International Journal of Latest Engineering Science (IJLES) E-ISSN: 2581-6659
DOI: 10.51386/25816659/ijles-v4i1p104
Volume: 04 Issue: 01 January to February 2021 www.ijlesjournal.org

An Approach to Assess Duplication Level of Texts using Fuzzy Weight

Trung Nguyen Tu[1], Ngan Tran Thi Thu[2]
[1]*Thuy Loi University, Hanoi, Vietnam*
[2]*Foreign Trade University, Hanoi, Vietnam*

Abstract *– Content Duplication of Text is a common issue on newspapers, news websites and creations. Therefore, detecting duplication which examines the similarities among documents is very essential. However, it is not easy to deal with this problem, which has been also interested by many researches recently. Nowadays, there are a lot of methods which are researched to cope with it. In this paper, we recommend an improvement on similarity measure based on fuzzy logic and practical application in detecting content duplication of the articles.*

Keywords:*Text, Duplication Detecting, Similarity Measure*

I. Introduction

Content duplication of text is a popular issue in our life. For many reasons, a great number of texts are plagiarised or quoted, which makes the same content or written words among the websites. These things raise us a question of finding and detecting the duplication of content. For example, an author would like to check whether his works are illegal used or not. Likewise, checking plagiarism in the areas of music, literature and etc is interested by a lot of experts. For document storage systems, it wastes memory when we keep the documents with high duplication.For search engines used to collect data from the Internet, if the duplication of new data is well evaluated and comparedto documents in other sources, we can avoid downloading and storing documents with duplicate content. Therefore, the issue of detecting duplication is currently of concern.

Detecting the duplication is not easy for many reasons. Texts are not copied totally but partly and there may be some changes in the copied parts, for instance, the order of words or part of text. To check the duplicate content, we can use some techniques such as: Shingling[1], I-Match[2], Random Projection, SpotSigs, the similarities between two documents and so on.

It can be seen that detecting the duplication actually means calculating the similarity in the content of the required texts to be compared with the texts available in the corpus. The similarity is based on the index as below:

- The similarity of meaning of texts: keywords, TF – IDF;
- The similarity of sentences, paragraphs;
  - The similarity of grammar of texts: part of speech, syntax sentence structures, …
  - The similarity of HTML tabs of the websites.

Using a criterion to evaluate the similarity of texts in a corpus is becoming increasingly ineffective sinceboth Internet users and copying tools are now getting smarter. Therefore, recent researches focus on collaborating criteria of examining similarity to increase the accuracy of similarity checking tools, search engines and so on.

[4] refers that Muneer and his partners proposed an algorithm for establishing clusters of duplicate sites. Besides, Fresno and his partners recommended FCC weight Function as fuzzy system for the act of foisting particular weights and their combination [5][3].

In Vietnam, there were some researches in detecting the duplicate content in Vietnamese text corpus [9],[6],[8]. The researches show that using combined criteriaincreases the accuracy of algorithms. However, those researches have also indicated that the improvement needs researching more to optimize the combination of the criteria and enhance the accuracy of detecting duplication.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

This paper proposes the improvement of evaluating the similarity between two texts. The remainders of this paper are presented as follows: Section 2 will present theduplicate content checker tooland the similarity measure between two texts. Section 3 willshow the new similarity measure which is improved by using fuzzy logic. Some practical experiments are presented in Section 4. Then Section 5 will conclude the paper.

II. The sytem of Duplicate content checking

[6] refers a model of the system duplicate content checking. It examines an online newspaperwhetherits content is the same or almost the same as the newspaperspreviously collected. Data taken from electronic articles are written in Vietnamese.

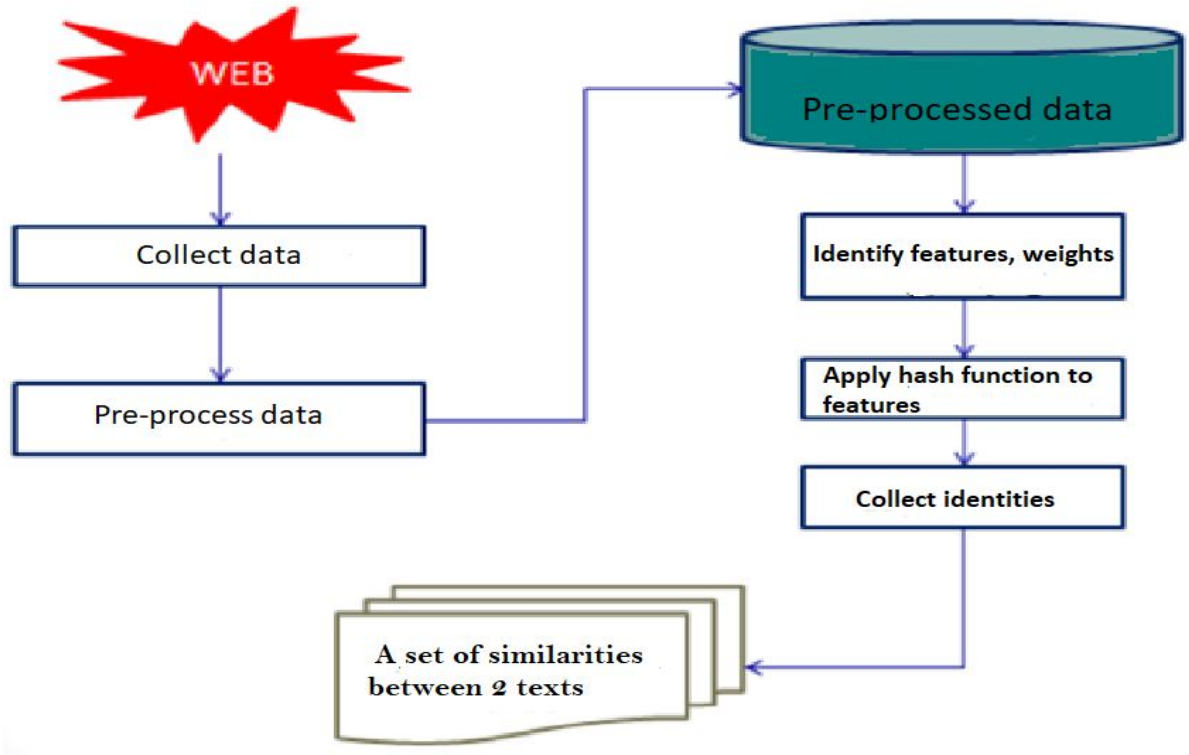

Fig 1. Experimental model to detect the content duplication of the texts [6].

The steps are as follows:

Step1: Collect the online newspapers

Step2: Pre-process the content of the newspapers

Step3: Process to do Shingling for each required document

Step4: Collect the particular identities of each required document

Step5: Compare and show the result

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

International Journal of Latest Engineering Science (IJLES) E-ISSN: 2581-6659
DOI: 10.51386/25816659/ijles-v4i1p104
Volume: 04 Issue: 01 January to February 2021 www.ijlesjournal.org

III. Similarity measure

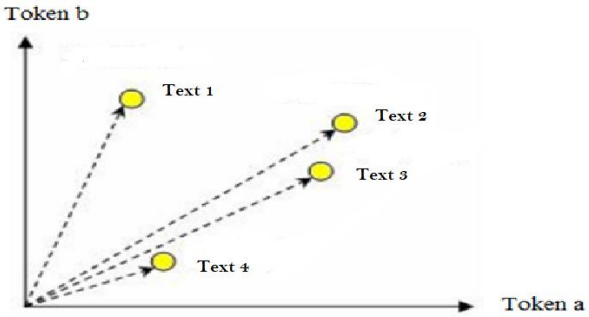

Fig 2. A vector model oftexts [7].

Text is shown in the shape of Vector basing on Frequency modulation [7], for example, the methods base on Term Frequency (IF) and inverse document frequency(IDF). Figure 2 shows the text with 2 tokens. In general, the number of tokensisvarious, for example, the total of the syllables (using syllable particulars), the total of words (using word particulars).

We use some methods such as cosine similarity, Jaccard coefficient, Euclidean distance, Pearson Correlation coefficient [7]. In this paper, we consider Euclide distance as the formula below:

$$d_{Euclide}(A,B) = \sqrt{\sum_i (A_i - B_i)^2} \quad (1)$$

IV. The clustering algorithm

In case of big database, searching in the corpus makes execution speed very slow. Therefore, we have to proceed hierarchical clustering data of texts in advance to boost the searching speed up. Besides, Hierarchical clustering technique still helps to divide the text database into various levels.

KMeans Clustering [10] includes 4 steps as below:

Input:Observationx_i(there are n observations), subject to i={1, 2,..., n}and cluster of c

Output:Cluster C_j (j={1,...c}), target formula E gains its minimum value:

$$E = \sum_{j=1}^{c} \sum_{x \in C_j} d^2(x, C_j) \quad (1)$$

Step 1: Create

Choose a set of observations C_j (j={1,2,…c}) (there are k observations) is the first centroidof K Input Clusters (random or experienced choice).

Step 2: Point the centroidcluster by distance

For each observation xi (i={1,…n}), calculating the distance from it to each centroidof Cj (j={1,…c}). The observation belonging to Cluster Cs makes the distance from centroidCs to it gain its minimum value

$$d(x, C_S) = \min d(x, C_j), \; j = 1..c \quad (2)$$

Step 3: Update the centroidof Cluster

For j={1,…,c}, get the update of centroidof Cj by deciding the average of observative vectors, which were pointed to the centroid.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

International Journal of Latest Engineering Science (IJLES) E-ISSN: 2581-6659
DOI: 10.51386/25816659/ijles-v4i1p104
Volume: 04 Issue: 01 January to February 2021 www.ijlesjournal.org

$$C_{jk} = \frac{\sum_{x \in cluster(j)} x_k}{count(cluster(j))} \quad (3)$$

Step 4: Duplicate and check the stop condition

Step 2 and 3 are repeated until convergences remain unchangeably between 2 continuous repetitions.

- $d(x, C_j)$: distance from x to centroidCj
- C_{jk}: k – the number kin a series of component ofcentroidCj
- x_k: k – the number k in a series of component of observation x

V. Aimproving of similarity measure

At present, similar measures can evaluate the features with the same role and only rely on frequency to distinguish characteristic values of each document. It doesn't make sense if we only use syllable-level feature which has no meaning. However, if we use the word-level feature, this is unreasonable. The reason is that common words will be used more often than proper names, numbers ... so the repeatability is much higher. In contrast, abbreviations have very low repeatability in texts. In other words, in terms of repeatability, common words have the greatest ability to repeat while abbreviations have the least. Thus, if we classify features according to different levels, we can evaluate text similarity more accurately.

From the above, the authors propose a set of rules to determine the influence level on different features as below:

1) Very great – if the feature is a common word
2) Great – if the feature is a proper name
3) Medium – if the feature is a person's name or a name of an entity
4) Small – if the feature is a percentage or a number
5) Very small if if the feature is an abbreviation.

Supposing$F_i$as i^{th}feature, the effect functionof feature$F_i$is$effect(F_i)$. Then, the fomula to measure the similarities (1)becomes:

$$d_{Euclide}(A, B) = \sqrt{\sum_i (effect(A_i)A_i - effect(B_i)B_i)^2} \quad (2)$$

VI. Experiments

Text similarity measure is used in duplicate text lookups. In particular, the authors use a corpus containing over 500 articles (title, abstract). For new articles, the system compares the similarity of the new abstracts with the others in the database. Then the corpus is clustered. In lookup phase, the system will list 5 articles with the highest similarity from the clusters. The system offers 2 ways to find out the duplicate texts. The first case is applied when the number of texts in the corpus is not too much, all the texts can be browsed and the similarity can be compared with the input text. The second case is applied when there are too many texts in the corpus. The lookup phase consists of 2 steps. In step 1, the system compares the similarity of the input text with the data clusters in the corpus. In step 2, the system continues comparing the similarity of the input text with the most the most similar texts in each cluster.

A. Evaluate the corpus clustering

To evaluate the corpus clustering, refering to [11], the authors use*F(I)* [13], following the cluster uniformity criteria [11] [12] to compare clustering results of algorithms. The smaller the value of F (I) is, the higher the uniformity is. This index is calculated as follows:

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

International Journal of Latest Engineering Science (IJLES) E-ISSN: 2581-6659
DOI: 10.51386/25816659/ijles-v4i1p104
Volume: 04 Issue: 01 January to February 2021 www.ijlesjournal.org

$$F(I) = \frac{1}{1000(N\times M)}\sqrt{R}\sum_{i=1}^{R}\frac{e_i}{\sqrt{A_i}} \quad (8)$$

Table 1 and Figure 6 show the statistics comparing the quality of text clustering in the cases of using and without using fuzzy weight in the cases of 3, 5, 6 and 8 clusters.

Table 1:Comparing the similarity of clusters

	3	5	6	8
Not Fuzzy	0.00303	0.00485	0.00521	0.0072
Fuzzy	0.0016	0.00266	0.00316	0.00415

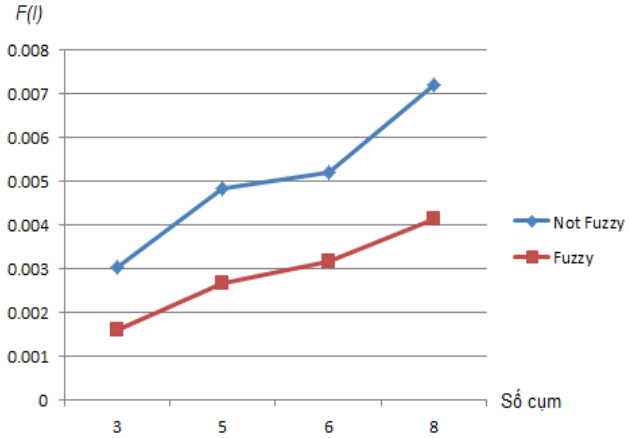

Fig 3. Comparing the unifomity of clusters

B. Detect the similar texts

I) Experiment 1

Table 2: Input text in experiment 1

Title	Summary
Study of social behavior, memory and learning in experimental animals injected with drugs that cause schizophrenia	Evaluate the activities of moving, social interaction, spatial memory of Swiss rats before and after chronic ketamine injection with dose range of 10-35 mg / kg / day; Build an experimental schizophrenia model with appropriate ketamine doses, and then treat with antipsychotic drugs. Assess changes in animal behavior, memory and learning before and after the treatment.

Table 3: Searching results of experiment

Title	Similarity
Study of social behavior, memory and learning in experimental animals injected with drugs that cause schizophrenia	100%
Researching and applying geographic information system (GIS) and SWAT model to forecast the flow rate and soil erosion in the sub-basin of On Luong - Hop Thanh	71%

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

river.	
Research on microorganism applied to biogas production to increase efficiency in brackish water and salt water conditions.	70%

II) Experiment 2

Table 4: Input text in experiment 2

Title	Summary
Impacts of climate change on droughts in the south-central region of Vietnam, predictability and response solutions	Overview of drought situation and studies on drought, drought prediction based on greenhouse gas emission scenarios; Experiment and select an appropriate meteorological drought index, thereby determining the degree of drought changes in the South-Central region in the past, trend of future changes according to scenarios of greenhouse gas emissions then proposing response solutions.

Table 5: Searching results of experiment 2

Title	Similarity
Impacts of climate change on droughts in the south-central region of Vietnam, predictability and response solutions.	100%
Analyzing the geochemical and petrological characteristics of the coal rock and the coal clay of the miocene sediments in the north of the Red River sedimentary basin	71%
Researches and application of geographic information system (GIS) and SWAT model to forecast runoff flow and soil erosion in On Luong - Hop Thanh sub-basin.	70%
Research on microorganism applied to biogas production to increase efficiency in brackish water and salt water conditions.	70%
Establishing a scientific basis for environmental protection planning in Phu Loc district, Thua Thien Hue province.	70%

III) Experiment 3

Table 6: Input text in experiment 3

Title	Summary
Improving the financial capacity of joint stock commercial banks in Vietnam.	Systematize andcomplete the basic theories about financial capacity of commercial banks such as giving views on finance, financial capacity of commercial banks. Especially, the thesis has focused on analyzing the basis to give explanations leading to theoretical presentation of financial capacity of commercial banks; Analyze more clearly the basis and show the meaning of the evaluation criteria of the financial capacity of commercial banks, at the same time, the approach to analyzing the influencing factors also shows the logic and the system with the solutions; On the basis of researching experiences in enhancing the financial capacity of banks in some countries around the world, the main causes leading to weaknesses in banking and financial capacity arequick credit growth and

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Title	Summary
	unsustainable development. On the other hand, in order to improve the financial capacity of commercial banks, apart from the efforts of commercial banks themselves, they are still in need of the support from the Central Bank and the Government. These are also essential lessons in improving the financial capacity of Vietnam's commercial banks.

Table 7:Searching results of experiment 3

Title	Similarity
Improving the financial capacity of joint stock commercial banks in Vietnam.	100%
Completing the preparation and presentation of consolidated financial statements in the steel manufacturing enterprises of the Vietnam Steel Association	46%
Corporating financial risk management in Vietnam	46%
Analyzing the geochemical and petro logical characteristics of the coal rock and the coal clay of the miocene sediments in the north of the Red River sedimentary basin	45%
An enhanced K-Means clustering algorithm for unsupervised multi-spectral segmentation	45%

VII. Conclusion

In this paper, we have proposed to improve similarity measure between two texts based on fuzzy logic. The fuzzy logic is constructed to generate the weights affectedby features. The results show that the improved measure is well applied to the comparison of Vietnamese texts. In addition, we apply improved similarity measure in searching duplicate texts.

In the following study, we plan to analyze more deeply and do research on the role and position of features in sentences to show the influence level according to specific contexts.

References

[1] A.Z. Broder, S.C. Glassman, M.S. Manasse, G. Zweig, Syntactic Clustering of the Web, Computer Network, 1997.

[2] E. Uyar, Near-duplicate news detection using name entities, 2009.

[3] M. A Hearst,Clustering versus faceted categories for information exploration, In Communications of the ACM, 2006.

[4] Muneer K., Syed Farook K, An Innovative Approach for Clustering of Web Pages Based on Transduction, International Journal of Advanced Research in Computer Science & Technology IJARCST, Vol. 2, Issue 3, 2014, pp. 241-244.

[5] Xuemin Lin Chuan Xiao, Wei Wang. Efficient similarity joins for near duplicate detection, In 17th international conference on World Wide Web, 2008.

[6] Phạm Kim Hồng, Luận văn thạc sĩ, Phát hiện sự trùng lặp nội dung của các bài báo, 2013.

[7] Lê Mạnh Hùng, Luận văn thạc sĩ, Tra Cứu Văn Bản Tiếng Việt Dựa Trên Kỹ Thuật Phân Cụm, 2012.

[8] Dương Thăng Long, Báo cáo đề tài nghiên cứu: Nghiên cứu độ đo tương tự trong văn bản tiếng Việt và ứng dụng đánh giá việc sao chép bài điện tử.

[9] Nguyễn Tuấn Anh, Luận văn thạc sĩ, Phát hiện trùng lặp văn bản và và xây dựng chỉ mục hiệu quả cho WebCrawler, 2009.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

[10] http://www.onmyphd.com/?p=KMeans.clustering.

[11] Valliammal N., S.N.Geethalakshmi, Leaf Image Segmentation Based On the Combination of Wavelet Transform and K Means Clustering, International Journal of Advanced Research in Artificial Intelligence, Vol. 1, No. 3, 2012.

[12] R. H. Haralick, and L. G. Shapiro, Image segmentations techniques, Computer Vision Graphics Image Processing 29, pp. 100-132, 1985.

[13] J. Liu, and Y. H. Yang, Multiresolution color image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.16, no.7, pp.689-700, Jul 1994.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)