Levenshtein Distance and k-Nearest Neighbor Based COVID-19 Hot Spot Detection

Minseok Kim
Department of Electrical Engineering
Pusan National University
Busan, South Korea
minskey@pusan.ac.kr

Seongsu Park
Department of Computer Engineering
Pusan National University
Busan, South Korea
tjdtnsu@pusan.ac.kr

Qikang Deng
Department of Computer Engineering
Pusan National University
Busan, South Korea
dengqikang@pusan.ac.kr

***Abstract*—Recently, with the spread of the coronavirus disease 2019 (COVID-19) in China, the number of confirmed cases and deaths has increased rapidly. To prevent the spread of COVID-19, many countries take measures such as refusing to enter the country for foreign and imposing travel restriction. However, despite the government's efforts, the virus is likely to spread because it can be infected easily by droplets or physical contact. In addition, it is difficult to predict when the vaccine development will be completed, so it is necessary to identify the traces of the confirmed patient and avoid it in advance. Therefore, in order to prevent and avoid indirect contact with COVID-19, we propose a method to search for traces and information of confirmed patients corresponding to the region of interest in this paper. The proposed method is, it obtains province, city, gender, and birth from the user, and determines the province and region of shortest distance with the entered query (e.g. province, city) with Levenshtein distance. After that, search the traces and information of contacted persons with high similarity between the provided query and the determined province, city using kNN. This paper used data collected by the Korea Centers for Disease Control and Prevention (KCDC). As a result of the experiment, it was confirmed that the proposed method can effectively obtain the traces and related information of confirmed patients under similar conditions.**

***Keyword*—COVID-19, Levenshtein distance, k-Nearest Neighbor, Data mining**

I. Introduction

From December 2019 the first respiratory infection detected in Wuhan, China, the number of confirmed cases and deaths has increased rapidly with the spread of the coronavirus disease 2019 (COVID-19) [1]. According to National Institutes of Health (NIH), COVID-19 has asymptomatic infection and high infectivity from at least two days prior to symptom expression [2] and can infect more than 40% of all contacts before symptom expression. Accordingly, European Union (EU) take measures such as refusing to enter the country for foreign and imposing travel restriction to prevent the spread of COVID-19. The case of South Korea, the government is trying to prevent the spread of COVID-19 by blocking regional infection rather than using extreme solutions such as barring the entry of foreign. However, despite the government's efforts, the number of confirmed cases in South Korea is on the rise continually because of asymptomatic transmission characteristics of COVID-19. For example, in Daegu and Gyeongsangbuk-do,

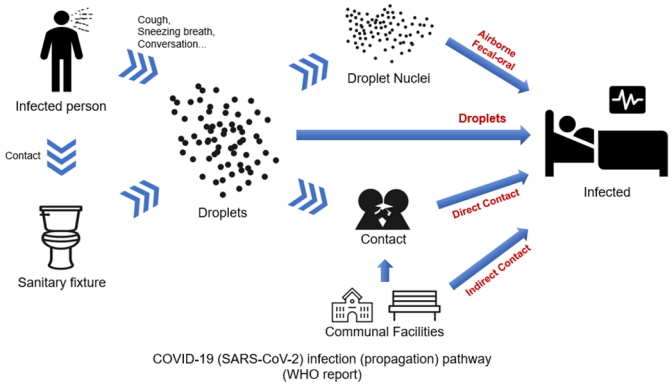

Fig. 1. Infection route of COVID-19.

thousand of group confirmers occurred in a short time through one super-spreader in the church [3, 4].

Assumptions about the infection route of COVID-19 are summarized broadly in two. The first one is propagation by large and supersize droplets (i.e. particles released during sneezing, coughing, or talking). The other one is propagation of direct contact between the hand or body [5]. In addition to the above two assumptions, the scientific community is also considering possibility of spreading faecal-oral in human feces. Although the ultimate solution for COVID-19 is completion of vaccine, it is considered that early vaccine completion is difficult. Because sufficient procedures must be performed to verify the safety of the vaccine, such as neutralizing antibodies and immune responses. Recently, identifying movement of confirmed patient work is proceeding to prevent droplet infection or direct contact by using various techniques such as machine learning or deep learning for temporary plan for vaccine development.

In this paper, we propose a searching method for provinces and places visited by COVID-19 confirmers according to queries. It applies a machine learning method, k-nearest neighbor (kNN) using data collected by Korea Centers for Disease Control and Prevention (KCDC). The procedure of the proposed method is demonstrated in Fig. 1. First, it obtains province, city, gender, and birth from the user, and determines the province and region of shortest distance with the entered

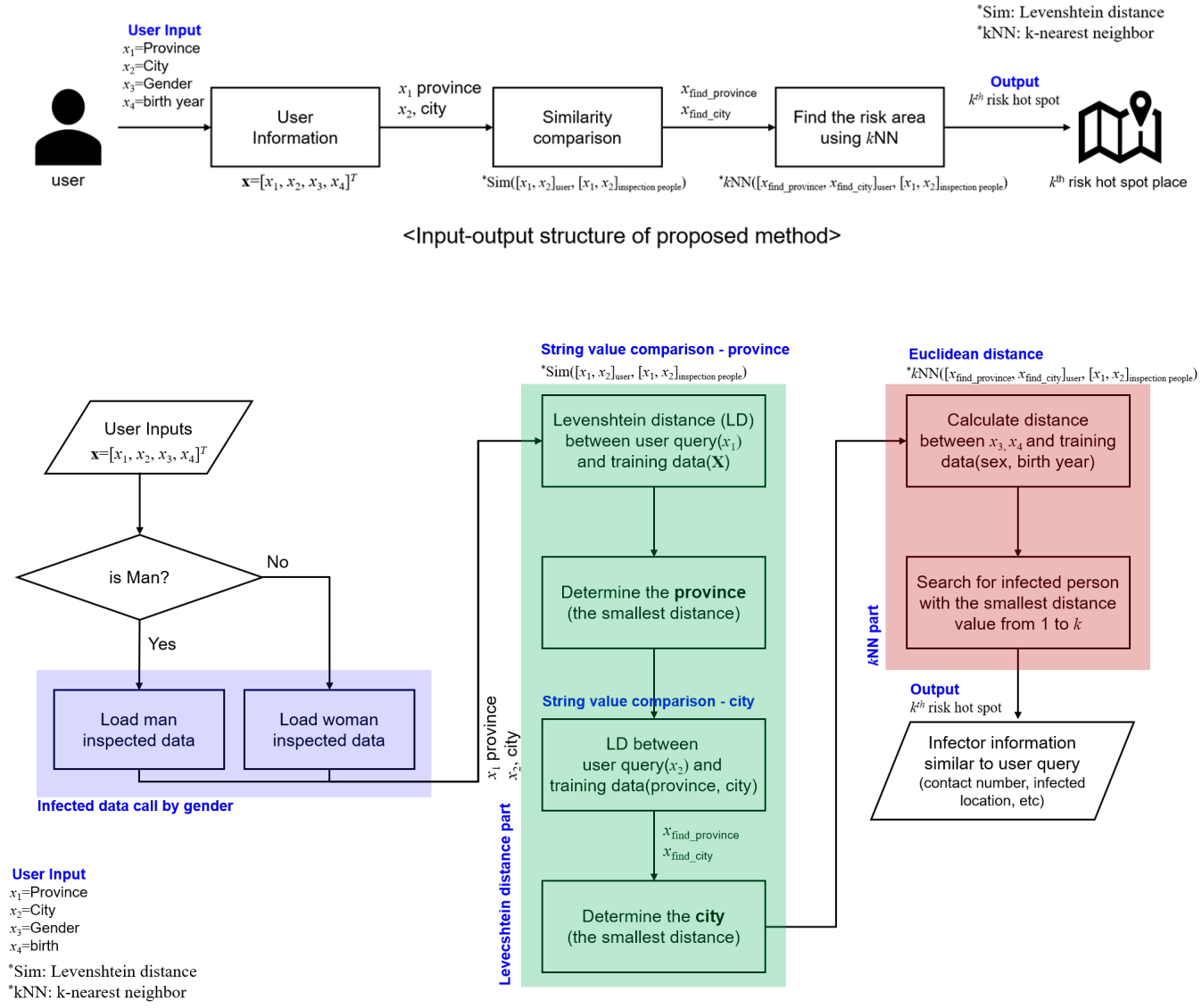

Fig. 2. Procedure of Proposed Method.

query (e.g. province, city) with Levenshtein distance. After that, select k-visited places and the numbers of contacted persons with high similarity between the provided query and the determined province, city using kNN.

kNN is a method of searching for k datas existing at the closest distance from the query, and is used not only for data mining techniques such as classification and outlier detection, but also for similar location search [6]. Especially, kNN has the advantage that it can be applied to COVID-19 confirmed patient's movement search because it can search immediately according to changes of queries.

The rest of the paper is organized as follows. In Section 2, we introduce how to search the area visited by confirmed patients using Levenshtein distance and kNN. After that, we deal with COVID-19 confirmatory data for this paper in Section 3. Experimental results will be presented in Section 4. Finally, conclusions and further investigations in Section 5.

II. Hot Spot Detection Method of Confirmed Case

A. Levenshtein distance

Levenshtein distance (LD) is representative string distance metric method to calculate distance of two strings. Levenshtein distance between two strings, s1 and s2 is defined minimal number of times of operation (i.e. insertion, deletion, replacement) to convert the string s1 to s2 [7]. Generally, LD is used to correct misspellings, can replace non-right words with the nearest right words based on the string distance when there is a dictionary value for right words. Levenshtein distance also referred to as edit distance. The process for calculating LD is detailed in Algorithm 1.

Algorithm 1 The Levenshtein Distance Algorithm

Input: Two strings: char **s[**1..m**]**, char **t[**1..n**]**
Output: Levenshtein distance: **d[m, n]**

Initialisation : int d[] = new int [0..m, 0..n]
for i = 0 to m **do**
  d[i, 0] := i
end for
for j = 0 to m **do**
  d[0, j] := j
end for
for i = 1 to m **do**
  for j = 1 to n **do**
    if (s[i] = t[j]) **then**
      cost := 0
    else
      cost := 1
    end if
    d[i, j] := Min(d[i-1, j] + 1, d[i, j-1] + 1, d[i-1, j-1] + cost)
  end for
end for
return d[m, n]

Algorithm 2 The K-Nearest Neighbors (KNN) Algorithm

Input: An arbitory dataset: **vector<char[]> dataset**
A number of nearest neighbors: **k**
A searching input we want to find it's nearest neighbors: **char s[]**
Output: k nearest neighbors: **vector <pair<int, char[]>> result**
// **d** is the Levenshtein between each patient and new searching input **s**.
Initialisation : vector <pair<int,char[]>> d
for i = 1 to $dataset.Length$ **do**
  d.push(pair<Distance(dataset[i], s), dataset[i]>)
end for
// sort the d vector according to the float value of **pair<int, char[]>**(descending sort)
vector <pair<int, char[]>> result = Sort(d)
// return the k nearest neighbors
return result[1..k]

B. k-Nearest Neighbor

Unlike machine learning methods, such as Support Vector Machine (SVM) and Bayesian inference, which classify targets by calculating targe function, k-nearest neighbor (kNN) is a method that calculates distances to a target and searches for similar k values. Since kNN can search for unclassified or duplicated data effectively, it can be applied to search for places visited by confirmed patients like the query. kNN is widely used because it can produce simple and excellent results. The process for selecting k confirmed patients from the query is followed in Algorithm 2.

TABLE I
EPIDEMIOLOGICAL DATA OF COVID-19 PATIENTS WITH LOCATION AND STATISTICAL DATA OF THE REGIONS IN SOUTH KOREA

x_i	Variable	Unit
x_1	Index	-
x_2	Patient id	-
x_3	Global num	-
x_4	Sex	male/female
x_5	Birth year	year
x_6	Age	year
x_7	Country	-
x_8	Province	-do
x_9	City	-si/-gun/-gu
x_{10}	Underlying disease	TRUE/FALSE
x_{11}	Infection case	-
x_{12}	Infection order	-
x_{13}	Infected by	-
x_{14}	Contact number	-
x_{15}	Symptom onset date	day
x_{16}	Confirmed date	day
x_{17}	Released date	day
x_{18}	Deceased date	day
x_{19}	State	-
x_{20}	Elementary school count	-
x_{21}	Kindergarten count count	-
x_{22}	University count	-
x_{23}	Academy ratio	%
x_{24}	Elderly population ratio	%
x_{25}	Elderly alone ratio	%
x_{26}	Nursing home count	-

III. EXPERIMENTAL RESULT

In this paper, We used 3,154 COVID-19 confirmed patient information data collected by KCDC. Table I shows a total of 25 process variables selected through KCDC. Among the total of 25 variables, after excluding 11 variables such as patient ID (x_1 variable) and symptom data (x_{15} variable), which are irrelevant for searching the place the confirmed patient visited, we used 14 variables (e.g. sex (x_4), birth year (x_5), and province (x_6)). Especially, since the clear location of the COVID-19 confirmer data is not provided for personal information security, we search for the confirmed patient's location similar to the query of sex(x_4), birth(x_5), and age(x_6) with kNN. In kNN, k is set to 10.

Fig. 3 shows the similarity comparison result of character strings using LD. Fig. 3 (a), (b) are LD results of province queries similar to the corresponding query, and Fig.3 (c) shows two provinces with a large distance among the provinces provided by KCDC. Derived distance value through LD is calculated in metric, it shows that the number of insert, delete, and replace operations to convert the query to the province of KCDC is derived as a distance value. The proposed method shows that applying LD for string comparison and correction of misspelling results that the target string (Busan) can be searched even if there is a similar string (Ulsan). As a result, it was confirmed that the LD used in this paper finds similar provinces even if the query is incorrect.

Fig. 4 shows the result of the distance value from the city query after determining the target province based on province query as shown in Fig 3. In other words, This is the result of joining the province and city queries. The city query

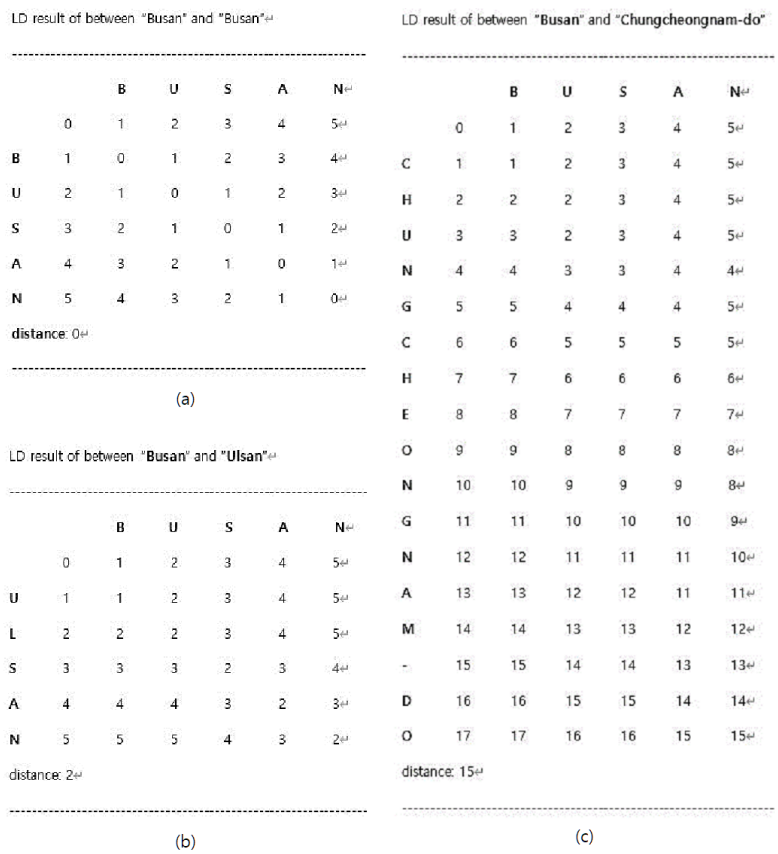

Fig. 3. LD result between query province and KCDC confirmed patient's province

(a) "Busan"–"Busan", (b) "Busan"–"Ulsan", (c) "Busan"–"Chungcheongnam-do".

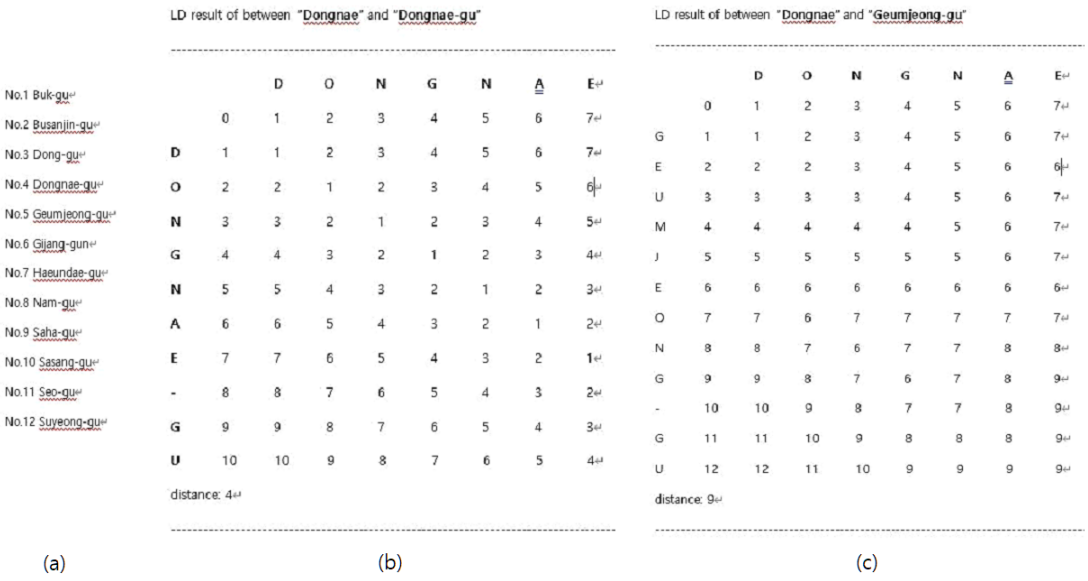

Fig. 4. LD result between the query city that satisfies the query province and the KCDC confirmed patient's city

(a) the number of cities corresponding to the province of the query, (b) "Dongnae"–"Dongnae-gu", (c) "Dongnae"–"Geumjeong-gu".

Region with the highest similarity: **Busan**, District: **Dongnae-gu**

Number of elementary schools: 22 Number of elementary schools: 31

Number of universities: 0 Number of institute rate: 1.98

kNN analysis result (k=10)

No.1	Sex: male	Contact number: 3
	Confirmed date: 2020-02-24	Infection case: Onchun Church
No.2	Sex: male	Contact number: 27
	Confirmed date: 2020-02-22	Infection case: Onchun Church
No.3	Sex: male	Contact number: 25
	Confirmed date: 2020-02-23	Infection case: Onchun Church
No.4	Sex: male	Contact number: 5
	Confirmed date: 2020-03-18	Infection case: overseas inflow
No.5	Sex: male	Contact number: 3
	Confirmed date: 2020-03-25	Infection case: overseas inflow
No.6	Sex: male	Contact number: 28
	Confirmed date: 2020-02-23	Infection case: etc
No.7	Sex: male	Contact number: 28
	Confirmed date: 2020-02-23	Infection case: etc
No.8	Sex: male	Contact number: 28
	Confirmed date: 2020-02-23	Infection case: etc
No.9	Sex: male	Contact number: 2
	Confirmed date: 2020-03-03	Infection case: contact with patient
No.10	Sex: male	Contact number: 2
	Confirmed date: 2020-03-03	Infection case: contact with patient

Fig. 5. Infection information of k confirmed patients who satisfy both the province and city conditions of the query.

compares whether the province query value is the same, and then calculates the city LD in the same province. As can be seen from the results in Fig. 4, after finding the city like Fig. 4 (a) belonging to the query (Busan), the city with the closest distance value is selected through LD operation as shown in Fig. 4 (b). Fig. 4 (c) shows that the two most distant cities in Fig. 4 (a).

Fig. 5 shows the results of searching for the number of contacts, confirmed date, cause of infection of the confirmed patient with the closest gender and age provinces and cities selected with kNN through Fig. 3, 4. Fig. 5 shows the number of elementary schools, universities, and institutes in the city. Through this information, it makes not only determine the risk of spread of infections from confirmed patients in the city, but also figure out the risk of educational facilities according to the occurrence of confirmed patients. In addition, it is possible to identify and avoid dangerous places visited by confirmed patients with similar conditions by searching for traces of them corresponding to age or gender. Therefore, the proposed method shows that it can provide useful information such as prevention of COVID-19 spread and avoidance.

IV. Conclusions

COVID-19 is the virus that can endanger the whole world like SARS (SARS-CoV-1, 2003), MERS (MERS-CoV, 2015). Especially, COVID-19 is likely to spread rapidly because it can be infected easily by droplets or physical contact. Unfortunately, the COVID-19 vaccine is currently in development, the best solution is to find and avoid the place where you are likely to contact with confirmed patients. In this paper, we proposed a method to search for the infection routes, the date of confirmation, etc. of the confirmed patients, which is similar to the information in the desired area, to prevent contact beforehand using LD and kNN. In order to verify the performance of the proposed method, it is applied to the confirmed patient data collected by KCDC, and we found that the number of contacts, confirmed date and infection route of them belonging to the query (city and age) were searched. Various solutions can be sought through the information of the similar confirmed patients.

V. Future Work

In this paper, Levenshtein Distance and k-Nearest Neighbor algorithm are used to build a Hot spot Detection. Levenshtein is a classical algorithm to measure the difference between two string. But it takes a large time cost and storage cost. In future work, other distance function will be used to computing speed and reduce storage space such as Monger-Elkan distance. Also, In this paper, kNN algorithm just calculates the distance between one input data and all other patients. In the next step, k-means algorithm can be used before kNN, the speed of distance calculation will increase significantly. In addition, the data this paper used from Korea Centers for Disease Control and Prevention (KCDC) includes some latitude and longitude related to confirmed patients, latitude and longitude can be used to improve the accuracy of hot spots. The ultimate solution for COVID-19 is the completion of vaccine, but finding a method to avoid contact with potential patients is an efficient way to prevent the spread of COVID-19 and protect people's lives. The future work also including searching the information of similar confirmed patients such as infection route and traces not only within Korea but also other countries simultaneously.

References

[1] Petrosillo, Nicola, et al. "COVID-19, SARS and MERS: are they closely related?." Clinical Microbiology and Infection (2020).

[2] World Health Organization. "Coronavirus disease 2019 (COVID-19): situation report, 72." (2020).

[3] Kang, Yun-Jung. "Lessons learned from cases of COVID-19 infection in South Korea." Disaster Medicine and Public Health Preparedness (2020): 1-20.

[4] Park JY. Corona Corresponds to Korea, Being Caught by Sinchon and Conservatives. New 1; February 29, 2020. [cited March 6, 2020]. https://www.news1.kr/articles/?3858187. (Korean).

[5] Korea Centers for Disease Control and Prevention. February 4, 2020. [cited March 6, 2020]. http://ncov.mohw.go.kr/baroView.do?brdId=4&brdGubun=&dataGubun=&ncvContSeq=&contSeq=&board_id=. (Korean).

[6] Samuel, Jim, et al. "Covid-19 public sentiment insights and machine learning for tweets classification." Information 11.6 (2020): 314.

[7] Chour, William, et al. "Shared Antigen-specific CD8+ T cell Responses Against the SARS-COV-2 Spike Protein in HLA A* 02: 01 COVID-19 Participants." medRxiv (2020).

[8] Zhang, Shengnan, Yan Hu, and Guangrong Bian. "Research on string similarity algorithm based on Levenshtein Distance." 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017.

[9] Konstantinidis, Stavros. "Computing the Levenshtein distance of a regular language." IEEE Information Theory Workshop, 2005.. IEEE, 2005.

[10] Dudani, Sahibsingh A. "The distance-weighted k-nearest-neighbor rule." IEEE Transactions on Systems, Man, and Cybernetics 4 (1976): 325-327.

[11] Seidl, Thomas, and Hans-Peter Kriegel. "Optimal multi-step k-nearest neighbor search." Proceedings of the 1998 ACM SIGMOD international conference on Management of data. 1998. Tan, Songbo. "Neighbor-weighted k-nearest neighbor for unbalanced text corpus." Expert Systems with Applications 28.4 (2005): 667-671.

[12] Zhang, Min-Ling, and Zhi-Hua Zhou. "A k-nearest neighbor based algorithm for multi-label classification." 2005 IEEE international conference on granular computing. Vol. 2. IEEE, 2005.

[13] Weinberger, Kilian Q., and Lawrence K. Saul. "Distance metric learning for large margin nearest neighbor classification." Journal of Machine Learning Research 10.2 (2009).

[14] Hajebi, Kiana, et al. "Fast approximate nearest-neighbor search with k-nearest neighbor graph." Twenty-Second International Joint Conference on Artificial Intelligence. 2011.

[15] Anggraini, Nenny, and Muhammad Jabal Tursina. "Sentiment Analysis of School Zoning System On Youtube Social Media Using The K-Nearest Neighbor With Levenshtein Distance Algorithm." 2019 7th International Conference on Cyber and IT Service Management (CITSM). Vol. 7. IEEE, 2019.

[16] Yujian, Li, and Liu Bo. "A normalized Levenshtein distance metric." IEEE transactions on pattern analysis and machine intelligence 29.6 (2007): 1091-1095.

[17] Wagner, Robert A., and Michael J. Fischer. "The string-to-string correction problem." Journal of the ACM (JACM) 21.1 (1974): 168-173.

[18] Withum, Timothy O., Kurt P. Kopchik, and Oren I. Oxman. "Modified Levenshtein distance algorithm for coding." U.S. Patent No. 7,664,343. 16 Feb. 2010.