A Fuzzy Approach to Identity Resolution

Asif Nawaz and Hassan Kazemian

Intelligent Systems Research Centre School of Computing and Digital Media
London Metropolitan University London, UK
asn0144@my.londonmet.ac.uk

Abstract. Identity resolution is crucial for law enforcement agencies globally and a difficult task to match the real-world identity in big data due to data inconsistency e.g. typographical errors, naming variation, and abbreviations. The fuzzy approach to identity resolution has been introduced that uses Soundex and Jaro-Winkler distance algorithms in a cascaded manner to calculate an aggregate score for the full name. While the Edit-distance algorithm is used to score the address and ethnicity description attributes. The Soundex code has been modified to numbers only and increased the code length to 6-digits for this fuzzy approach. This allowed the matching algorithm to overcome some of the Soundex code limitations of name matching. The approach accommodates three different variations of name and an iterative search process retrieves matched records based on inputs. In the experiment, searching for a suspect in two different cases, the initial search retrieved 173 and 52 records for each target suspect. These records were grouped using the Mean-Shift clustering technique based on the similarity of three attributes. For further analysis, the segmentation process of records matched 16 and 22 records for each case respectively, and graph analysis matched the target suspect identity out of other matched identities with links association to different addresses. The overall matching performance of this fuzzy approach is encouraging, and it can benefit law enforcement agencies to speed up the investigation process and most importantly can help to identify the suspect with even minimal information available.

Keywords: Fuzzy string matching, Identity resolution, Graph analysis, Soundex, Jaro-Winkler.

1 Introduction

Identity Resolution is not just matching information but to detect, identify and consider past information or associations for the target entity. Fraud and other crimes are a globally major ongoing threat for law enforcement agencies and other institutions to identify the real-world identity from a pool of false or similar identities. Identity is the property and characteristics of an entity that helps to differentiate entities from each other. Each entity [1] has attributes and can be identified by ID numbers, names, and date of birth as key attributes but there are many reasons which make it difficult to find the correct identity. But matching two records using limited attributes is not sufficient for the true identity of any entity. For example, matching records by name will not resolve this issue as there can be many similar names in the database. One of the main issues is a huge amount of unstructured, incomplete, and incorrect data available to extract the required information for a particular individual [2].

The techniques, used for record matching and record linkage, normally classified records into three categories i.e. “Match”, “No Match” and “Possible Match”. But match or possible match might not be accurate if the information changes or not updated with time as this can lead to incorrectly flagged as a match or possible match [3]. The record matching in [4] refers to the entity resolution as information extraction using names while refers to identity resolution as a technique to determine the extracted information belongs to whom and how it is linked to others in real-world associations. This statement leads to find uniqueness and commonality between different datasets from different sources to answer who is who and who knows whom [5].

Data quality in the database makes it difficult to identify one real-world entity due to various similar multiple entries in the database. Remove or resolve duplicate and similar entries can be achieved by merging and de-duplicating records in the database representing the same entity. This is referred to as record matching in [6], [7]. But missing or incomplete information in the database leads to a huge amount of manual work to guess the matching record [8]. The data grows with the passage of time and traditional record matching techniques cannot accurately find a relationship between the records.

2 Literature Review

Fuzzy string matching is a technique to match strings based on approximate pattern similarity for entity resolution. A survey about duplicate record detection [9] explores the string similarity techniques developed for fuzzy string matching. The most common technique is Levenshtein distance also known as edit distance to calculate the distance between two strings by applying three edit operations on the strings. The three edit operations are inserting, deleting, or substituting the characters to match any given strings. If the distance of strings is less than the set threshold value after applying edit operations, then the strings are considered a close match with slight variation. But edit distance fails in the situation where strings are written in short form or abbreviation, instead of a complete word as it results in distance incorrectly. Jaro distance is another fuzzy string-matching technique, primarily to compare the short strings e.g. first and last names. This technique finds the common characters between strings while tracking the order of the characters [10]. The algorithm was enhanced by Winkler in 1990 by giving name prefix higher weighting. This variation was named as Jaro-Winkler distance. Jaro and Jaro-Winkler distance [9] algorithms cannot perform well if there is a positional difference between two strings and is more than allowed change. For example, strings “Alice bruce Bob” and “Bob bruce Alice” have allowed a change of 6 positions, but the character ’B’ in the string “Bob” has a difference of 12 positions. In this case, only string “Bruce” will match between two strings, and the algorithms will not find the better match.

For records matching, the record linking data association method [11] was introduced to match the criminal records referring to the same individual record. This method compares two records and calculates the total similarity measure as a weighted sum of the similarity measures of all corresponding featured values. The approach requires more computing power to calculate the total similarity as the dataset grows with time. Another similar work proposed in [12] to compare four personal attributes such as name, date of birth, social security number, and address for detecting identities by combining the overall similarity score. But the approach is limited and cannot produce accurate results if one or more attributes are missing from the dataset. It does not filter the referring records efficiently and can also ignore needed records on fewer similarity measurements. Another method discussed in [13] to remove duplicates from the dataset by applying dimensional hierarchy over the link relations such as city, state, and country. This approach matches the identity record only if both identities belong to the same area otherwise the record does not match against other similar entries in the dataset for different areas. The foreign keys in [14] were utilized in a relational database using a probabilistic relational model (PRM) for citation matching. The approach is rule-based and relies on the quality of data in the dataset and if the dataset has incomplete or missing information then the true match cannot be generated accurately.

To eliminate the duplicate records, [15] proposed another rule-based model called conditional random field model (CRF) to measure the associations among other different entities. However, [16] suggested that this approach fails and cannot find the links between similar entities. One of the interesting graph-based methods was proposed by [17] in which between each pair of the reference entity, the relational graph created, matching is based on similarity with the same attribute that matches with similarity measure. Furthermore, the approach was enhanced [18] by adding a collective entity resolution algorithm to match social information based on already matched records to reuse it for matching more records instead of only comparing two records. To match entities from social media websites e.g. Facebook and Twitter, the model in [19] combines user profiles using different attributes into a graph by detecting the social linkages between two user profiles using a CRF-based approach. The recent work in [20] proposed a rule-based approach to score attributes and analyze links between identities using a graph-based approach. The majority of the previous researches are rules-based techniques that can be very time-consuming to create a set of rules for big data.

Therefore, in this research, a fuzzy approach to identity resolution has been introduced that uses an iterative search with cascaded string similarity algorithms (Soundex code and Jaro-Winkler distance) to generate an aggregate score for the name variation. The fuzzy approach utilizes minimal attributes to retrieve the matching records from the dataset. The matched records then go through clustering processes to put similar records into clusters. For further analysis, the matching of these records is done by segmentation and graph analysis. This research carries forward previous work of [20] by using string similarity techniques with clustering, segmentation, and graph analysis.

3 Problem Definition for Identity Resolution

In the entity and identity resolution process, record matching and de-duplication is a difficult task to identify duplicates due to different attribute values. The techniques that require a human expert for manual tuning are better but unfeasible for large databases. Such databases can be referred to as big data. The techniques require training data samples to generate results for different situations. Other issues are related to the quality of data and techniques are not capable to use all the string similarity metrics to complete the matching process. This leads to unsatisfactory results because every single string similarity metric (fuzzy matching technique) is domain-specific that solves a certain problem. Missing one or more similarity metrics does not help to achieve better results. The discussed approaches do not consider the followings to achieve better results based on fuzzy matching:

- Use different similarity metrics techniques to calculate a matching score to extract records.
- Use similarity metric on data at different stages to output better entity matching results for record linkage and analysis.
- Use the clustering technique with help of segmentation for identity resolution.

4 The Proposed Fuzzy Approach to Identity Resolution

The fuzzy approach has been applied to identity resolution shown in Fig.1 by using cascaded string similarity techniques to retrieve an approximate entity match. The fuzzy approach utilizes Soudex, Jaro-Winkler algorithms to calculate the aggerate score for names and Edit-distance to score the other attributes e.g. ethnicity description and addresses. The aim is to match names simply by using similarity metrics and analyze retrieved records for similarities using clustering, segmentation, and graph analysis. This fuzzy approach is implemented using Python 3.7 using PyCharm (community version) IDE and the anonymized policy data is stored in MS SQL Server Express 2017. Pandas (a Python data analysis library) is used to clean data and store datasets retrieved during different stages. The NetworkX and matplotlib libraries are used for graph analysis and visualization.

4.1 Policing Dataset

An anonymized criminal dataset from police has been used in this research. The dataset consists of 1,146,212 records containing duplicates, typographical errors, incomplete or missing information. Some records are partial duplicates of other records with only one or more different attribute values. Each reported crime has a unique crime reference number and everyone in the dataset has a unique nominal reference number. But there are individuals with multiple nominal reference numbers. For example, there are 6,032 individuals with 5 or fewer similar nominal reference numbers assigned. The other attributes such as forename, surname, date of birth, gender, address, and ethnicity are representing an individual. In the dataset, there are 309,518 duplicate records based on surname, forename, and date of birth.

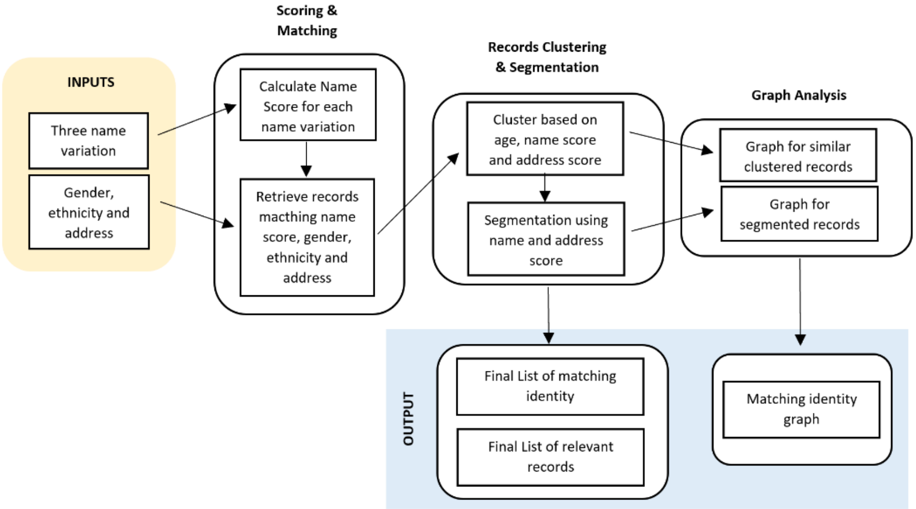
Scoring & Matching
Records Clustering & Segmentation
Graph Analysis
INPUTS
Three name variation
Gender, ethnicity and address
Calculate Name Score for each name variation
Retrieve records macthing name score, gender, ethnicity and address
Cluster based on age, name score and address score
Segmentation using name and address score
Graph for similar clustered records
Graph for segmented records
OUTPUT
Final List of matching identity
Final List of relevant records
Matching identity graph

Fig. 1. The block diagram of a fuzzy approach to identity resolution.

4.2 Data Pre-processing

The data pre-processing is used to eliminate missing values from the dataset for each record attribute. In the dataset, street_name, town_name, and district_name and postcode are four different fields representing the address. But in this research, the three fields are combined as one address field while keeping the postcode field separate. Similarly, the surname and forename are combined as name filed. There are records defined as 'Unknown' gender rather than 'M' for male' and 'F' for female, so in this fuzzy approach, these records are considered as missing values. Overall, there is a total of 430,293 missing values that are removed from the dataset. Therefore, the dataset after data cleaning still has 715,919 records. Considering the upper case and lower-case attribute values in the dataset, all attribute values are converted to lower case. The fuzzy approach utilizes name, gender, ethnicity description, and address attribute to start the initial search of records by generating the aggregated name score.

4.3 String Matching & Aggregated Score

This fuzzy approach uses Soundex, Jaro-Winkler similarity techniques to calculate the aggregate score for name matching. The Soundex code algorithm has been modified by removing the name first character as a constant letter from the code and changing the length to 6-digits. This generates a numerical Soundex code of 6-digits to help eliminate the Soundex first character mismatch issue. By increasing the length of the code helps to reduce many false-positive retrievals as compared to the 4-digits code. Later, the Jaro-Winkler technique is applied to this Soundex code to get a Soundex fuzzy score. All these scores are then used to calculate the aggregate score. The aggregate score is normalized between the fuzzy score of 0 and 1. The aggregate score is calculated with the following equation.

$$\mathrm{Agg}_{\mathrm{Score}} = (\mathrm{S}_{\mathrm{score}} + \mathrm{JW}_{\mathrm{score}}) * 0.5 \qquad (1)$$

4.4 Searching & Matching Criteria

It is very important to establish search and match criteria for retrieving entities by comparing the calculated score of selected attributes. The fuzzy approach focuses on eliminating most no-match entities during the initial search, based on aggregate score and Soundex code match. To find the results as accurately as possible is important as each record contains different attributes that help to differentiate one entity from other entities. But if there are issues in the value of the attributes e.g. incomplete information or typographical error then matching of records becomes difficult. This fuzzy approach proposes an iterative process by taking three different name variations as an input for the target entity. The approach also uses gender, ethnicity description, and address attributes to retrieve records. These selected attributes help to reduce the number of records retrieved by reducing processing time.

The matching of gender is done by an exact match while matching of ethnicity description is done by a partial matching of the string. For matching addresses, the edit distance technique is used to score addresses. Each iteration process is a combination of these selected attributes to generate search results as three data-subsets. Once the iteration process is completed, all three data-subsets are merged and compared for duplicate records to form one resultant dataset. All duplicate records are removed from this retrieved dataset. At this stage, records are retrieved even with the low aggregated score (e.g. score of 0.50 or 0.60) but have matching Soundex code. This to make sure any possible or close matched records are not ignored or dropped during the initial search.

4.5 Clustering, Segmentation, and Graph Analysis

In this fuzzy approach, the labeling of data from a human expert is not required to group similar records. For this purpose, the Mean-Shift clustering algorithm has been used to group similar records based on age, name, and address score. Each record is automatically labeled with a cluster number. These clustered records are then matched and compared based on the highest name and address score to create segments of records. This is to make sure similar records are linked together even if in different clusters by extracting record(s) with maximum address scores from clustered datasets to form a segment of records. In the segmentation process, the records are matched for similar addresses from the initial retrieved dataset and the clustered dataset. The similar segmented records are merged into one dataset and any other relevant records are kept separate.

For further graph analysis, these segmented records are compared with the clustered dataset to match the final identity out of all other identities. The graph creation is layer-based by using different attributes from the dataset. The first graph is created using the entity name and the clusters label for the entity. This visually list all entities linked to each cluster. The second graph is created using the entity name and address from the segmented dataset. This graph data is then compared with the first graph to find the matched identity out of other identities. The third graph is created to simply removed all the false positives and only show the matched identity with associated addresses.

5 Experimental Evaluation

5.1 Target Entity

For this fuzzy approach to identity resolution, it is assumed the police investigating officer is working on two different criminal cases and during an investigation, he obtained some basic information about the suspects. The information obtained per case is listed in Table 1 and investigating officer use this information to search a suspect for each case using this fuzzy approach. The investigating officer does not know the correct name spelling of the suspect or believes the dataset has typographical errors, so he inputs the full name with three different variations. The gender of the suspect is known while ethnicity description and address details are partially known.

Table 1. The information available to investigating office about the suspect for each case

	First Case	**Second Case**
Target Search (Suspect)	BECH Jaunette	FAROS Abbidah
Available suspect information for input		
Name Variation 1	back janette	abidah firos
Name Variation 2	bach janet	abiddha feroz
Name Variation 3	beck janete	abidha firoz
Gender	f	f
Description	white	white
Address	town and close	brandearth hey

5.2 Searching Results

First case. Based on the inputs for the first case in Table 1, the initial search and matching criteria produced three datasets. The name first variation retrieved 95 records; the second variation retrieved 61 records and the third variation retrieved 95 records. It is worth noting, the first and third name variation generated similar Soundex code (125300) and the aggregate score (0.70) that retrieved the same number of records. But the second variation generated a different Soundex code (122530) and retrieved a different number of records compared to the other name variations. A

resultant dataset of 173 records is generated after merging all three datasets and removing duplicate records.
Second case. Based on the inputs for the second case in Table 1, the initial search and matching criteria produced three datasets. All three name variations retrieved 41 records and interestingly the resultant dataset of 52 records is generated after merging all three datasets and removing duplicate records. All these records retrieved have some similarity or are completely different from one another, but this will be differentiated later in the next stages.

5.3 Clustering, Segmentation & Graph Analysis Results

The dataset (produced from the searching results) is fed into the clustering algorithm. This created, clusters of records based on age, name aggregate score, and address score. For the first case, records are grouped into 4 clusters, and records for the second case are grouped into 5 clusters based on the similarities score of the attributes. The clustering of records is shown in Fig. 2 and Fig. 3 for both cases respectively.

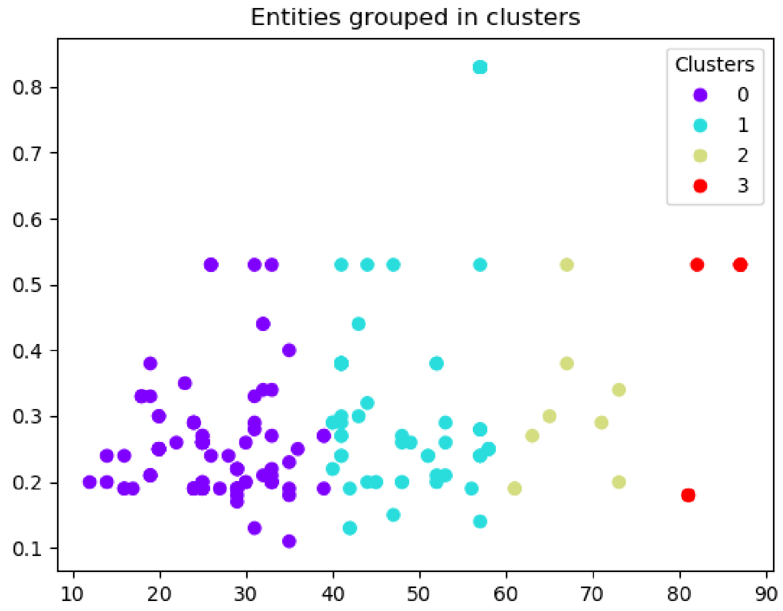

Fig. 2. First Case - Clustering based on age, aggregated name score, and address score similarity

The clustering of records does not identify the entity but provides a way to label and group records based on the score of attributes. Some of the records in clusters have a low score at the y-axis (address score) but are still required for the next stage to make sure not to ignore any matches, related or close matches. After clustering, the segmentation process picked the address with the highest score for the first case and second case as 0.83 and 1.0 respectively. For the first case, 7 records are retrieved with the

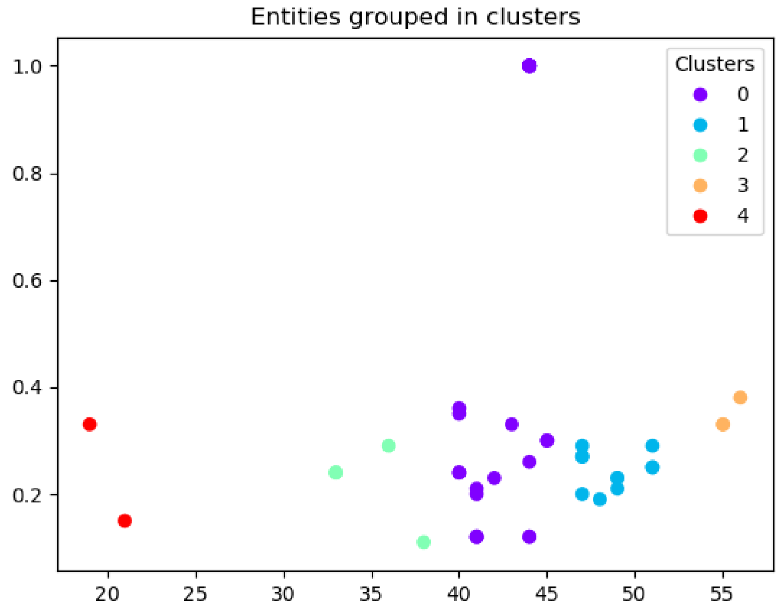

Fig. 3. Second case - Clustering based on age, aggregated name score, and address score similarity

highest address score and aggregated name score while 9 other records are retrieved with a similar name, low address score, and two different dates of birth. So, the final dataset retrieved for the first case contained 16 records with a combination of the same name, two different dates of birth, and different addresses. These records have some similarities but have unique crime reference numbers and duplicate nominal reference numbers for the same name associated with different addresses. For the second case, 16 records are retrieved with the highest address score and aggregate name score while 6 other records are retrieved with a similar name, low address score, and three different dates of birth. So, the final dataset retrieved for the second case contained 22 records with a combination of the same name, three different dates of birth, and different addresses. The records are with some similarities and have a unique crime reference number, duplicate nominal reference numbers associated with different addresses.

For graph analysis, the clustered dataset is used to create the graph with cluster ids as main data points associating several different entities linked to each cluster. Another graph is created from a segmented dataset based on the name and address. Both graphs are merged and compared as shown in Fig. 4 for the first case and Fig. 5 for the second case. The suspect is identified out of other entities and is highlighted with Red color in the graph associated with different addresses.

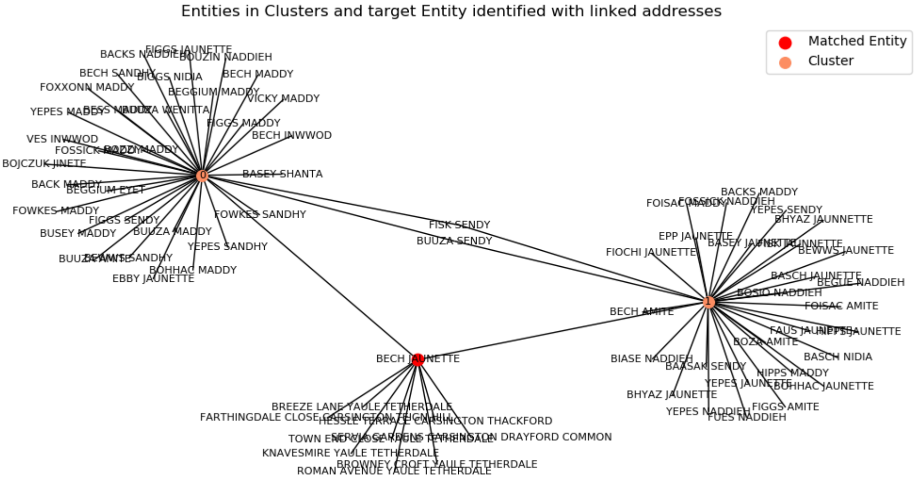

Fig. 4. First Case – Graph analysis, the suspect identified as "Bech Jaunette" highlighted red and associated to different addresses

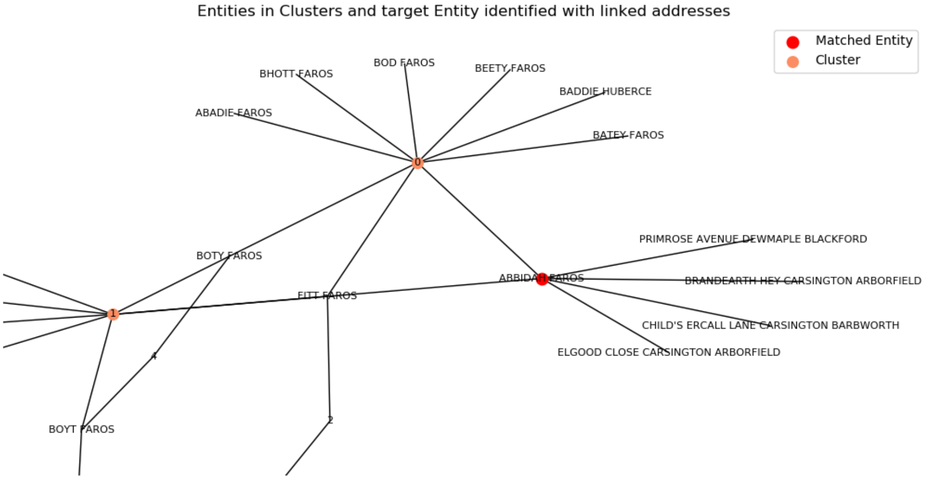
Entities in Clusters and target Entity identified with linked addresses
● Matched Entity
● Cluster
BHOTT FAROS
BOD FAROS
BEETY FAROS
BADDIE HUBERCE
ABADIE FAROS
BATEY FAROS
0
PRIMROSE AVENUE DEWMAPLE BLACKFORD
BOTY FAROS
ABBIDAH FAROS
BRANDEARTH HEY CARSINGTON ARBORFIELD
FITT FAROS
1
CHILD'S ERCALL LANE CARSINGTON BARBWORTH
ELGOOD CLOSE CARSINGTON ARBORFIELD
4
2
BOYT FAROS

Fig. 5. Second Case - Graph analysis, the suspect identified as "Abbidah Faros" highlighted red and associated to different addresses

5.4 Results Summary

The results discussed show that the fuzzy approach to identity resolution successfully identifies target identity (suspect) out of false identities from a huge dataset. During the evaluation, the results show the reduced number of records retrievals during the initial search and passing through different processes a final dataset was reduced significantly to make the matching process easier. This ensures law enforcement agencies can easily identify a suspect out of false or linked identities and can speed up the investigation process with minimal information in hand.

6 Conclusion

Identity resolution is a very important and crucial task for the law enforcement agency to identify the real identity of a suspect. This research introduced a fuzzy approach to identity resolution using string similarity techniques with a combination of clustering, segmentation of records, and graph analysis. This research is conducted on an anonymized policing dataset of 1,146,212 records and after data cleaning, a dataset of 715,919 complete records was obtained. The other main feature of this approach is minimal information available for selected attributes e.g. full name, gender, ethnicity description, and part of the address. The similarity algorithms are used in a cascaded manner to calculate the full name aggregate score. The iterative search retrieved records based on the name variation and matching of selected attributes. These records are merged into one dataset and duplicates are removed making the dataset for clustering of records using the Mean-Shift algorithm. Based on the experiment, the search for two suspects created 4 and 5 clusters for records respectively. Later, the segmentation process picked the records with the highest address score to search for similar addresses from the clustered dataset and initial search generated dataset. This process ensured to pick any relevant records that may have missed during the initial search for the suspect's identity. The graph analysis linked all identities on basis of selected attributes and after comparing graphs the suspect identity was identified associated with different addresses. Considering overall results, the fuzzy approach to identity resolution can be very handy for law enforcement agencies to find real identity using minimal information available about any suspect.

In future research, this fuzzy approach can be improved by introducing a weighting system for attribute scores and complete the incomplete records with available information in the dataset or complete records with predicted information for better identity resolution. It will be worth using machine learning techniques to generate a knowledge base that grows with each identity search and simplifies the future search process.

References

1. Wang, G., Chen, H. & Atabakhsh, H., 2004. Automatically detecting deceptive criminal identities. *Communications of the ACM - Homeland security,* 47(3), pp. 70-76.
2. Barkay, D. & Dror-Rein, E., 2015. Achieving Cyber Identity Resolution via Electronic Warfare Techniques. Sinapore, RSAConference
3. Duncan, J. et al., July 2015. Building an Ontology for Identity Resolution in Healthcare and Public Health. *Online Journal of Public Health Informatics,* 7(2).

4. Roth, D. & Ratinov, L., 2009. Who's Who in Your Digital Collection: Developing a Tool for Name Disambiguation and Identity Resolution. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science (DHCS),* pp. 1-17.
5. Clendenen, C., 2009. A New Approach to Workers Compensation Fraud. *IAIABC Journal: Introducing Identity Resolution,* 46(1), pp. 103-114.
6. N, M. R. and Alankar, R. (2016) 'Detection of Fuzzy Duplicates in High Dimensional Datasets', pp. 1423–1428.
7. Bilenko, M. and Mooney, R. J. (2003) 'Adaptive duplicate detection using learnable string similarity measures', *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 39–48. doi: 10.1145/956755.956759.
8. Mon, A. C., Mie, M. and Thwin, S. (2013) 'Effective Blocking for Combining Multiple Entity Resolution Systems', 2(4), pp. 126–136.
9. Elmagarmid, A. K., Ipeirotis, P. G. and Verykios, V. S. (2007) 'Duplicate record detection: A survey, IEEE Transactions on Knowledge and Data Engineering, 19(1), pp. 1–16. doi: 10.1109/TKDE.2007.250581.
10. Gomaa, W. and Fahmy, A. (2013) 'A survey of text similarity approaches', International Journal of Computer Applications. Foundation of Computer Science (FCS), 68(13), pp. 13–18. doi: 10.5120/11638-7118.
11. Brown, D. E. & Hagen, S., 2003. Data association methods with applications to law enforcement. Decision Support Systems, 34(4), pp. 369-378.
12. Wang, G., Chen, H. & Atabakhsh, H., 2004. Automatically detecting deceptive criminal identities. Communications of the ACM - Homeland security, 47(3), pp. 70-76.
13. Ananthakrishna, R., Chaudhuri, S. & Ganti, V., 2002. Eliminating Fuzzy Duplicates in Data Warehouses. Hong Kong, China, VLDB Endowment, pp. 586-597.
14. Pasula, H., Marthi, B., Milch, B., Russell, S. and Shpitser, I. (2003) 'Identity uncertainty and citation matching', Advances in Neural Information Processing Systems, pp. 1425–1432. doi: 10.1.1.15.8644.
15. Culotta, A. & McCallum, A., 2005. Joint deduplication of multiple record types in relational data. Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 257-258.
16. Li, J. & Wang, A. G., 2015. A framework of identity resolution: evaluating identity attributes and matching algorithms. *Security Informatics - A SpringOpen Journal,* pp. 1-12.
17. Bhattacharya & Getoor, L., 2006. Entity resolution in graphs. In: Mining graph data. s.l.:Wiley-Blackwell, p. 311.
18. Bhattacharya & Getoor, L., 2007. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(5)
19. Bartunov, S., Korshunov, A., Park, S., Ryu, W. and Lee, H. (2012) 'Joint Link-Attribute User Identity Resolution in Online Social Networks Categories and Subject Descriptors', The Sixth SNA-KDD Workshop Proceedings.
20. Phillips, M., Amirhosseini, M.H., Kazemian, H.B., 2020. A Rule and Graph-Based Approach for Targeted Identity Resolution on Policing Data, in: 2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020. Institute of Electrical and Electronics Engineers Inc., pp. 2077–2083. doi:10.1109/SSCI47803.2020.9308182