International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 8 26 – 30

Identity Resolution across Different Social Networks using Similarity Analysis

[1]Neha Talokar, [2]Swati Mali
Dept. Of Computer Engineering
K. J. Somaiya College of Engineering
VIDYAVIHAR, MUMBAI
[1]*neha.talokar@somaiya.edu*, [2]*swatimali@somaiya.edu*

Abstract—Today the Social Networking Sites have become very popular and are used by most of the people. This is because the Social Networking sites are playing different roles in different fields and facilitating the needs of its users from time to time. The most common purpose why people join in to these websites is to get connected with people and sharing information. An individual may be signed in on more than one Social Networking Site so identifying the same individual on different Social Networking sites is a task. To accomplish this task the proposed system uses the Similarity Analysis method on the available information details.

Keywords-*Social Networking Sites, Identity Resolution , Similarity Analysis.*

*****

I. INTRODUCTION

The Social Networking Sitesare very popular today as they are contributing in most of the fields in some or the other way. The information available from these Social Networking sites abides to the rules and laws of privacy.Every social network stands for different purpose, and so, a user may be present on various online social networks, not necessarily under the same name.

Identity resolution is a process where an identity is searched and analysed between disparate data sources to find a match/resolve identities.

Social Networking Sitesare also termed as Online Social Networks (OSNs). OSN can be defined as a network of social interactions and personal relationships. It is a dedicated website or application which enables users to create and share information, ideas, career interests and other forms of expression.

Today most of the users are available on the social networking sites. The user's publically available information can be captured from various OSNs to create a detailed user profile. Generally human beings identify others with their name as identification mechanism. Sometimes more than one user share same profile name. This creates a need to identify the correct user profile from all the rest of the ones. Identifying the correct profile of an individual from the available profiles can be termed as identity resolution on social networks.The growing number of users on social networks and also the growth in cybercrimes, marketing, fake profile proliferation, cyberbullying, trolling has raised the need of identity resolution across social networks.

Identity Resolution across Social Network datasets aims at identifying the same user across two different social networking sites based on the similarity analysis. Similarity analysis is done on the dataset obtained from publically available user profile information. Similarity analysis deals with the demographic attributes (Username, Email id, Location, Birthplace, Gender, Age, Education, Hobby, Status, etc.) and the post content matching.

The work done by the researchers and various existing model information is mentioned in section II. The proposed model and its details are mentioned in section III.Section IV mentions some results obtained after partial implementation of the proposed system.

II. RELATED WORK

The techniques for Identity Resolution across OSNs are evolving due to the threats and challenges imposed by the growth in number OSN users.The approach to Identity Resolution involves extracting the information of individuals from OSNs, storing in database and applying the techniques for resolving the Identities.

The study of large online social footprints by collecting data on 13,990 active users is presented in [1]. The data from 10 of the 15 most popular social networking sites is parsed to find general statistics of how much user data is publically available and the possibility of reconstruction of the user profile by using the users' online footprint by attacker.

In [2] the system applies various automated techniques along with the online digital footprints of individual on one social network to identify the same individual on other social networks.It finds UserID and

Nameas the most discriminativefeatures. It helps toanalyze and compare two different social networks.

In [3] novel identity search methods are proposed to increase the accuracy of identity resolution process in online social networks. It proposes to divide Identity Resolution into two sub-problems, namely Identity Search and Identity Matching. Identity search includes Profile Search, Content Search, Self-mention Search and Network Search for searching identity on Facebook by exploiting information on Twitter.It also mentions Syntactic Matching and Image Matching as Identity Matching methods.

The system for searching and matching individuals on Facebook and MySpace is discussed in [4]. The system describes search methodology using threshold. It uses classifier to find match among candidate sets and describes the Matching methodology, Matching Validation and Threshold Selection and Analysis of Data to draw conclusions.

In [5] the automated methods for Identity Resolution across heterogeneous social platforms are discussed which include the Identity Search and Identity Linking methodologies along with the evaluation. For profile linking it uses cascaded framework where Classifier I extracts current username features and use an existing method to classify username sets while Classifier II extracts proposed username set features and use a supervised classifier to re-classify username sets labelled as negative by Classifier I.

The HYDRA framework [6] is proposed for the Social Identity Linkage based on the Heterogeneous behavior. The Hydra framework consists of three steps which include (a) learning the similarity of attributes and long-term topical distribution analysis (b) building structure consistency model to maximize structure and behavior consistency model on users' social structure and linkage could be performed on the basis of group of users (c) a normalized-margin-based linkage function is proposed where it learns multi-objective optimization is performed on the results of above two step towards Pareto optimal solution.

A framework that gives more importance to some attributes and allows users to assign a different similarity measure to each attribute [7] in order to find proper match is proposed and experimented to compare the superiority of results with the current ones.

The username is always unique to a user and is commonly visible to all users. The user name as it not secret and can be used to identify and link profiles [8]. The profiles are sometimes linked but have a different user names for different OSN. So here the uniqueness of username could be studied to link the records to get better results based on the calculation of probability that the username belongs to same user.

The methodology (MOBIUS) [9] to identify unique behavioral patterns and redundancies across different sites and then constructing the feature set to exploit the information redundancies that can be used by machine learning for user identification is defined. The different user behavioral patterns while selecting the usernames are studied.

The similarity measures could be used to identify the duplicate records [10]. The field matching techniques that could help detect duplicate records are discussed. It also mentions the various ways to improve efficiency and scalability of duplicate data detection algorithms.

Brief mention of existing tools and the problems are also listed and discussed. The various techniques and similarity measures that could be applied to various types of data are compared [11] to find the proper matched result. The performance of various matching methods on Census data and results are mentioned which helps comparison of methods for better understanding.

The various String metrics for matching Names and Records that are available in Java toolkit and the distance metrics which include distance functions are described and compared to draw results based on the performance of the distance .metrics [12].

The Levenshtein method in combination with the Smith Waterman algorithm is used for Plagiarism Detection [13]. It describes how Levenshtein distance can be used to change the likely scarcity, which can improve both time and space efficiency.

The Levenshtein algorithm is used for detecting Question Bank Similarity [14].It discusses the problem of Levenshtein algorithm along with the solution to improve the performance by overcoming the problem using segmentation.

The user profiles are compared using vector based comparison algorithms and Vector Field matching [15].The Vector Field Matching are mentioned based on their method of matching.

III. PROPOSED SYSTEM

The proposed system aims at identifying the profile of same individual on different Social Networking Sites and integrating the profile information for building a detailed user profile.

The system considers two social networking sites say S1 and S2.The datasets d1 and d2 belong to the sites S1 and S2 respectively. Datasets d1 and d2 contains record of users of S1 and S2 respectively. The system has to process d1 and d2 such that the user with an identity in d1 finds a resolved identity in d2 provided the user is present on site S1 as well as on S2.

The proposed system is explained in the following sub-sections.

A. *Architecture*

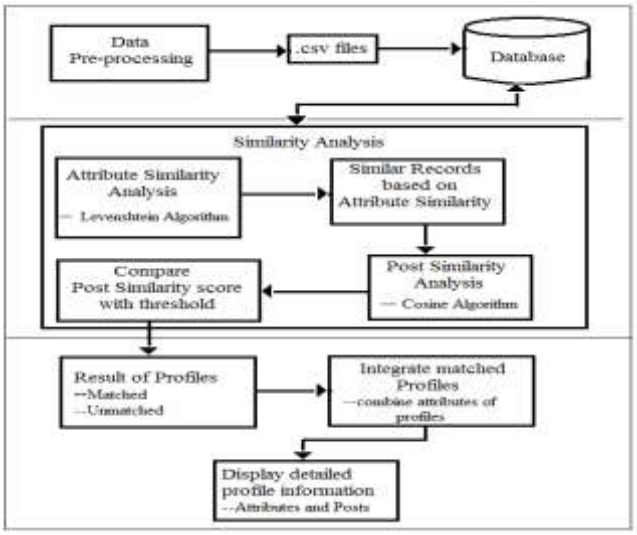

Figure 1. Proposed system architecture

Flow of the Proposed System

1. Creating and uploading datasets
2. Defining thresholds
3. Calculating Similarity Analysis Score
   a. Attribute similarity analysis score(S_A)
   b. Post similarity analysis score(S_P)
4. Display Result
5. Display detailed user profile

B. *Creating and uploadingdatasets*

The proposed system has two datasets as input. Each dataset consists of records from different Social Networking sites. The datasets are created manually by using the APIs' made available by the Social Networking sites to extract user information, by manually searching down the known user profiles on the particular Social Networking sites and by generating dummy datasets with the help of online data generation websites. The datasets created are pre-processed.Pre-processing involves the cleaning of data, removing redundancies or duplicate data, checking for missing data, etc.

The dataset under consideration is constructed such that the two datasets belong to two different social networks and have the following characteristics: (a) the user has accounts on both the social networking sites and shares information publically (b) the user has registered with correct information (c) the user is not biased and shows similarity in behavior on both the networks.The two datasets have different fields (attributes) as in the Table I.

TABLE I. FIELDS IN DATASET AND THRESHOLDS

Attribute No.	Dataset1	Dataset2	Threshold
1	User name	User name	0.5
2	Email id	Email id	0.4
3	Location	Location	0.2
4	Birth place	Birth place	0.5
5	Gender	Gender	1.0
6	Date of Birth	Date of Birth	1.0
7	Education	Education	-
8	Status	Status	-
9	*Job profile	*Hobby	-
10	*College name	*School name	-

The attributes under consideration are the publically available attributes that could be collected for different users provided there are no privacy restrictions applied by the user. Each of the two datasets contains 100 records. The dataset files are in .csv file format.

The posts belonging to each user are given as input to the system and the file is a .txt file.

C. *Defining threshold*

The threshold is the minimum acceptable value which is set in the system based on which the system is able to give decisions as per the logics defined. The proposed system has threshold defined for the attributes and the posts. The threshold values are decided based on the heuristic approach. Heuristic approach is an approach where the methods applied for achieving a goal are based on practical observations and experimentation. The outcomes of heuristic method approach may or may not be perfect and optimal but helps in achieving the desired goal. The threshold values are declared for the first 6 attributes and are specifically mentioned in Table I .

D. *Calculating Similarity Analysis Score*

The similarity analysis aims at calculating the similarity for finding a match between fields under consideration. The similarity analysis module calculates the similarity based on the similarity measure and algorithm defined at the time of designing the system. The selection of similarity measure and algorithm is very important as it affects the performance of the system and the results given by the system. The subsections below gives brief information on the analysis and selection of similarity measures.

a. Attribute similarity Analysis Score

The attributes under consideration for the record matching are listed in Table1. It is observed that the attribute values are in text format except that for Date of Birth and further the values are single words and not collection of words. So while selecting the similarity measure we select it such that it is capable of finding character based similarity irrespective of the data which may be in text or numeric form.

The attribute values for each record are checked to find a match between the records in different datasets. The match is found using the similarity measure which gives the percentage

similarity score (S_A) that ranges between 0 and 1 i.e. $0 \leq S_A \leq 1$.The similarity measure used in the system is Levenshtein distance.

b. Post Similarity Analysis Score
The post is in text form which is collection of words and sentences. So the similarity for these posts needs to be calculated using the similarity measures which are token based/term-based. The most commonly and widely used term-based similarity measure is Cosine-similarity and the proposed system also aims to use the same.

E. Display Result

The result displayed is the list of users' after calculation of the Attribute Similarity Analysis Score comparison with the thresholds and conditions with a view button to get the detailed user profile (Fig. 2).

F. Display detailed user profile

The detailed user profile displayed (Fig. 3) is the result obtained by integrating the similar user's profile details from dataset1 and dataset2.

IV. RESULTS

The partial results are presented here. The Attribute Similarity Analysis Score calculation is implemented and the results obtained are as follows:

a. Similarity Result: It is the result obtained after applying similarity measure algorithm to find similar user profiles based on demographic information in dataset1 and dataset2. Score represents the percentage similarity between the user profiles from site1 and site 2.

Site 1 Username	Site 2 Username	Score	View Profile
Benedict	Benedict	100	view
Imelda	Imelda	100	view
Ivana	Ivana	100	view
Jonah	Jonah	100	view
MacKenzie	MacKenzie	100	view
n	nt	83.33333333333334	view
neha	neha	98.88888888888889	view
ram	ram	89.2156862745098	view
Winifred	Winifred	100	view
xyz12	xyz1	95.47619047619048	view

Figure 2. Similarity Result

b. Detailed user profile: It is the result obtained by integrating the similar user profile details from dataset1 and dataset2 (social networking site1 and site2).

Attributes	Site1	Site2
Username	Ivana	
Email-ID	molls.vitae@cursusdiamet.ca	
Location	Samoa	
Birth Place	Georgia	
Gender	female	
Age	26	
Education	MCA	BE
Status	Married	Single
Job_Profile	Asset Management	-
Hobby	-	reading
College Name	Sardar Patel College of Engineering	-
School Name	-	Activity High School

Figure 3. Detailed user profile

The result in above figure (fig. 3) is explained as follows: if the profiles on site 1 and site 2 share same information then there is no entry under site 2 column and if the information differs then the entries are present under site 1 as well as site 2 column and '-'(dash) in the rows represent that there is no field available for the particular site.

V. CONCLUSION & FUTURE WORK

The proposed system aims to link profile records using similarity analysis applied on the details of the users available in the datasets. The datasets contain the information details which are generally available through the social networking APIs'. The proposed system is implemented partially and hence the results presented here are also partial. The results presented here include the results after calculating the similarity score and applying certain threshold conditions. Also it includes the result after integrating the profile records to create better and detailed user profile.

Further in future the aim is to implement complete proposed system i.e. by considering the post details of the users to identify and link them along with the demographic details.

REFERENCES

[1] Irani, Danesh, Steve Webb, Kang Li, and Calton Pu. "Large online social footprints--An emerging threat." In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 3, pp. 271-276. IEEE, 2009.

[2] Malhotra, Anshu, Luam Totti, Wagner Meira Jr, Ponnurangam Kumaraguru, and Virgilio Almeida. "Studying user footprints in different online social networks." In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 1065-1070. IEEE Computer Society, 2012.

[3] Jain, Paridhi, Ponnurangam Kumaraguru, and Anupam Joshi. "@ i seek'fb. me': identifying users across multiple online social networks." In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1259-1268. ACM, 2013.

[4] Motoyama, Marti, and George Varghese. "I seek you: searching and matching individuals in social networks." In *Proceedings of the eleventh international workshop on Web information and data management*, pp. 67-75. ACM, 2009.

[5] Jain, Paridhi. "Automated Methods for Identity Resolution across Heterogeneous Social Platforms." In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pp. 307-310. ACM, 2015.

[6] Liu, Siyuan, Shuhui Wang, and Feida Zhu. "Structured Learning from Heterogeneous Behavior for Social Identity Linkage." *IEEE Transactions on Knowledge and Data Engineering* 27, no. 7 (2015): 2005-2019.

[7] Raad, Elie, Richard Chbeir, and Albert Dipanda. "User profile matching in social networks." In *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pp. 297-304. IEEE, 2010.

[8] Perito, Daniele, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. "How unique and traceable are usernames?." In *International Symposium on Privacy Enhancing Technologies Symposium*, pp. 1-17. Springer Berlin Heidelberg, 2011.

[9] Zafarani, Reza, and Huan Liu. "Connecting users across social media sites: a behavioral-modeling approach." In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 41-49. ACM, 2013.

[10] Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. "Duplicate record detection: A survey." *IEEE Transactions on knowledge and data engineering* 19, no. 1 (2007): 1-16.

[11] Bilenko, Mikhail, Raymond Mooney, William Cohen, Pradeep Ravikumar, and Stephen Fienberg. "Adaptive name matching in information integration." *IEEE Intelligent Systems* 18, no. 5 (2003): 16-23.

[12] Cohen, William, Pradeep Ravikumar, and Stephen Fienberg. "A comparison of string metrics for matching names and records." In *Kdd workshop on data cleaning and object consolidation*, vol. 3, pp. 73-78. 2003.

[13] Su, Zhan, Byung-Ryul Ahn, Ki-Yol Eom, Min-Koo Kang, Jin-Pyung Kim, and Moon-Kyun Kim. "Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm." In *Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on*, pp. 569-569. IEEE, 2008.

[14] Shao, Min-Min, and Dong-Mei Qian. "The Application of Levenshtein Algorithm in the Examination of the Question Bank Similarity." In *Robots & Intelligent System (ICRIS), 2016 International Conference on*, pp. 422-424. IEEE, 2016..

[15] Vosecky, Jan, Dan Hong, and Vincent Y. Shen. "User identification across multiple social networks." In *Networked Digital Technologies, 2009. NDT'09. First International Conference on*, pp. 360-365. IEEE, 2009.