Indian Journal of Science and Technology, *Vol 8(28), DOI: 10.17485/ijst/2015/v8i28/73899, October 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Novel Fuzzy Logic Model to Identify Closeness for Alias Detection

M. Subathra[1*] and R. Nedunchezhian[2]

[1]Department of Computer Applications, PSG College of Technology, Bharathiar University, Coimbatore – 641004, Tamil Nadu, India; subaclient@gmail.com
[2]KIT-Kalaignar Karunanidhi Institute of Technology, Coimbatore – 641402, Tamil Nadu, India; rajuchezhian@gmail.com

Abstract

The detection of person alias names is important for improving the accuracy of the information quality. The fuzzy-based decision system is proposed for alias detection, which is a rule-based system that uses fuzzy logic to make a decision about the closeness between the given name pairs. A fuzzy logic is formulated by a set of linguistic variables based on score value of the features. An entity pair's association score values are calculated using string and link-based features like Hamming Distance, Leventein Distance, Normalized String Edit Distance, Common Friends, Normalized Dot Product and Co-occurrence Relevance. These features are transformed into fuzzy input variables and designed with triangular membership function. The proposed novel fuzzy inference system gives the decision of aliases closeness in the form of crisp values ranging from 0.0 to 1. In this work, the model achieves upto 90% accuracy compared to estimated accuracy.

Keywords: Alias Detection, Closeness Identification, Fuzzy Logic, String and link based features

1. Introduction

The alias name of an entity is used to improve the information retrieval quality while it is added in semantic web. Paul et al.[1] investigated many classification algorithms such as KNN, decision tree, SVM and logistic regression incorporated with string and semantic measures for alias detection and found that the logistic regression has better performance for alias detection. An et al.[2] proposed the probabilistic based logistic regression classifier to find the semantic aliases of an entity in the web corpus, using features like co-occurrence, alias and social relevance for finding the association score between the two entities. Anwar[3] proposed an automatic pattern based approach with n-gram technique for alias extraction of a given entity in web corpus. Danushka et al.[4] proposed a method to generate patterns automatically from the web search using a training set having names and their aliases. However, the pattern-based approach is not suitable for people intentionally hiding their names.

Shen et al.[5] introduced fuzzy set based Absolute Order-of-Magnitude model with link-based properties such as cardinality and uniqueness for alias detection. Bhat et al.[6] enriched Latent Semantic Analysis to find the alias names. However, the system is not applicable for a large text of corpus. Meijuan et al.[7,8] extracted alias names using email dataset.

2. The Proposed Fuzzy Model

In alias detection, the problem is defined as a closeness function $f(x) = Y\varepsilon R$, where Y represents the closeness value of the given name pair that is expected to be an alias for an entity. The closeness function is defined the closed interval [0, 1] and the output value is shown as the closeness score of given name pairs.

**Author for correspondence*

2.1 Input Features

The following similarity functions are selected based on their performance and verified with the related works[1,2].

- Levenshtein Distance (LD) is computed as the minimum number of single character edits which is required to change one string into another string.
- Hamming Distance (HD) compares two names of equal length and the number of positions at which the corresponding letters are different and it measures the minimum number of substitutions required to change one string into another.
- Normalized String Edit Distance (NSED) is calculated with the minimum number of single character insertions, deletions and replacements required to convert one string to another string i.e., one name to another name. Also, it is computed by length of two strings.
- Co-occurrence Relevance (CR) is defined as the person's name co-occurring with the different aliases.
- Common Friends (CF) is defined as the number of friends that co-occur with the given name pair.
- Normalized Dot Product (NDP) is computed like a DP but each name vector is normalized by its magnitude from the named vector values divided by the total magnitude value prior to DP.

2.2 Output Feature

- A customized feature for finding alias name closeness is "Accuracy" and it has two membership functions to derive the output from the integrated input features function.

The Mamdani based fuzzy inference is used and the values of the all the features are transformed into fuzzy linguistic values and defined with triangular membership functions as shown in Figure 1.

The proposed fuzzy system for alias detection has the following steps which are shown in Figure 2.

- Input the feature set values of an entity pair.
- Fuzzification of the feature set for qualitative description.
- The building of fuzzy inference system and development of fuzzy rules.
- Defuzzification that involves the transformation of fuzzy sets into crisp value.

From the experimental dataset, a person's names are extracted from terrorist related dataset from Auton Lab,

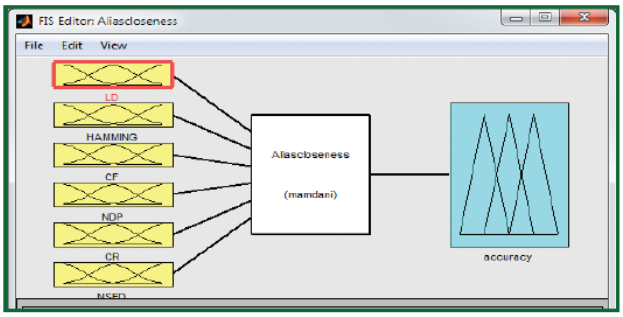

Figure 1. The developed Mamdani FIS.

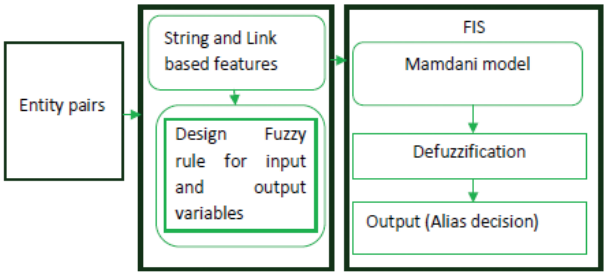

Figure 2. The proposed fuzzy-logic decision model.

carnegie mellon university. The name pairs are generated covering both positive and negative instances. The fuzzification process is started with the conversion of the crisp values of the extracted features into fuzzy sets and this process is defined by fuzzy logic membership functions. The triangular membership function is developed for the representation of the input and output parameters.

2.3 Fuzzy Membership Functions

The membership functions values are designed and normalized into the range [0 1]. The fuzzy input and output variables are simulated in MATLAB for alias detection as shown in Figure 3.

2.4 Fuzzy Rules Definition

The fuzzy depiction of linguistic variables is representing systems through fuzzy rules encoding the decision knowledge using a set of conditions that are satisfied and then a set of consequents can be inferred about the system. Sample fuzzy rules governing the alias detection are given in Figure 4.

2.5 Defuzzification

In a fuzzy system[9,10] it is easier to decide if the output is represented as a crisp value. The conversion of a fuzzy set to crisp value is done by defuzzification. The centroid

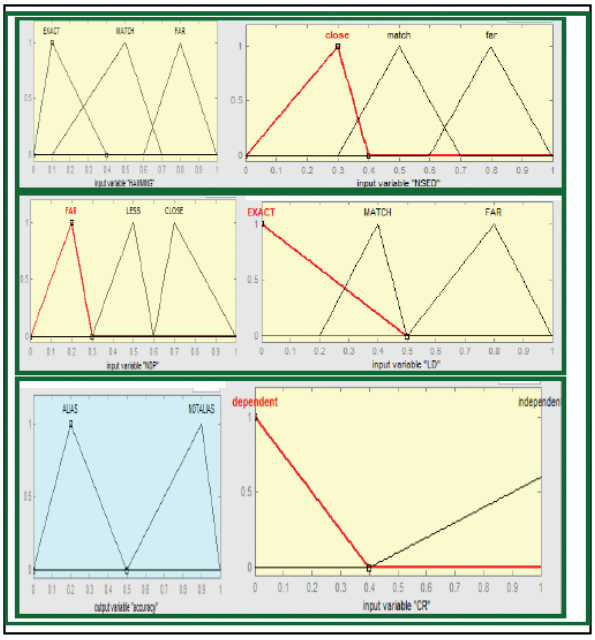

Figure 3. Input and output membership functions.

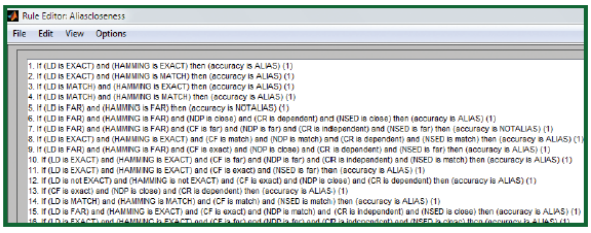

Figure 4. The sample developed fuzzy rules.

method is used in wider applications for fuzzy logic decision and obtains the centre of area occupied by the fuzzy set given by the formula

$$\text{c (centre of area)} = \frac{\sum_{i=1}^{n} = x_i \cdot \mu(x_i)}{\sum_{i=2}^{n} = \mu(x_i)}$$

Here, n represents the number of alias pair in the sample, x_i 's the elements and $\mu(x_i)$ is its membership function. In this work, n represents the total number of entity pairs and the closeness value is displayed in the form of crisp value as shown in Figure 5.

3. Result and Discussion

This section gives details of the experimental evaluation of the performance of the proposed system. The experiment is conducted on a core i5 with 4 GB RAM running on Windows 7 OS and alias dataset is taken from Auton Lab. The performance of the Alias closeness is also tested by determining the classification accuracy using the formula

$$\text{Accuracy} = \frac{\text{number of correctly identified}}{\text{Total number of true pairs}}$$

In this sample instance, it gains 90% accuracy for finding aliases. The performance of the system is verified with the given entity of abu_abdallah and its proposed alias names which are shown in Figures 6 and 7.

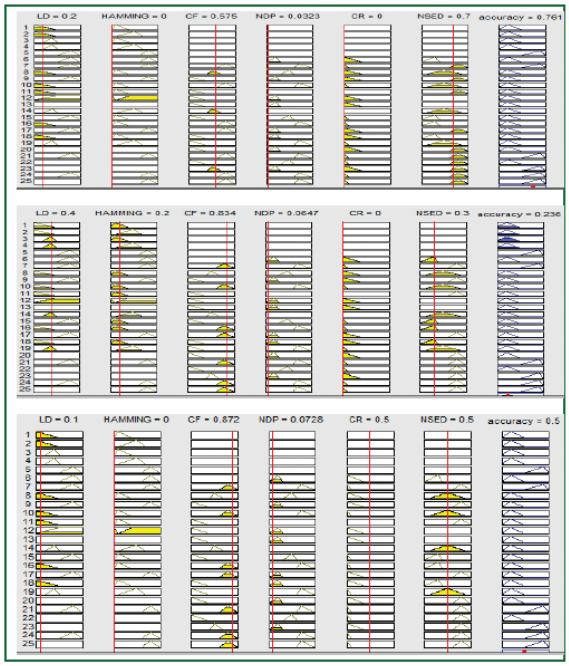

Figure 5. Sample defuzzification outputs.

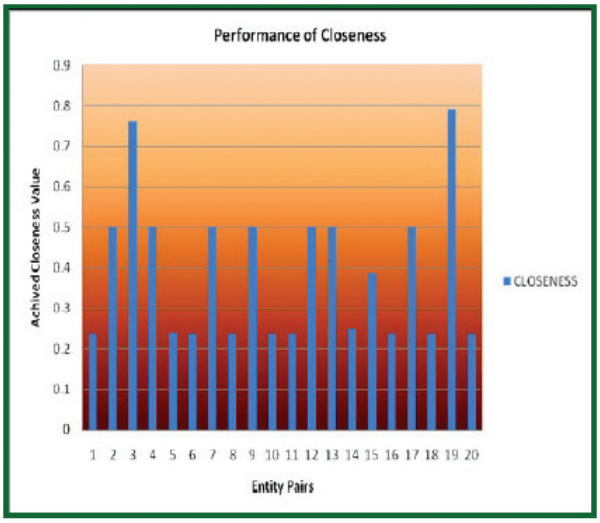

Figure 6. Performance of alias closeness.

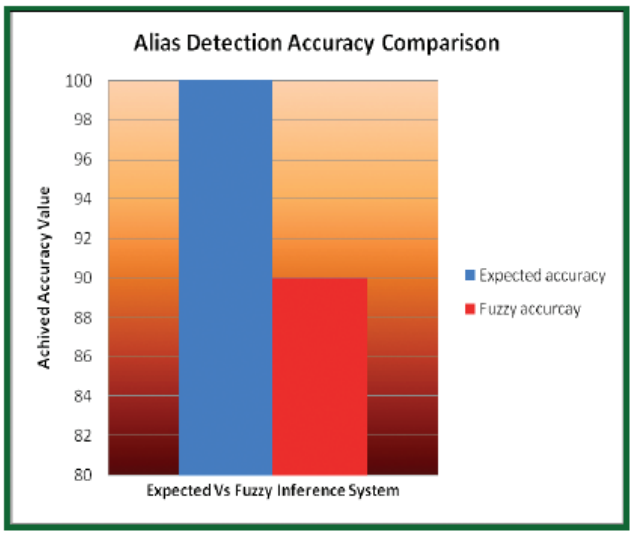

Figure 7. Alias detection accuracy of proposed fuzzy system.

4. Conclusion

Most of the real-time applications are fuzzy in nature, where there is a possibility of knowing the output in fuzzy value instead of the Boolean value.The fuzzy framework has involved the development of fuzzy inference rules using feature set, defuzzification and assessment of its performance. The achieved result shows that the proposed fuzzy-based framework has achieved an accuracy of up to 90%. These results imply that the developed fuzzy-based model is accurate and reliable to detect the person's alias names. In future, the proposed system is required to be assessed with other membership functions.

5. References

1. Paul H, Moore A, Neill D, Schneider J. Alias detection in Link Data Sets. Proceedings of the International Conference on Intelligence Analysis; Boston. 2005; 4:1–8.
2. An N, Jiang L, Wang J, Luo P, Wang M, Li B N. Towards detection of aliases without string similarity. Information science. 2014; 261(10):89–100.
3. Anwar T, Abulaishy M, Alghathbar K. Web Content Mining for Alias Identification: A first step towards suspect tracking. IEEE Conference on Intelligence and Security Informatics; Beijing. 2011. p.195–97.
4. Danushka B, Matsuo Y, Ishizuka M. Automatic Discovery of Personal Name Aliases from the Web. IEEE Transaction on knowledge and data Engineering. 2011; 23(6):831–42.
5. Shen Q, Boongoen T. Fuzzy Orders-of-Magnitude-Based Link Analysis for Qualitative Alias Detection. IEEE Transaction on knowledge and data engineering. 2012; 24(4):649– 63.
6. Bhat V, Oates T, Shanbhag V, Nicholas C. Finding aliases on the web using latent semantic analysis. Data and Knowledge Engineering. 2004; 49(2):129–43.
7. Meijuan Y, Xiaonan L, Jungong L, Xiangang L. A system for extracting and ranking name aliases in emails. Journal of software. 2013; 8(3):737–45.
8. Meijuan Y, Xiaonan L, Jungong L, Yongxing T. User Name Alias Extraction in Emails. International Journal on Image, Graphics and Signal Processing. 2011; 3(3):1–9.
9. Zadeh LA. Fuzzy Logic. IEEE computer. 1998; 22(4):83–93.
10. Santhanam T, Ephzibah EP. Heart Disease Prediction Using Hybrid Genetic Fuzzy Model. Indian Journal of Science and Technology. 2015; 8(9):97–803.