minimum metric refers to the similarity between document 1 and the reference paper while the maximum similarity metric is between document 4 and reference paper.

Table 4. Similarity values for Pearson correlation

	P	A1	A2	A3	A4	A5
P	1.00	0.74	0.91	0.88	0.94	0.90
A1	0.74	1.00	0.86	0.72	0.60	0.74
A2	0.91	0.86	1.00	0.93	0.81	0.86
A3	0.88	0.72	0.93	1.00	0.81	0.84
A4	0.94	0.60	0.81	0.81	1.01	0.90
A5	0.90	0.74	0.86	0.84	0.90	1.00

For Spearman coefficient, the metrics are minimum and maximum at 0.47 and 0.94 scores respectively. Using the two related methods, document 4 and the reference paper are less correlated while document 3 and 5 are referred to be highly correlated.

Table 5. Similarity values for Spearman correlation

	P	A1	A2	A3	A4	A5
P	1.00	0.78	0.78	0.78	0.78	0.78
A1	0.78	1.00	0.94	0.71	0.51	0.76
A2	0.77	0.94	1.00	0.82	0.59	0.85
A3	0.64	0.71	0.82	1.00	0.72	0.94
A4	0.47	0.51	0.59	0.72	1.00	0.67
A5	0.67	0.76	0.85	0.94	0.67	1.00

Ⅴ. Conclusion

Cosine similarity measure and Pearson correlation coefficients are mathematically related that's why their results are significantly not different. The experimental results stipulate the reality by using table 3 and 4. In this work the two related methods are parametric techniques that estimates the similarity coefficients with normality assumption while Spearman correlation is a non parametric method which estimates similarity scores without normality assumption. We conclude that the related parametric measures of similarity are suitable for normally distributed texts while non parametric measures are suitable for dataset with no any model assumptions.

ACKNOWLEDGMENTS

본 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF – 2017S1A6A3A01079869).

REFERENCES

[1] L. Yuhua, M. David, B. Zuhair, O. James, and C. Keeley,“Sentence Similarity Based on Semantic Nets and Corpus Statistics,”*IEEE Trans. on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 1138-1150, 2006.

[2] K. T. Tung, N. D. Hung, and L. T. M. Hanh, “A Comparison of Algorithms used to measure similarity between two documents,” *International Journal of Advanced Research in Computer Engineering and technology*, vol. 14 no. 4, pp.1118-1121, 2015.

[3] W. Gomaa, and A. Fahmy, “A Survey of text Similarity Approaches,” *International Journal of Computer Applications*, vol. 68, no. 13, pp. 305-332, 2013.

[4] M. K. Vijaymeena, and K. Kavitha, “A Survey on Similarity Measures in Text Mining,” *Machine Learning and Applications: An International Journal*, vol. 3, no. 1, pp. 19-28, 2016.

[5] L. M. Q Abualigah, “Feature Selection and Enhanced Krill herd Algorithm for text Document Clustering,”*Springer*, ISSN 1860-949X, 2018.

[6] V. Zhelezniak, A. Savkov, A. Shen, and N. Y. Hammerla, “Correlation Coefficients and Semantic Textual Similarity,” *Annual Conference Northern American. Association for Computational Linguistics*, pp. 951-962, 2019.

[7] C. Luo, J. Zhan, L. Wang, and Q. Yang “Cosine Normalization: Using Cosine Similarity Instead of Dot Product in neural Networks,” arXiv. 1702.05870v5.

[8] S. Hajeer, "Comparison on the Effectiveness of Different Statistical Similarity Measures," *International Journal of Computer Applications*, vol. 53, no. 8, pp. 14-16, 2012.

[9] S. Korkman, D. Goksuluk, and G. Zararsiz "Multivariate Normality Tests" *The R Journal*, vol. 6, no. 2, pp. 151-162, 2014.

저자 소개

존 믈랴히루 (John Mlyahilu)

2009년 12월 : BS Mathematics, University of Dar es Salaam
2014년 2월 : MS Statistics, Pukyong National University
2018년 09월 ~ 현재 : PhD Student, Pukyong National University

김 종 남 (Jong-Nam Kim)

1997년 2월 : 광주과학기술원 정보통신공학과 졸업(공학석사)
2001년 8월 : 광주과학기술원 기전공학과 졸업(공학박사)
2001년 8월 ~ 2004년 2월: KBS 연구원
2004년 3월 ~ 현재: 부경대학 IT 융합응용공학과 교수
관심분야 : 비디오압축, 영상처리, 컴퓨터비전 등