KICSP
한국 융합신호처리학회

융합신호처리학회논문지 Vol. 23, No. 3 : 143~149 September. 2022
DOI : 10.23087/jkicsp.2022.23.3.005

Similarity Measurement Between Titles and Abstracts Using Bijection Mapping and Phi-Correlation Coefficient

John N. Mlyahilu[1], Jong–Nam Kim[1]*

[1]Division of Computer Engineering and AI, Pukyong National University

Abstract This excerpt delineates a quantitative measure of relationship between a research title and its respective abstract extracted from different journal articles documented through a Korean Citation Index (KCI) database published through various journals. In this paper, we propose a machine learning-based similarity metric that does not assume normality on dataset, realizes the imbalanced dataset problem, and zero-variance problem that affects most of the rule-based algorithms. The advantage of using this algorithm is that, it eliminates the limitations experienced by Pearson correlation coefficient (r) and additionally, it solves imbalanced dataset problem. A total of 107 journal articles collected from the database were used to develop a corpus with authors, year of publication, title, and an abstract per each. Based on the experimental results, the proposed algorithm achieved high correlation coefficient values compared to others which are cosine similarity, euclidean, and pearson correlation coefficients by scoring a maximum correlation of 1, whereas others had obtained non-a-number value to some experiments. With these results, we found that an effective title must have high correlation coefficient with the respective abstract.

- Key Words : Text mining, Machine learning, Zero-variance, Document-term-matrix, Corpora

Received 13 September 2022, Revised 28 September 2022, Accepted 30 September 2022

* **Corresponding Author** Jong-Nam Kim, Division of Computer Engineering and AI, Pukyong National University, 45, Yongso-ro, Nam-gu, Busan, Korea. jongnam@pknu.ac.kr

Ⅰ. Introduction

Title and abstract are the most vital elements that define a part of a paper, book, and an article that appears at the top and usually reads the first. A title defines the research theme whereas an abstract gives a concise summary of a research paper or an entire thesis. Without doubt, a title should basically be composed precisely. If the title is too long, consequently, it possesses too many unnecessary words whereas a short one, uses general words that can lead to ambiguity for the readers understanding [1].

On the other hand, an abstract is a brief summary of a research paper or thesis of the whole project that gives insight on what is yet done, methodology to solve the existing problem, expected and comparative results about the previous studies[2]. A detailed abstract sounds as a short summary whether published or unpublished research article, it usually has 6-7 sentences bounded at 150-250 words written to draw the gist of the paper to the reader and decide whether they can read the whole article or not[2].

However, there are some occasions where an abstract and a title of a paper mismatch, i.e., it doesn't draw the attention of the reader to keep on reading, and doesn't attract the audience even though it might have good results and strong methodologies. The situation might be caused by juniority in performing research, poor writing skills, and others alike. A confusing title can not be in line with the abstract when it has repeated words, and much more details, is too long, unspecific, noun-heavy, acronyms, filler words, jargon, hyphens, and question marks whereas an abstract might be confusing when it is too short or too long to be understood easily, an unorganized flow that causes skipping of information that could mislead readers. Additionally, an abstract containing much information from others without proof and any sign of authenticity. A poor abstract sometimes has language difficulty of a high order, it does not reflect the impact of the importance of the work and others alike.

In this work, we propose a supervised machine learning-based approach for binary problems that determine the relationship between a title and an abstract using the most occurring words in a title that corresponds with the words in an abstract. The advantage of using this metric is; it is a non-parametric statistic that measures the strength of association between two random variables. It also avoids the normality assumptions for the independent variable. Additionally, it surpasses the assumption that, for correlation between two random variables to exist, the random variable should have finite variance i.e. the second moment should exist[3-4].

This work is organized into five sections whereby the second section presents the related works, the third section presents the proposed work. The experimental results and discussion are in the fourth section while the last section describes the concluding remarks.

Ⅱ. Related Work

Analyzing and visualizing variables at a time requires extraordinary skills such that observable relationships are substantively portrayed to give insights to readers. If relationships are substantively portrayed analyses and conclusions can be easily reached when performing exploratory data analysis to determine how the variables in the dataset interact with respect to each other. There are various techniques that are used to analyze relationships numerically and visually whereby scatter-plots are the common and popular methods for analysis. They normally perform better in analyzing relationships based on covariance and correlations however, they are limited to data overlaps[4]. When the dataset has overlaps, then scatter plots are no longer useful because it is somehow difficult to determine the relationships and hence covariance metrics have to be opted for.

The famous methods for inferring relationships that rely on covariances are Pearson, Spearman, and Kendall correlation coefficients. The methodologies solve the problems associated with the quantification of relationships. They also help in quantifying the relationship if the variables have different units or different distributions. The methods normally transform each value of a random variable to a standardized one and measure how many times a data point is distant from the mean to obtain the standardized scores[5-7]. Even though the methodologies perform better for some kinds of datasets but they work well in datasets that have either positive or negative linear variations. Additionally, they work much better on interval-scaled datasets and those which happen to be normally distributed[8]. They are also limited to datasets that have much more outliers leading to impractical relationships between two or more variables. They are impractical for imbalanced datasets.

Apart from the limitations mentioned thereof, the relationship between two or more variables depends on data entries that are not unique. If the data entries are unique, the second moment existence theorem becomes violated because the covariance becomes zero and the zero variability of the unique entry points returns a 'NaN' result that hinders the determination of the relationship. This is caused by the denominator of the Pearson correlation coefficient being zero. If the standard deviation of one variable among others is zero, then the Pearson correlation coefficient does not exist. However, studies and research in statistics have introduced slack variables to datasets with no variability called jitter that can foster calculable Pearson correlation coefficients that sometimes return unreliable high correlation values using polychoric metric[9].

The other methods for measuring matrix or vector similarities include Jaccard and Cosine distances whereby the former calculates the similarity and diversity of sample sets, and the latter solves problems that are associated with high dimensionality[10]. The main disadvantage of the Jaccard distance is that; it is influenced by the size of the datasets whereby the larger the datasets the big the impact on the index it could affect the association. The main limitation of cosine distance is that it uses the normalized inner products of vectors and its magnitude is of zero importance hence when used to measure similarities the magnitude of vectors is not considered, merely their direction. This proves that the different values in two or more vectors are not taken into account.

The theorem of second-moment existence states that, the moment-based correlation metrics between two or more random variables exist when the second moment of each random variable exists. Most of the titles have had unique terms (unique elements) causing the inexistence of the second-moment. This limitation gives a way to propose a measure of similarity and association for two binary variables known as the phi-coefficient alias Mathew's Correlation Coefficient (MCC). The advantages of using MCC include; generates a high-quality score for prediction of the correctly classified instances with any balanced or imbalanced dataset[11]. Additionally, it is also very useful in determining the relationship between words, such as n-grams, searching for connectivity in graphs and networks, and determining weak and strong ties in a giant cluster, a list that is too long to mention[12]. Therefore, in this work, we propose a machine learning-based algorithm to determine the relationship between titles and abstracts because we have an imbalanced dataset i.e. the number of words in every title is not equal to the number of words in the document matrix formed by the respective abstract as mathematically described in section three.

Ⅲ. Proposed Work

The proposed algorithm stems from the fact that, the second moment correlation coefficients does not exist due to the violation of the theorem mentioned in[12]. This situation is influenced by non-variability datasets due to imbalanced groups, and unique data entries within the text data indicating that every term appears just once in a title rather than having multiple occurrences as mathematically presented here below.

Let D be a collection of documents such that, $D = D_1, D_2, D_3, \ldots, D_N$ whereby is the total number of documents in the corpus. For simplicity we label$D = \{D_i\}; \forall_{i \in N}\{D_i = (TA)_i\}$ where T represents 'Title', A represents 'Abstract', and i is the position of the abstract in the corpus. Therefore, a corpus D is represented by the following Eq. (1).

$$D = \forall_{i \in N}(T_i + \gamma)(A_i + \gamma) \tag{1}$$

where γ defines the stop words in both titles and abstracts. We define the DTM model as an algorithm that defines the importance(weights and probabilities) of every keyword in a matrix as shown in Eq. (2).

$$DTM = \{M_T M_A\}_i|(\gamma) \tag{2}$$

where $\{M_T\}_i$ and $\{M_A\}_i$ are the term-frequency matrices for titles and abstracts respectively, with exemption of stopwords. The two matrices are mathematically defined in Eq. (3) and (4) as shown here below.

$$\{M_T\}_i = tf(\forall_{i \in N}[T_i]) \tag{3}$$

$$\{M_A\}_i = tf(\forall_{i \in N}[A_i]) \tag{4}$$

where $T_i = t_1, t_2, t_3, \ldots, t_i$ and $A_i = w_1, w_2, w_3, \ldots, w_n$ are the column vector matrices with terms arranged in descending order.

We perform a 1-2-1(Bijection) mapping so that we can truncate the words that do not match the terms in titles and calculate the moment-based correlation as shown in Eq. (5).

$$r = \begin{cases} \exists r_{[-1,1]}, & -1 \leq r \leq 1 \\ otherwise, \ NaN \end{cases} \tag{5}$$

where $r = \dfrac{\sum_{i=i}^{N}(t_i - \bar{t})(w_i - \bar{w})}{\sqrt{\sum_{i=1}^{N}(t_i - \bar{t})\sum_{i=1}^{N}(w_i - \bar{w})}}$ and the term otherwise represents the condition that, there are some chances the second moment existence theorem is violated such that $\exists \sigma^2 > 0$. This situation happens only when most of the titles contain a column matrix with unique entries (terms). Then we apply a machine learning-based algorithm known as phi coefficient as shown in Eq. (6).

$$\phi = \frac{t_{ii}w_{ii} - t_{1i}w_{i1}}{\sqrt{w_{1i}t_{i1}(t_{ii} - t_{1i})(w_{ii} - w_{i1})}} \tag{6}$$

where t_{ii} and w_{ii} are the corresponding mapped frequencies for terms and words in the two corpora after the application of Bijection mapping.

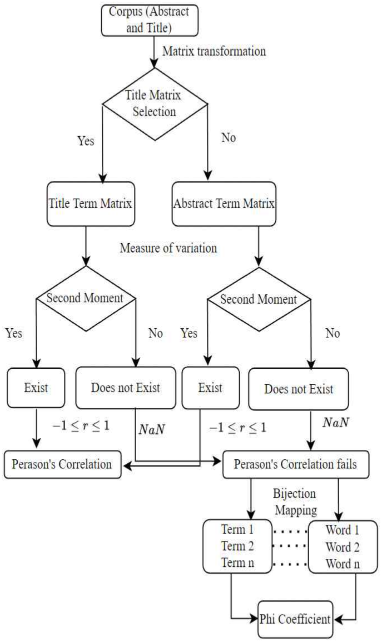

Fig. 1. Workflow of the proposed algorithm.

Additionally, we use the proposed one to compare between the existing algorithms which are used for relationship coefficient between terms in a title and words in an abstract as shown through Fig. (1).

From Fig. 1, the workflow of the proposed algorithm starts with a block named 'Corpus' which represents a collection of abstracts collected from the Korean Citation Index database repository that was published from 2018 until early 2022 regardless of specialties. The second block with 'Document Term Matrix' consists of occurrences, probabilities, and weights. The third block with 'Title Terms Matrix' and 'Abstract Term Matrix' represents two matrices the former bears all the information about every term in every title and the latter bears all information about the abstract. The fourth block with the 'Second Moment' block represents the variance values between the most frequent terms from the title matrix and words from the abstract matrix. If the second moment exists for the two matrices, then Pearson' s correlation exists for the two matrices. If the second moment does not exists then the Pearson correlation coefficient becomes 'NaN' value meaning that 'Not a Number' is obtained in block six. The seventh block allows 'Bijection Mapping' whereby the mapping between the title term matrix and abstract term matrix are matched based on the maximum number of terms in the title term matrix. In block eighth, we calculate the phi-coefficient by using the input matrix as the column matrix with unique elements to predict the corresponding elements in a column matrix for every abstract.

Ⅳ. Experimental Results

The experiments in this work were done through a gaming computer with high performance operating on windows 10. The computer is installed with R-programming language version 4.2 with built-in libraries, 'here' for setting the working environment, 'glue' for interpreting literal strings, 'tm' for text mining analytics, 'snowballc' for stemming and lemmatization of words, and 'mltools' for machine learning functions assisting for the exploratory analysis phase. The dataset was composed of abstracts and their respective titles from the KCI journals database published from 2018 to early 2022 making a total of 107. The journal articles are of broad diversity in such a way that most have no apparent relationship with others. Therefore, in this work, the proposed algorithm formulates a general corpus of all abstracts and presents the implementation from the first block to the fifth block as per the results shown in Fig. 2.

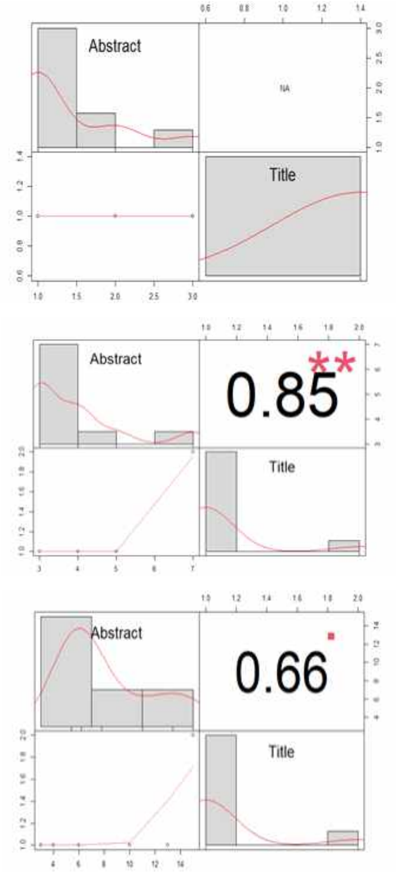

Fig. 2. Correlation coefficients obtained by Pearson.

From Fig. 2 above the results are arranged in a column presenting the relationship between the title and abstract were correlated at 66% and the sixth abstract was correlated at 85%. Among six abstracts, just two had shown similarity metrics and the rest had a ‘NaN’ relationship. This was caused by the existence of vectors with unique entry data points causing zero variability. The following table shows the title and abstract relationship based on phi-correlation coefficients for 100 sampled journal articles. For brevity purposes, we abbreviated the abstracts with their serial numbers as ‘AN’ representing the n-th abstract and N is a serial number from 10 - 64.

Table 1. Results from the proposed algorithm for 50 papers.

A10	A11	A12	A13	A14	A15	A16
1.000	1.000	0.777	0.945	0.715	0.951	0.960
A17	A18	A19	A19	A20	A21	A22
0.967	0.878	0.878	0.919	0.951	0.810	0.930
A23	A24	A25	A26	A27	A28	A29
0.650	0.843	1.000	0.967	0.967	0.960	0.945
A30	A31	A32	A33	A34	A35	A36
0.929	0.884	0.700	1.000	1.000	1.000	1.000
A37	A38	A39	A40	A41	A42	A43
0.905	0.905	0.750	1.000	0.690	0.715	1.000
A44	A45	A46	A47	A48	A49	A50
1.000	0.884	1.000	1.000	0.919	0.978	0.959
A51	A52	A53	A54	A55	A56	A57
1.000	1.000	0.777	1.000	0.828	0.884	0.715
A58	A59	A60	A61	A62	A63	A64

Table 2. Comparative results with conventional algorithms.

No	Pearson	Cosine	Euclidean	Proposed
81	NA	0.38	1.00	0.96
82	NA	1.00	1.00	0.89
83	NA	0.66	1.00	0.94
84	0.38	0.67	1.00	1.00
85	1.00	1.00	0.37	0.97
86	NA	1.00	0.62	0.63
87	NA	1.00	0.17	0.79
88	0.66	0.67	1.00	1.00
89	1.00	0.33	1.00	0.99
90	0.34	0.58	0.35	0.87
91	0.27	0.67	0.54	1.00
92	NA	0.43	0.66	0.94
93	0.59	0.89	0.71	0.87
94	0.99	1.00	0.33	0.99
95	0.53	0.69	0.67	0.92

From Table 1 and Table 2, the correlation and comparison test results indicate that the performance of our proposed algorithm is better than the rest because of its consistence in the results. Unlike other methods Pearson correlation coefficient was limited to zero variance problem and imbalanced dataset.

V. Conclusion

In this paper, we proposed a statistical machine learning-based algorithm to determine the relationship between the topic and abstract for Korean journal articles published from 2018 to early 2022. The proposed algorithm involved the creation of a corpus containing 107 abstracts, matrices for every title and abstract were treated as column vectors to allow a bijection mapping. This method is applicable where Pearson correlation fails due to the zero-variability behavior of the dataset. With such a failure, we employed a pairwise correlation coefficient to solve the zero-variability problem.

As verified in section three, our proposed algorithm solves the zero-variance problem that could limit the performance of the Pearson correlation coefficient and the rest. Our proposed algorithm also performs better for the imbalanced dataset problems as shown through the experimental part above. We avoided recommending cosine similarity because of its biasness, and it dictates the similarity by considering angles and not the magnitude of the vector itself, therefore we recommend the proposed algorithm for text similarity measurements.

ACKNOWLEDGMENTS

본 논문은 2017년 정부(교육부)의 재원으로 한국연구재단의 지원(NRF-2017S1A6A3A01079869)을 받아 수행되었습니다.

REFERENCES

[1] S. Tullu. "Write the title and abstract for research paper: Being concise, precise, and meticulous is a key" Saudi Journal of Anaesthesia, Vol. 13, No. 1, pp. 12 - 17, 2019.

[2] C. Andrade. "How to write a good abstract for a scientific paper or conference presentation," Indian journal of psychiatry, Vol. 53, No. 2, pp. 172-175, 2011.

[3] J. Soch, "The Book of Statistical Proofs," Online Edition, pp. 16, 2022.

[4] C.E. Paiva, J.P. Lima, B.S.R. Paiva. "Articles with short titles describing the results are cited more often," Clinics, Vol. 67, No. 5, pp. 509-513, 2012.

[5] C. Andrade. "How to write a good abstract for a scientific paper or conference presentation," Indian journal of psychiatry, Vol. 53, No. 2, pp. 172-175, 2011.

[6] D. Ruffell. "Writing a great abstract: tips from an Editor," FEBS Letters," Vol. 593, No. 2, pp. 141 - 143, 2019.

[7] K. Vijay. "Data Science ¦ Data Exploration," pp. 39-64, 2019.

[8] R.J. Janse, T. Hoekstra, K.J. Jager, C. Zoccali, G. Tripepi, F.W. Dekker, M. Diepen. "Conducting correlation analysis: important limitations and pitfalls," Clinical Kidney Journal, Vol. 14, No. 11, pp. 2332 - 2337, 2021.

[9] R. Aggarwal, P. Ranganathan. "Common pitfalls in statistical analysis: The use of correlation techniques," Perspectives in clinical research, Vol. 7, No. 4, pp. 187-190, 2019.

[10] J. Ekström. "A Generalized Definition of the Polychoric Correlation Coefficient," 2011.

[11] L. Zahrotun. "Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering with Shared Nearest Neighbor Method," Computer Engineering and Applications, Vol. 5, No. 1, (2016), pp. 11-18.

[12] D. Chicco, G. Jurman. "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, Vol. 21, No. 6, pp. 1 - 3, 2020.

저자소개

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

관심분야 : 멀티미디어 및 영상처리, 인공지능 등

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 컴퓨터인공진능공학부 교수

관심분야 : 머신러닝, 영상처리, 컴퓨터비젼 등