Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 5, 1898 – 1909

Enhance Data Similarity Using a Fuzzy Approach

[1]Dhuha Kh. Altmemi, [2]Imad S. Alshawi

[1]Department of Computer Science, College of Computer Science and Information Technology, University of Basrah, Basrah, IRAQ, dhuhakhalaf95@gmail.com
[2]Department of Computer Science, College of Computer Science and Information Technology, University of Basrah, Basrah, IRAQ, emad.alshawi@uobasrah.edu.i

Abstract

Text similarity is critical in a variety of applications, including word processing, signal processing, imagery, data mining, wireless sensor networks, etc., where text similarity measurements can detect whether texts are lexical or semantic similar. Semantic text similarity is the term that uses to describe similarities based on meaning. Although this function is very challenging, it remains an active subject of study due to the complexities of natural language. The second type is lexical similarity whereby this type can be used to eliminate repetition by grouping similar texts together provided that two texts are very similar. It is important to remember that traditional text similarity approaches only look at the actual words in a phrase to compare two texts. Depending on the use case, it's easier to build and manage and offers a better trade-off. This paper examines current work on text similarity and divides it into four categories. Techniques based on strings, Corpus, knowledge, or hybrid similarities, these categories are all comparable. There are also examples of different combinations of these techniques for matching text and finding similarities between two texts. A smart method is proposed to find out the similarity between two texts called the fuzzy data similarity (FDS), and to prove the efficiency of the proposed method, it was compared with the most famous methods, where the results showed an accuracy of the FDS about 93%.

Keywords: Text Similarity, String-Based, Corpus-Based, Knowledge-Based, Hybrid Similarity, Fuzzy Data Similarity.

1. INTRODUCTION

The ability to easily locate the most relevant material from the vast sea of information available on the Internet is becoming more and more important. To find the most relevant information from a large group, we may use a variety of strategies aimed at studying the similarity of texts; Such as image processing [1], signal processing [2], artificial intelligence [3], wireless sensor networks [4], data mining [5], machine learning [6] and so on. Text summarization relies heavily on the similarity of sentences and paragraphs to effectively perform tasks such as retrieving information, compiling documents, and clarifying word meaning. The result is returned based on how well the user query text matches the content in the document results. In addition, text similarity plays a significant influence in the classification of both text and document classifications. The similarity of phrases, words, paragraphs, and documents can be measured to classify them effectively. Using this classification, we can find the most relevant content for the user's search. In this paper, several similarities are discussed, such as lexical and semantic[7], the first is a lexical similarity that can be used to remove redundancy by grouping similar texts together. If two texts are very similar, you can always get rid of redundant information. Consider

duplicate product listings or the same person in your database with little name difference or even HTML pages that are close to duplicating. You can use well-established methods such as BM25, PL2, etc., but you can also develop your idea using a scale such as a cosine or Jaccard and dice also for short texts. The second type focuses on how similar the two terms are in their meanings. This was the primary focus of the study of NLP in the past. Even if the phrases “I like fruit” and “I ate a lot of fruit” sound the same, they refer to something completely different. It usually takes a greater degree of examination to decipher the meaning of the phrase. This is just one humble example. The syntax is used by the majority of individuals to identify semantic similarities[8]. Figure 1 shows the text-similarity metrics that are divided into four categories. Finding word similarity is an essential aspect of text similarity, which is then used to compare phrases, paragraphs, and texts.

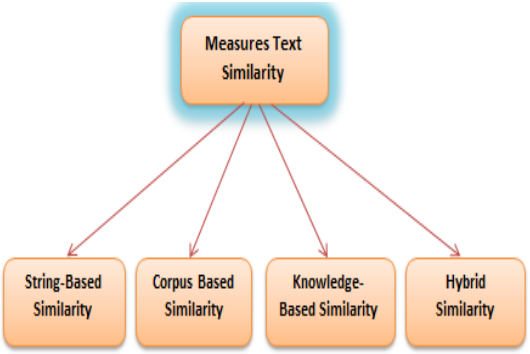

Figure 1. *Four major Measures of text similarity*

This paper uses a variety of methods, first string-based to introduce lexical similarity, then corpus-based and knowledge-based to introduce semantic similarity, and finally hybrid [9]. To introduce semantically similarities, corpus-based and knowledge-based algorithms are applied. Character composition and string sequences are used to calculate string-based measures (a string metric compares or matches two text strings based on their similarity or dissimilarity distance). Corpus-Based Similarity is a semantic similarity metric that examines words using data from a huge corpus. Knowledge-Based Similarity (KBS) is a semantic similarity measure that uses information from semantic networks to determine the degree of similarity between words[9]. Each category s most popular items will be briefly explained.

In this paper A smart new method is proposed to find out the similarity between two texts called the fuzzy data similarity(FDA), and to prove the efficiency of the proposed method, it was compared with the most famous methods, where the results showed an accuracy of the FDA about 93%.

The format of this paper is as follows: Section 2 explains the string-based similarity algorithms in detail. The Corpus-Based and Knowledge-Based techniques are discussed in Sections 3 and 4, respectively. Section 5 presents the hybrid similarity procedures. Section 6 presents the new proposal called Fuzzy Data Similarity. Finally, Section 7 summarizes the results of the similarity methods.

2. STRING-BASED SIMILARITY

Character composition and string sequencing are both governed by string similarity metrics. A string metric is either a match or a mismatch, The comparison match measures the similarity or difference between two strings' distance. This paper includes the most common measurements of series similarity seven of the algorithms will be character-based, while the others will be term-based distance scales, as indicated in Figure 2.

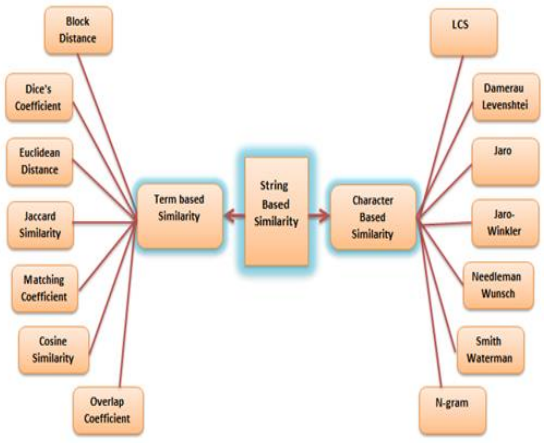

Figure 2. *String-Based Similarity Measures*

2.1. Character-Based Similarity

1) Longest Common Substring (LCS) is the matching of the first string to the second string based on the appearance of the characters in both strings [10].

2) Damerau Levenshtein is a distance algorithm that calculates the number of operations necessary to change one string to another to find the distance between two strings. An operation is described as the addition of a single character, the deletion of a character, the replacement of two adjacent characters, or the substitution of two adjacent characters[11].

3) Jaro is used to connect the two strings by counting the number and sequence of similarities of characters. It is used extensively in the field of linking records and calculating the usual spelling differences [12].

4) Jaro Winkler is a Jaro space extension where it uses a prefix scale that is equivalent to strings that match from the start the length of that prefix[13].

5) Needleman Wunsch is a type of dynamic programming that is used for the first accuracy to compare biological sequences. It finds the optimal alignment between two sequences by using a global alignment. Since chains can be considered equal in length if they have a high similarity ratio, they work well[14].

6) Smith-Waterman finds the best match between the conserved domains of two sequences using a local alignment. Smith-Waterman is useful in contrasting sequences that are similar to co-patterns or similarities that are within the larger sequence[15].

7) N-gram is a text sequence s n-item sub-sequence. The n-grams of each character or word in two strings are compared using n-gram similarity algorithms. Divide the number of linked n-grams by the maximum number of n-grams to get the distance[16].

2.2. Term based Similarity

1) Jaccard similarity is the intersection of two sets divided by the union of two sets. It uses a lexical method to determine similarity by comparing characters, words, strings, and sentences. Similarities between two sentences according to Jaccard[17].

2) Block distance is also known as Manhattan distance. Here, the distance between two data points is determined by following a grid-like route. The total of their component inconsistencies determines the Block distance between two items[18].

3) The cosine Similarity of two vectors in an inner product space is a measure of similarity that takes into consideration the cosine of their angle[19].

4) The method for obtaining the Dice Coefficient is to divide the total number of terms in both series by the number of similar words[20].

5) Euclidean Distance is the square root of the total squared differences between the similar members of two vectors, or L2 distance[21].

6) The matching parameter is a vector-based technique for determining the number of words in both strings where the matching parameter uses a non-zero vector[22].

7) The overlap coefficient is one of the types of term-based scales where the two coefficients are equivalent in the case of the first series being a subset of the second[23].

2.3. Results and Test String-Based similarity Algorithm

The results show the calculation of similarity and accuracy based on the test of the two texts (s1, s2) where s1 represents the original text and s2 represents the text that contains a missing part or an error resulting from the receipt of the text. Three algorithms of both types were chosen to calculate the percentage of similarity between the two texts as shown in the two tables. Table 1 shows the results of character-based algorithms, while Table 2 shows the results of term-based algorithms. The purpose of this test is to find out the percentage of similarity between two texts between [0 …

1], while the accuracy represents the accuracy taken to implement the approach.

Table 1. *Character-Based Similarity Measures*

Input Text	Algorithm /Approach	Similarity	Accuracy
S1 = Data mining is very important S2 = aata iiining is very important	Jaccard similarity	0.25	48%
	Cosine similarity	0.87	83%
	Overlap Coefficient	0.88	84%

Table 2. *Term-based Similarity Measures*

Input Text	Algorithm / Approach	Similarity	Accuracy
S1 = Data mining is very important S2 = aata iiining is very important	Damerau Levenshtein	0.88	84%
	Jaro	0.84	80%
	Jaro-Winkler	0.91	88%

3. CORPUS-BASED SIMILARITY

Corpus-Based Similarity is a semantic similarity metric that uses data from large corpora to determine word similarity. In language studies, the term "corpus" refers to a large collection of written or spoken items[24]. Figure 3 depicts the Corpus-Based Similarity Measures.

1) Hyperspace Analogue to Language (HAL) is a semantic space formed utilizing the frequent recurrence of words, each matrix element in a word matrix shows the degree of linkage between the words in the row and the words in the column. The technique s user may then choose to remove the low entropy columns from the matrix. When examining the text, the emphasis word is put at the beginning of a ten-word frame that captures neighboring terms that are classified as common. To construct the matrix values, co-occurrence is weighted inversely proportionate to the distance from the focal word; closer words are thought to represent more of the meaning of the word focus and are weighted higher. In addition, HAL maintains track of word order by interpreting frequent occurrences differently depending on whether the boundary word occurs first or last[25].

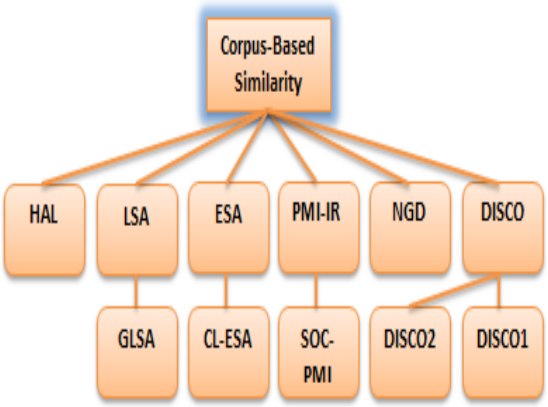

Figure 3. *Corpus-Based Similarity Measures*

2) Latent Semantic Analysis (LSA) is the most widely used body-based similarity approach and LSA believes that words with similar meanings will appear in connected textual works. A word count matrix for each paragraph, with rows denoting different words and columns denoting each individual. A paragraph consists of several sentences and a large block of text. The mathematical idea is referred to as the "singular value". System Decomposition (SVD) technology is used to reduce the number of components in the system. Keeping the similarity of the class structure afterward, the cosine of the angle is used to compare the words. Between any two rows and the two vectors, it creates[26].

3) Generalized Latent Semantic Analysis (GLSA) is known as the approach to creating meaning-driven phrases and document vectors. By concentrating on term vectors rather than the dual document-term representation, it extends the LSA approach. GLSA necessitates a dimensionality reduction strategy as well as a semantic relationship measure between concepts. In the GLSA approach, any similarity

measure on the space of words can be paired with any appropriate dimensionality reduction method. In the last phase, the weights for the linear combination of term vectors are calculated using the standard term-document matrix[27].

4) Explicit Semantic Analysis (ESA) is a method is a vector representation of text individual words or full texts) that employs a document corpus as a knowledge base in natural language processing and information retrieval. In ESA, a word is represented as a column vector in the text corpus s TF–IDF matrix, and a document a string of words) is the centroid of the vectors representing its words[28].

5) CLESA is an extension of ESA that supports several languages' cross-language explicit semantic analysis). To represent a document as a language-independent idea vector, CL-ESA employs a multilingual document-aligned reference collection. The degree of resemblance between two articles published in different languages may be determined by the cosine similarity between matching vector representations[29].

6) PMI-IR Point-wise Mutual Information - Information Retrieval) is a technique for finding word similarity that computes probability using AltaVista s Advanced Search query syntax. The PMI-IR similarity score is based on the proximity of two words on a web page[30].

7) SOC-PMI is an algorithm that has the advantage of being able to calculate the similarity between two words which doesn’t happen very often but does so in the same close terms. PMI is wise mutual information for classifying lists of adjacency terms keywords for the two target words for a huge set[31].

8) Normalized Google Distance (NGD) is a semantic similarity metric generated using Google s number of hits for a collection of keywords. In terms of Normalized Google Distance, keywords with similar or identical meanings in natural language are "close," and words with different meanings are "far[32].

9) DISCO is a method for extracting distributional similar words using co-occurrences. Words with comparable meanings occur in similar contexts, according to distributional similarity. To obtain distributional similarity, large text collections are statistically evaluated. DISCO is a method for calculating word distributional similarity by counting co-occurrences using a three-word context frame. When DISCO calculates the exact similarity between two words, it simply takes their word vectors from the indexed data and uses the Lin measure to compute the similarity[33]. DISCO returns the second-order word vector for a given word when the most distributional comparable word is requested. DISCO1 uses collocation sets to compute the first-order similarity between two input words, while DISCO2 uses collocation sets to compute the second-order similarity between two input words. DISCO2 calculates the second-order similarity between two input phrases by combining sets of distributional related keywords[34].

3.1. Results and Test Corpus-Based Similarity Algorithm

We tested a second class model Corpus-Based Similarity Point Mutual Information Approach-Information Retrieval, Table 3 shows the result of the approach test PMI-IR.

Table 3: *Corpus-Based Similarity Measures*

Input Text	Algorithm/ Approach	Similarity	Accuracy
Text= this is a foo bus red car foo bus blue car foo bar barred car shep bus blue	PMI-IR	Is & a = 4.34, this & is = 4.36,a & foo =2.85, car & shep =2.86, red & car =2.85, bar & bar =2.39, bar & red =2.31, car & foo = 2.25, shep & bus = 2.02, bus & blue = 2.03, blue & car = 1.88, foo & bar = 1.86, foo & bus = 1.46,bus & red = 1.03, bus & bus = 0.79	82%

4. KNOWLEDGE-BASED SIMILARITY

Figure 4 shows a semantic similarity measure that is based on evaluating the degree of similarity between terms using the semantic network s knowledge and based on semantic similarity between them. Semantic association, on the other hand, is a larger sense of resemblance that encompasses both the idea and its form[35]. Measurements of semantic similarity and measures of semantic relatedness are the two types of knowledge-based measures of similarity. Six measures of semantic similarity exist. Three are based on the substance of the information, while the other three are based on the length of the journey. WordNet and the Natural Language Toolkit are the two most common programs for knowledge-based similarity measures (NLTK). In the realm of knowledge measurement, the most often used semantic network is WordNet[36]. Words That Are Similar WordNet is a massive English lexical database. Cognitive synonyms (synsets) are collections of nouns, verbs, adjectives, and adverbs that communicate different concepts. Conceptual, semantic, and lexical links link synsets together.

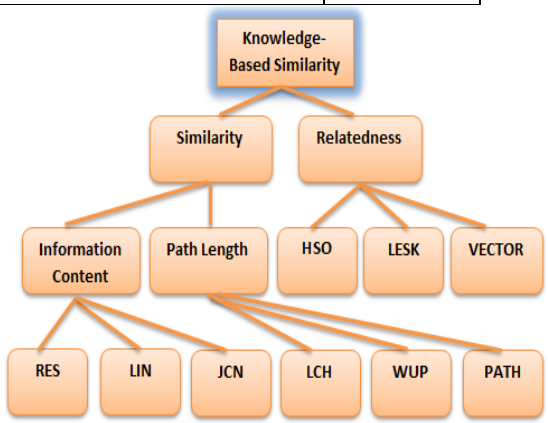

Figure 4. *Knowledge-Based Similarity Measures*

4.1. Measures of Semantic Similarity

Six scales of semantic similarity are divided by semantic similarity scales. Three of them are based on the contents of the information, while the other three are based on the length of the path.

4.1.1 Information-Based Metrics

1) RES is used for less common information content. The information content of nouns and verbs in WordNet is created using the repetition of concepts or tokens defined in a text or data set[37].

2) LIN scale multiplies the information content of the least common dependent by the total of the information content of concepts A and B, such that the Least Common dependent's information content is measured by this amount[38].

3) JCN is comparable to LIN Similarity in that both are inversely proportional to JCN Distance. The resulting score indicates how much information is required to express the similarity between the two ideas or synsets [39].

4.1.2 Path-Length-Based Metrics

Path similarity is the method of finding the shortest path between two groups or ideas of knowing path similarities. The result is discrete and unnatural, and there are no weights on the edges[40].

1) LCH gives a score indicating how close two-word senses are based on the shortest route connecting the senses and the maximum level of the taxonomy in which the senses exist[41].

2) WUP is a statistic that examines the taxonomy s placement of concepts or synsets c1 and c2 about the Least Common Subsume[42].

3) Path measurement provides a score indicating how close the two-word senses are based on the shortest path connecting the two senses in the classification[43].

4.2 measures of Semantic Relatedness

1) HSO is a two-word lexical string finder. The maximum achievable degree of correlation is 16[44].

2) LESK scale looks for luminosity overlaps between the two synchronization groups[45].

3) Vector is the representation of each word used in WordNet semantics from a given set by a co-occurrence vector, and each luminosity/concept is represented by a vector representing the average of these frequency vectors[46].

4.3. Results and Test Knowledge-Based Similarity Algorithm

From the third category, we tested the knowledge-based similarity in three ways to find out the degree of similarity between words. Table 4 shows the similarity with the implementation accuracy of the approach.

Table 4: *Knowledge-Based Similarity Measures*

Input Text	Algorithm /Approach	Similarity	Accuracy
S1= rat S2=lion	RES	4.66	26%
	JCN	0.08	12%
	LIN	0.52	43%

5. HYBRID SIMILARITY MEASURES

The Hybrid Similarity Scale combines one of the group techniques listed below with one of the individual similarity measures listed above. Multiple similarity measures are used in hybrid methods, and several studies have been conducted in this area. In[47] eight measures of semantic similarity were examined, two of the scales were based on a group, while the other six measures were based on knowledge. These eight algorithms were first evaluated independently and then combined. The best results were obtained by adopting an approach that combines different measures of similarity into one measure. In [48]describes a method for determining the semantic similarity of highly abbreviated phrases or texts using semantic information and word order. First, the lexical knowledge base and the set of texts are used to determine semantic similarity. Second, the proposed strategy takes into account the effect of word order in a sentence on meaning. The number of distinct words, as well as the number of word pairs in different sequences, is measured by the similarity of the derived word order. In [49]he described an approach that its creators called Semantic Text Similarity (STS). This approach uses a combination of semantic and grammatical information to determine the similarity between two texts. They considered two required functions string similarity and tag similarity as well as one optional function common word order similarity). In [50]a revised vector space vector

model is proposed to remove the ambiguity of names. The experimental results indicate that additional factors identified in the publications may be very useful for resolving ambiguity in names with a high degree of certainty. The mixed name problem was found to be resolved fairly easily. With an F1 score of 0.97, rating scales have seen a significant improvement. In [51]the idea was to combine two modules to create a promising correlation between manual and instrumental similarity results. The first unit uses N-gram-based similarity to calculate the similarity between sentences, while the second unit uses the WordNet-based Concept Similarity Scale to determine the similarity between concepts in two sentences. In[52] an approach for similarity discovery using information from Tamil Indo WordNet, Tamil Wiktionary, and Oxford Tamil Dictionary is proposed. To determine similarity, we used the definitions and model sentences for each word provided by each of these resources. Human evaluated Miller Charles and Rubenstein Good enough data sets are used to test the proposed technology.

6. THE PROPOSED METHOD

Fuzzy Data Similarity (FDS): the proposed method represents the process of extract of the similarity data based on match text and extracting useful information. a fuzzy Algorithm is a powerful approach that is used for solving more problems such [53], [54], [55]. In the context of data mining, a similarity measure is a distance with dimensions indicating object characteristics. That is, if the distance between two data points is minimal, the objects will have a high degree of resemblance, and vice versa. Most similarity approaches use distance measures to assess the differences between a pair of objects, and one of the most common distance scales with which they are compared with the algorithm proposed (Jaccard similarity, Cosine similarity, Overlap Coefficient) [56], [57],[58], [59], [60].

The fuzzy values are handled by the inference engine, which includes a rule base and a variety of techniques for inferring the rules, with a total number of 52=25 for the fuzzy rule base. As an example, IF FE(n) is high and SE(n) is low THEN the similarity (n) is Medium. All these rules are processed in a parallel way by a fuzzy inference engine. Then, the process of removing fuzzy values is performed to bring out a single clear and precise value from the fuzzy solution region. The CoG method of defogging is carried out by [61].

$$\mathbf{CoG} = \frac{\sum_{i=1}^{n} \boldsymbol{u}_i \boldsymbol{c}_i}{\sum_{i=1}^{n} \boldsymbol{u}_i}$$

Where CoG is the center of gravity and Ui is the output of rule base i, Ci is the center of the output membership function for n rule base number.

Figure 5 shows a "fuzzy data similarity " process.

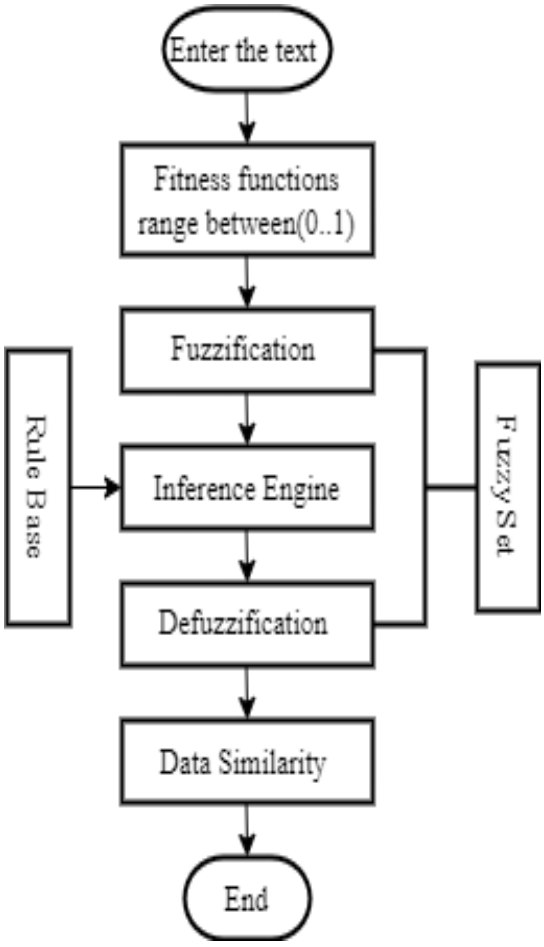

Figure. 5. *fuzzy data similarity process*

To test the work of the proposed algorithm, two texts are compared based on accuracy and similarity. Table 5 shows a comparison of the fuzzy data similarity algorithm with other methods.

Table 5. *Accuracy of the information*

Input Text	Algorithm /Approach	Similarity	Accuracy
S1 = Data mining is very important S2 = aata iiining is very important	Jaccard similarity	0.25	%48
	Cosine similarity	0.87	%83
	Overlap Coefficient	0.88	%84
	Fuzzy Data Aggregation	0.96	%93

As for choosing the similarity threshold, the similarity threshold is determined based on the type of application, It is clear through this that the higher the similarity percentage, the higher the accuracy of the information [58], [59]. Figure 6 shows the application of a similarity threshold ranging from (0..1) to the fuzzy data similarity algorithm.

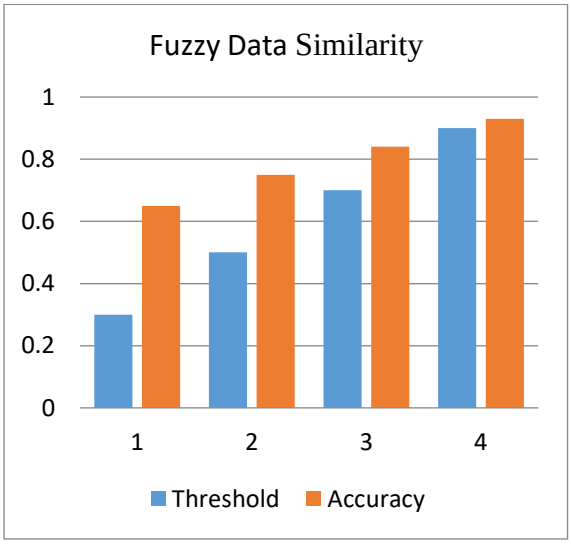

Figure 6. *Percentage of accuracy after applying the threshold*

7. CONCLUSION

Data can be extracted by knowing the percentage of similarity between two scripts to apply the necessary processors such as deleting duplicate data that occupies a large percentage of memory and others. In this paper, four textual approaches to problems are discussed; String Based similarity, Corpus, knowledge, and Hybrid. There are fourteen algorithms in the first category that indicates string-based similarity in two groups. first in term-based groups. Second, character-based. In the second category, nine algorithms Corpus-Based representing semantic similarity was presented, In the third category, nine knowledge-based similarity algorithms were presented, this group represents two types, the first type represents six based on semantic similarity, and the other two types, the first represents three algorithms based on the essence of information, while the other is along the path. The second group includes three measures of Semantic Relatedness. Finally, in the fourth category, the Hybrid Similarity Scale combines measures to address problems and obtain the best results, and several studies have been conducted in this area. In this light, a smart method is proposed to find out the similarity between two texts called the fuzzy data similarity(FDS), and to prove the efficiency of the proposed method, it was compared with the most famous methods, where the results showed an accuracy of the FDS about 93%. In future work, I will use an FDS algorithm to handle problems that occur in the data and troubleshoot and process errors based on the similarity of texts.

References

[1] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 201–216.

[2] X. Zhao and B. Ye, “The similarity of signal processing effect between SVD and wavelet transform and its mechanism analysis,” Acta Electron. Sin., vol. 8, 2008.

[3] J. Chandra, A. Santhanam, and A. Joseph, “Artificial intelligence based semantic text

similarity for rap lyrics," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1–5.
[4] R. Wan and N. Xiong, "An energy-efficient sleep scheduling mechanism with similarity measure for wireless sensor networks," Human-centric Comput. Inf. Sci., vol. 8, no. 1, pp. 1–22, 2018.
[5] C. C. Aggarwal, "Mining text data," in Data mining, 2015, pp. 429–455.
[6] V. Hatzivassiloglou, J. L. Klavans, and E. Eskin, "Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning," 1999.
[7] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," Int. J. Comput. Appl., vol. 68, no. 13, pp. 13–18, 2013.
[8] Z. Al-Huda, B. Peng, Y. Yang, and M. Ahmed, "Object scale selection of hierarchical image segmentation using reliable regions," in 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2019, pp. 1081–1088.
[9] S. Inzalkar and J. Sharma, "A survey on text mining-techniques and application," Int. J. Res. Sci. Eng., vol. 24, pp. 1–14, 2015.
[10] T. Kociumaka, J. Radoszewski, and T. Starikovskaya, "Longest common substring with approximately k mismatches," Algorithmica, vol. 81, no. 6, pp. 2633–2652, 2019.
[11] N. Hamidah, N. Yusliani, and D. Rodiah, "Spelling Checker using Algorithm Damerau Levenshtein Distance and Cosine Similarity," Sriwij. J. Informatics Appl., vol. 1, no. 1, 2020.
[12] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," J. Am. Stat. Assoc., vol. 84, no. 406, pp. 414–420, 1989.
[13] W. E. Winkler, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.," 1990.
[14] A. Naharuddin, A. D. Wibawa, and S. Sumpeno, "A high capacity and imperceptible text steganography using binary digit mapping on ASCII characters," in 2018 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2018, pp. 287–292.
[15] E. F. de O. Sandes and A. C. M. A. de Melo, "Retrieving smith-waterman alignments with optimizations for megabase biological sequences using GPU," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 5, pp. 1009–1021, 2012.
[16] G. Gledec, R. Šoić, and Š. Dembitz, "Dynamic N-Gram system based on an online croatian spellchecking service," IEEE Access, vol. 7, pp. 149988–149995, 2019.
[17] A. W. Qurashi, V. Holmes, and A. P. Johnson, "Document Processing: Methods for Semantic Text Similarity Analysis," in 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2020, pp. 1–6.
[18] X. Gao and G. Li, "A KNN model based on manhattan distance to identify the SNARE proteins," IEEE Access, vol. 8, pp. 112922–112931, 2020.
[19] J. Wang and Y. Dong, "Measurement of text similarity: a survey," Information, vol. 11, no. 9, p. 421, 2020.
[20] T. A. Basuki and B. G. Antaputra, "How Similar is Similar: A Comparison of Bahasa Indonesia and Bahasa Malaysia," in Proceedings of the 3rd International Conference on Electronics, Communications and Control Engineering, 2020, pp. 8–12.
[21] B. S. Lee, R. Phattharaphon, S. Yean, J. Liu, and M. Shakya, "Euclidean Distance based Loss Function for Eye-Gaze Estimation," in 2020 IEEE Sensors Applications Symposium (SAS), 2020, pp. 1–5.
[22] X. Chen, Y. Li, Z. Liu, J. Zhang, C. Chen, and M. Ma, "Investigation on matching relationship and plugging mechanism of self-adaptive micro-gel (SMG) as a profile control and oil displacement agent," Powder Technol., vol. 364, pp. 774–784, 2020.
[23] F. B. Allyson, M. L. Danilo, S. M. Jose, and B. C. Giovanni, "Sherlock N-Overlap: invasive normalization and overlap coefficient for the similarity analysis between source code," IEEE Trans. Comput., vol. 68, no. 5, pp. 740–751, 2018.
[24] E. S. Pramukantoro and M. A. Fauzi, "Comparative analysis of string similarity and corpus-based similarity for automatic

essay scoring system on e-learning gamification," in 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2016, pp. 149–155.
[25] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," Behav. Res. methods, instruments, Comput., vol. 28, no. 2, pp. 203–208, 1996.
[26] P. Kherwa and P. Bansal, "Latent Semantic Analysis: An approach to understand semantic of text," in 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), 2017, pp. 870–874.
[27] A. M. Olney, "Generalizing latent semantic analysis," in 2009 IEEE International Conference on Semantic Computing, 2009, pp. 40–46.
[28] T. Gottron, M. Anderka, and B. Stein, "Insights into explicit semantic analysis," in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 1961–1964.
[29] P. Sorg and P. Cimiano, "An experimental comparison of explicit semantic analysis implementations for cross-language retrieval," in International Conference on Application of Natural Language to Information Systems, 2009, pp. 36–48.
[30] J. Read, "Recognising affect in text using pointwise-mutual information," Unpubl. M Sc Diss. Univ. Sussex UK, 2004.
[31] A. Islam and D. Inkpen, "Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words.," in LREC, 2006, pp. 1033–1038.
[32] C. T. Lopes and D. Moura, "Normalized Google Distance in the Identification and Characterization of Health Queries," in 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), 2019, pp. 1–4.
[33] D. L. Horwitz, "Dance at the Ethiopian Disco: Tradition or Change," in The Beta Israel in Ethiopia and Israel, Routledge, 2013, pp. 205–214.
[34] P. Arpasat, P. Porouhan, and W. Premchaiswadi, "Improvement of call center customer service in a thai bank using disco fuzzy mining algorithm," in 2015 13th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015), 2015, pp. 90–96.
[35] R. Santhanam and J. Elam, "A survey of knowledge-based systems research in decision sciences (1980–1995)," J. Oper. Res. Soc., vol. 49, no. 5, pp. 445–457, 1998.
[36] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," Int. J. Lexicogr., vol. 3, no. 4, pp. 235–244, 1990.
[37] Z. Tan and L. He, "An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle," IEEE Access, vol. 5, pp. 27211–27228, 2017.
[38] B. Das, "Extracting Collocations from Bengali Text Corpus," Procedia Technol., vol. 4, pp. 325–329, 2012, doi: 10.1016/j.protcy.2012.05.049.
[39] S. Abujar, M. Hasan, and S. A. Hossain, Sentence similarity estimation for text summarization using deep learning, vol. 828. Springer Singapore, 2019.
[40] X. Bai and L. J. Latecki, "Path similarity skeleton graph matching," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 7, pp. 1282–1292, 2008, doi: 10.1109/TPAMI.2007.70769.
[41] S. Albitar, S. Fournier, and B. Espinasse, "An effective TF/IDF-based text-to-text semantic similarity measure for text classification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8786 LNCS, pp. 105–114, 2014, doi: 10.1007/978-3-319-11749-2_8.
[42] D. Choi, J. Kim, H. Kim, M. Hwang, and P. Kim, "A Method for Enhancing Image Retrieval based on Annotation using Modified WUP Similarity in WordNet," pp. 83–87.
[43] W. T. Yih, M. W. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf., vol. 1, pp. 1321–1331, 2015, doi: 10.3115/v1/p15-1128.
[44] P. Kolb, "Experiments on the difference between semantic similarity and

relatedness," Proc. 17th Nord. Conf. Comput. Linguist., pp. 81–88, 2009.
[45] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2276, pp. 136–145, 2002, doi: 10.1007/3-540-45715-1_11.
[46] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," J. Biomed. Inform., vol. 40, no. 3, pp. 288–299, 2007, doi: 10.1016/j.jbi.2006.06.004.
[47] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," Proc. Natl. Conf. Artif. Intell., vol. 1, pp. 775–780, 2006.
[48] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics," IEEE Trans. Knowl. Data Eng., vol. 18, no. 8, pp. 1138–1150, 2006.
[49] A. Islam and D. Inkpen, "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity," ACM Trans. Knowl. Discov. Data, vol. 2, no. 2, pp. 1–25, 2008, doi: 10.1145/1376815.1376819.
[50] T. Arif, R. Ali, and M. Asger, "Author name disambiguation using vector space model and hybrid similarity measures," 2014 7th Int. Conf. Contemp. Comput. IC3 2014, pp. 135–140, 2014, doi: 10.1109/IC3.2014.6897162.
[51] D. Buscaldi, R. Tournier, N. Aussenac-Gilles, and J. Mothe, "IRIT: Textual similarity combining conceptual similarity with an N-Gram comparison method," *SEM 2012 - 1st Jt. Conf. Lex. Comput. Semant., vol. 2, pp. 552–556, 2012.
[52] D. Karuppaiah and P. M. D. R. Vincent, "Hybrid approach for semantic similarity calculation between Tamil words," Int. J. Innov. Comput. Appl., vol. 12, no. 1, pp. 13–23, 2021.
[53] I. S. Alshawi, M. H. K. Jabbar, and R. Z. Khan, "Development of Multiple Neuro-Fuzzy System Using Back-propagation Algorithm," Int. J. Manag. Inf. Technol., vol. 6, pp. 794–804.
[54] I. S. Alshawi, A.-K. Y. Abdulla, and A. A. Alhijaj, "Fuzzy dstar-lite routing method for energy-efficient heterogeneous wireless sensor networks," Indones. J. Electr. Eng. Comput. Sci., vol. 19, no. 2, pp. 1000–1010, 2020.
[55] E. A. Al-Hussain and G. A. Al-Suhail, "A Fuzzy Based Clustering Approach to Prolong the Network Lifetime in Wireless Sensor Networks," in International Conference on Intelligent Computing & Optimization, 2021, pp. 97–107.
[56] G. Sahar, K. A. Bakar, F. T. Zuhra, S. Rahim, T. Bibi, and S. H. H. Madni, "Data Redundancy Reduction for Energy-Efficiency in Wireless Sensor Networks: A Comprehensive Review," IEEE Access, 2021.
[57] D. Ruby and J. Jeyachidra, "An Analysis of Repository and Similarity Measures in Meteorological Data of Underwater Wireless Sensor," Proc. 3rd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2019, pp. 964–968, 2019, doi: 10.1109/ICECA.2019.8822008.
[58] R. Maivizhi and P. Yogesh, "Spatial Correlation based Data Redundancy Elimination for Data Aggregation in Wireless Sensor Networks," 2020 Int. Conf. Innov. Trends Inf. Technol. ICITIIT 2020, pp. 0–4, 2020, doi: 10.1109/ICITIIT49094.2020.9071535.
[59] J. M. Bahi, A. Makhoul, and M. Medlej, "Data aggregation for periodic sensor networks using sets similarity functions," IWCMC 2011 - 7th Int. Wirel. Commun. Mob. Comput. Conf., pp. 559–564, 2011, doi: 10.1109/IWCMC.2011.5982594.
[60] S. Dhimal and K. Sharma, "Energy Conservation in Wireless Sensor Networks by Exploiting Inter-Node Data Similarity Metrics," Int. J. Energy, Inf. Commun., vol. 6, no. 2, pp. 23–32, 2015, doi: 10.14257/ijeic.2015.6.2.03.
[61] T. A. Runkler, "Selection of appropriate defuzzification methods using application specific properties," IEEE Trans. fuzzy Syst., vol. 5, no. 1, pp. 72–79, 1997.