Empirical Investigation of Finding 'Person' in Social Network using Clustering

Trupti S. Indi/Gunge
M.E. (CSE) - appearing
Walchand Institute of Technology
Solapur University
India

Raj B. Kulkarni
Asst. Prof. Dept. CSE
Walchand Institute of Technology
Solapur University
India

Dr. S.K.Dixit
Professor Dept. E&TC
Walchand Institute of Technology
Solapur University
India

ABSTRACT

We are trying to identify person as an entity from social network data by grouping the person with some characteristics. It is very difficult to identify people on social networks especially for specific names in different cultures like India, and other ethnic places whose names are difficult to trace. This is a small effort to bridge the gap of identifying person with features like name, date of birth and address containing state, city etc. We are trying to make use of clustering algorithms to uniquely identify people/person on social networks. An empirical analysis of the same is presented here.

KEY WORDS

Social network, clustering, features, CURE

1. INTRODUCTION

There are different ways to identify person entity from collection of persons. We are proposing the system to group the person using valuable features and get the results. When any action becomes more complex we try to simplify that action by dividing an action into number of sub-actions. We are using the similar concept to solve the complex problem that is 'identifying the person from collection of persons'. Here we are dividing clustering process in two phases. In initial phase, we are using some common features of person, like name, date of birth and address, for grouping the person with identical features. Once the persons are clustered with the initial characteristics then from defined clusters again we split the persons using some external features available in the database and get minute level clusters which gives more accurate or better result than previous stage. The system presented here uses the tagged and preprocessed data. The favorable point in this system is that after first level of grouping it allows adding external features for clustering which will not be fixed to the system, user can vary these features as and when needed but following rules of the system.

2. RELATED WORK

There is considerable work going on personalized search, discovering and using groups to improve personalized search, extracting the social networks and contact information from email and the Web, an advanced social network extraction system from the Web etc. In the mid-1990s, H. Kautz and B. Selman developed a social network extraction system called the Referral Web [7].To better understand whether groups of people can be used to benefit personalized search, Teevan et al. explored the similarity of query selection, desktop information, and explicit relevance judgments across people grouped in different ways [4]. Culotta et al. presented the system which identifies unique people in email, finds their Web presence, and automatically fills the fields of a contact address book using conditional random fields—a type of probabilistic model well-suited for such information extraction tasks [5]. Matsuo et al. proposed a social network extraction system called POLYPHONET, which employs several advanced techniques to extract relations of persons, to detect groups of persons, and to obtain keywords for a person. Search engines, especially Google, are used to measure co-occurrence of information and obtain Web documents [6]. Jussi Kurki defined a simple ontology for describing people and organizations. The model is based on FOAF and other existing vocabularies. In actor ontology, using ONKI people service for searching and disambiguating people. An ONKI person is nothing but a centralized repository of persons and organizations [8].

There are different clustering techniques explained by Jain et al. in "Data Clustering: A Review" ACM Computing Surveys, Vol. 31, No. 3, September 1999 [1]. The paper presents an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners.

Traditional clustering algorithms either favor clusters with spherical shapes and similar sizes, or are very fragile in the presence of outliers. Guha et al. proposed a new clustering algorithm called CURE that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction [3]. The clustering algorithm presented here also based on the CURE method. It basically contains cluster representative which used to compare new item for decision making.

DOUGLAS H. FISHER presented the concept of conceptual clustering that organizes the data in such a way that to maximize the inference ability. This paper describes the problems in conceptual clustering like clustering problem, characterization problem and COWEB system which is an incremental conceptual clustering system [9]. There are systems implemented to generate some visual representation of social network using clustering algorithms. For example Chan Stefani et al. implemented system to generate visual representation of a social network from data, like author, co-author and citation, obtained from the web. In this they have worked on algorithms to cluster authors based on mutual research interest [10].

We are trying to develop a system which helps to identify person in database. These databases can be building from any other easily available database like social network database or government agencies database or telephone directories information or electricity billing system database. Furthermore Teevan and group presented the concept of personalized search based on the data from other people that is group of people information used to personalized search. They used different grouping types like task-based and trait-based. We are representing a concept which concentrates only on person's individual information and not other people information relate to in at initial level for search.

3. PERSON CLUSTER

In person clustering, cluster is formed by group of "person" elements and to decide which person belongs to which cluster different characteristics or features of a person is used. Here each feature carries some weight to decide nearest possible cluster.

Person, an entity, bounded with various features. After Person born, the "name" gets attached with him along with its other facial features like eyes, face, color etc. When kid starts going to school then the "name of school" means to which school he belongs gets attached with him. Like this when as a student, as a professional, as a social entity of society the role and characteristics of the person changes.

Based on role and application, the properties by which we identify the person will differ. We have worked on some test data of the student data collected from college and applied the clustering algorithm discussed in this paper.

A group of persons form network to share knowledge, experience, to make friends etc. Based on the aim of forming the network the features used to search the person in the network changes. We presented some basic network formed by collection of person.

3.1 Different Networks Formed By Person Entity:

We are describing basic three networks where person (user) gets identified as a unique entity.

1. Social Network
2. Professional Network
3. Business Network

3.1.1 Social Network:

In the social networks, like Orkut and Facebook, user enters his/her profile details by its own knowledge. The information entered in such social network forms the "personal profile" of the user.

In such networks the information entered by user is sometime incomplete or may be incorrect. The reasons are many like user is not interested in completing the information or he is not frequent user of computer etc. Each user has its own reason to join these networks like messaging to friends, making new friends, knowing interest of the person which will open new business opportunities, forming specific community of interest etc.

Examples of social networks are Orkut, Facebook.

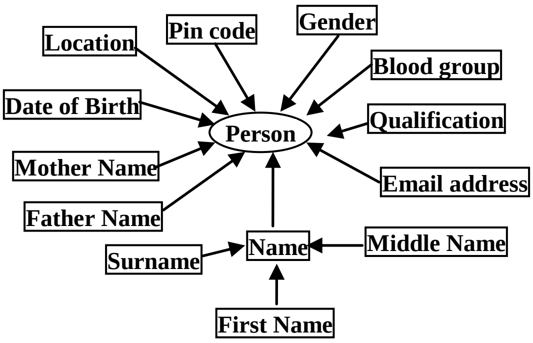

Figure1: Social Network related features forming personal profile

3.1.2 Professional Network:

Professional network is a network build from collection of professionals. LinkedIn website is an example of professional network where user, a person who is engaged in some of learned profession, can enter his/her professional details and then he can collaborate with other professionals through this network. The details entered in professional network by professional forms "professional profile" of user.

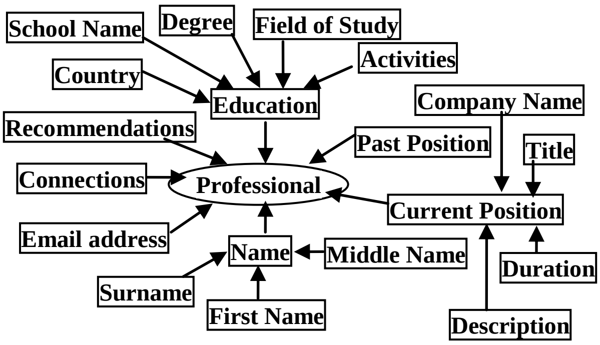

Figure2: Professional Network related features forming professional profile

In such networks people with same thinking collaborate with each other and share their business ideas, their experience in their profession, get recommendations from other people, knowledge sharing, etc. which in turn helps in growing as professional as well as in business. As a Professional the features get considered here from a person are different than in social network. In these kind of networks, more highlighted features will be user's education, past experience, technologies he know, his current working location, any research work he did, etc.

3.1.3 Business Network:

Business network is one form of social network but not same. In these networks, the people with similar domain get interact with each other. These networks can be formed by the people from same business community. These are a kind of social networking sites for business owners, entrepreneurs and professionals. The need of user in such sites is analyzed differently than in a social network.

Examples of business networks are Bizlink, Ryze.

Ryze is a business networking community which helps users to organize themselves by interest, location, and current and past employers. Organization can set up a network on Ryze to help their members connect with each other and also it helps to recruit new members and publicize your events.

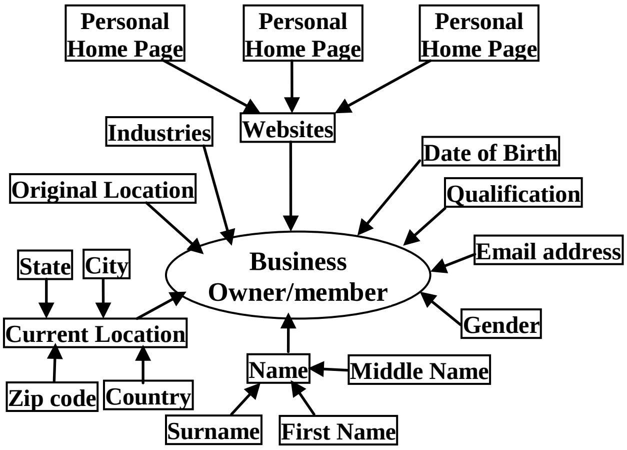

Figure3: Business Network related features forming business profile

4. CLUSTERING

Clustering has been used in many application domains, including biology, medicine, anthropology, marketing, and economics. Cluster applications include many different applications like plant and animal classification, image processing, disease classification, document classification and pattern recognition [2].

Clustering of data is a method by which large set of data are grouped into clusters of smaller sets of similar data.

We are dividing clustering algorithm into mainly two phases.

1. Phase-I
2. Phase-II

Along with these phase, we are considering features of two categories:

1. Internal features and
2. External features.

The Internal features can be used in first phase of clustering and external features used in second phase of clustering and third phase of clustering should be present or not will be decided based on number of available external features.

4.1 Person Clustering Algorithm:

Phase-I:

In this phase, internal features which are non-modifiable used are the following one:

f1: Name (Surname and First Name)
f2: Date Of Birth (dd-mm-yyyy)
f3: Address (Country, State, and City)

Using above feature set, the first set of clusters (C1, C2, C3… Cn) will be formed.

Phase-I Clustering Algorithm:

```
Input:
    P    = {P1, P2, …, Pm} //Set of elements
Output:
    C    = {C1, C2, …, Kn}            //Set of clusters

Algorithm Steps:
1.  Start
2.  $C_{map}$ : empty
3.  Add $C_1$ to $C_{map}$ as Representative
4.  For i = 2 to m do
        i. If $C_{map}$.IsExists($C_i$) Then
        ii. continue;
        iii. Add $C_i$ to $C_{map}$ as Representative
        iv. For j = i+1 to m do
        v. If $C_i$ equalTo($C_j$) Then
            a. Add $C_j$ to $C_{map}$ where $C_i$ is Representative
        vi. End If
        vii. End of j For
        viii. End of i For

5.  End
```

Phase-II:

For this phase, the number of clusters formed in Phase-1 will be the input along with external features set like education details or social information or other attributes.

f4: email address (String)
f5: Father Name (String)
f6: Mother Name (String)
f7: School Name (String)
f8: School Address/city name (String)
And many more

Using above features set, input clusters C1, C2, C3, … Cn will be further split into new clusters (C11, C12, C21, C22, C23… Cn). This will be final set of clusters or input to next phase if needed.

Phase-III:

As mentioned above, this phase will be considered if further splitting of clusters is needed. If cluster quality needed is achieved from phase-2 itself then third phase execution is not necessary.

4.2 Similarity Measures:

We talk about single data item "p" used for clustering algorithm, it consists of vector of m measurements like p1, p2, … pm. The individual scalar components pi of pattern p is called *features* or *attributes*. A *distance measure* is a metric to quantify the similarity of patterns on the feature space. Many of clustering methods uses similarity measure as fundamental to measure the similarity between two patterns drawn from the same feature space. The distance measure should be chosen properly because there are variety of feature types and scales. As well we can define dissimilarity between two patterns using distance measure [1].

In first phase, similarity measure between two person items is defined by using features 'name', 'date of birth' and 'address'. The formula used (as per approach two given below) is as follows:

$\mathbf{d_N}$: distance measure between names
$\mathbf{d_{DOB}}$: distance measure between date of births
$\mathbf{d_{ADD}}$: distance measure between addresses

$$\textbf{Approximate Value} = \mathbf{dN} + \left[\frac{\mathbf{dDOB + dADD}}{\mathbf{2}}\right]$$

5. SYSTEM ARCHITECTURE

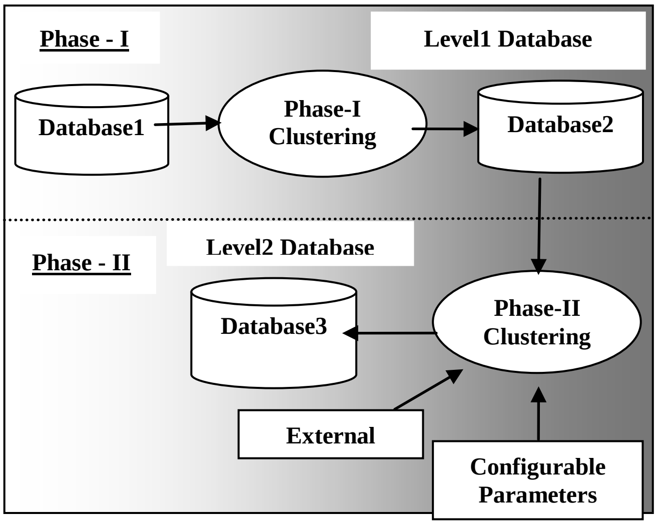

Figure4: System Architecture

Clustering process is divided into two phases I & II, as shown in Figure4. Phase-I clustering uses the proper data from database1 where data is already loaded into the database in table form. Phase-I uses internal fixed features for clustering so no other external data except database1 is used. The output of Phase-I clustering is set of clusters which getting stored into database2 in the form of tables and relations. Phase-II Clustering using configurable features, such as features need to be considered while clustering in phase-II and rules for clustering. We are saying rules means it includes data type of features and matching conditions and formula and threshold values.

6. EXPERIMENT AND RESULT ANALYSIS

Phase-I Analysis:

Consider, features f1: name, f2: dob, and f3: address (country, state, city).

To calculate the distance measure between two clusters C1 & C2 whose values are given below:

C_1: [NAME: First Name (Preeti) Surname (Jain), DOB: dd (09), mm (04), yyyy (1989), ADDRESS: Country (India), State (Maharashtra) and City (Solapur)]

C_2: [NAME: First Name (Dipti) Surname (Jain), DOB: dd (07), mm (04), yyyy (1989), ADDRESS: Country (India), State (Maharashtra) and City (Solapur)]

Basic Rules for Comparison of Feature values:

```
1. NAME
Convert all characters of string to lower case letter

If C1.Surname == C2.Surname AND C1.FirstName == C2.FirstName Then
    nameMatch = 1.0;
Else
    nameMatch = 0.0;
End If
In above example nameMatch = 0.0;

2. DATE OF BIRTH
If C1.yyyy == C2.yyyy AND C1.mm == C2.mm AMD C1.dd == C2.dd Then
    dobMatch = 1.0;
Else
    dobMatch = 0.0;
End If
In above example dobMatch = 0.0;

3. ADDRESS
Convert all characters of string to lower case letter
If C1.add == C2.add Then
    addMatch = 1.0;
Else
    addMatch = 0.0;
End If

In above example addMatch = 1.0;
```

Here, address itself contains three fields like country, state and city. There is need to first match all these three sub-fields and then we are saying address field of two items is matching.

Proximity Value Calculation:

Approach One:

Formula:

$$\textbf{Average} = \left(\frac{\textbf{NameMatch + DOBMatch + ADDMatch}}{\textbf{3}}\right)$$

Table 1. Test case results according to approach one

Case	nameMatch	dobMatch	addMatch	Avg
T1	0.0	0.0	0.0	0.0
T2	1.0	0.0	0.0	0.3333
T3	1.0	1.0	0.0	0.6666
T4	1.0	0.0	1.0	0.6666
T5	0.0	1.0	0.0	0.3333
T6	0.0	1.0	1.0	0.6666
T7	0.0	0.0	1.0	0.3333
T8	1.0	1.0	1.0	1.0

T1: No name match, no dob match, no city name match
T2: Name match, no dob match, no city name match
T3: Name match, dob match, no city name match
T4: Name match, no dob match, city name match
T5: No name match, dob match, no city name match

T6: No name match, dob match, city name match
T7: No name match, no dob match, city name match
T8: Name match, dob match, city name match

Here, if we use the threshold value 0.5 then in the case (T6) it gives incorrect result that is name is not matching and date of birth and birth place are matching. And any other threshold values also give many other incorrect results.

So we are not following this approach.

Approach Two:

In above approach we are calculating average value by adding three features match values and dividing by number of feature values. In this approach we will separate the "nameMatch" feature from "dobMatch" and "addMatch" features and so the formula formed is as follows

Formula:

Δ = [(DOBMatch + ADDMatch) / 2]

Average = NameMatch + Δ

Table 2. Test case results according to approach two

Case	nameMatch	dobMatch	addMatch	Δ	Avg
T1	0.0	0.0	0.0	0.0	0.0
T2	1.0	0.0	0.0	0.0	0.0
T3	1.0	1.0	0.0	0.0	1.0
T4	1.0	0.0	1.0	0.0	1.0
T5	0.0	1.0	0.0	0.0	0.0
T6	0.0	1.0	1.0	1.0	1.0
T7	0.0	0.0	1.0	0.0	0.0
T8	1.0	1.0	1.0	1.0	2.0

Here, we will be going with the threshold value 1.0 then all above cases gives the correct results.

In our algorithm we are using this approach in phase-I clustering.

Example:

Item1:

Name:	Preeti Jain
Date of Birth: (dd-mm-yyyy)	09-04-1989
Address :	<country>India</country> <state>Maharashtra</state> <city>Solapur</city>

Item2:

Name:	Preeti Jain
Date of Birth: (dd-mm-yyyy)	07-04-1989
Address:	<country>India</country> <state>Maharashtra</state> <city>Solapur</city>

As per approach-2, we are calculating the following values:

$$\begin{aligned} &\text{nameMatch} = 1.0 \\ &\left.\begin{aligned}&\text{dobMatch} = 0.0\\ &\text{addMatch} = 1.0\end{aligned}\right\} \ 0.0\ [(\text{DOB \& ADDRESS}] \end{aligned}$$

Final Avg = 1.0 which is not greater than threshold value 1.0 so both are not belong to same cluster.

Table 3: Snapshot of Sample data:

ID	Surname	First Name	DD	MM	YYYY	country	state	city
7	Jain	Anil	18	12	1989	India	MH	MUM
8	Jain	Sunil	6	12	1990	india	MH	MUM
9	Jain	Preeti	9	4	1991	india	MH	MUM
10	Jain	Dipti	7	4	1992	India	MH	MUM
11	Jain	Priyanka	26	9	1990	India	MH	SUR
12	Sharma	Amit	18	12	1990	India	MH	PUN
13	Sharma	Ajit	12	12	1991	India	MH	PUN
14	Sharma	Amol	12	12	1993	India	MH	PUN
15	Sharma	Priyanka	9	9	1990	India	MH	PUN
16	Sharma	Amita	18	12	1990	India	MH	PUN
17	Patil	Anil	18	12	1989	India	MH	SUR
18	Jain	Preeti	9	4	1990	India	MH	SUR
19	Jain	Dipti	7	4	1992	India	MH	MUM
20	Jain	Priyanka	26	9	1990	India	MH	SUR
21	Patil	Sachin	20	1	1989	India	Maharashtra	SUR
22	Patel	Scheta	22	2	1990	India	Maharashtra	SUR

Table 4: Snapshot of Level1 Database:

ID	Surname	First Name	DD	MM	YYYY	country	state	city
7	Jain	Anil	18	12	1989	india	MH	MUM
19	Jain	Dipti	7	4	1992	India	MH	MUM
10	Jain	Dipti	7	4	1992	India	MH	MUM
18	Jain	Preeti	9	4	1990	India	MH	SUR
9	Jain	Preeti	9	4	1991	india	MH	MUM
11	Jain	Priyanka	26	9	1990	India	MH	SUR
20	Jain	Priyanka	26	9	1990	India	MH	SUR
8	Jain	Sunil	6	12	1990	india	MH	MUM
22	Patel	Scheta	22	2	1990	India	MH	SUR
17	Patil	Anil	18	12	1989	India	MH	SUR
21	Patil	Sachin	20	1	1989	India	MH	SUR
13	Sharma	Ajit	12	12	1991	India	MH	PUN
12	Sharma	Amit	18	12	1990	India	MH	PUN
16	Sharma	Amita	18	12	1990	India	MH	PUN

In Table3 and Table 4, following notations are used:
MH means "Maharashtra"
MUM means "Mumbai"
SUR means "Solapur"
PUN means "Pune"

Table4 shows the output of the phase-I clustering. In this table, using color green and brown two clusters are represented. Green color represents cluster of Surname

"Jain" and First Name "Dipti" having same address and date of birth information but different ID that is "10" and "19". Brown color cluster consists of Surname "Jain" and First Name "Priyanka" having same address and date of birth. But the items represented by red color are a special case which is not handled properly. In these items all feature information are matching except First Name. In first Name, one item is holding "Amit" and other item holding "Amita". So by looking at the data it is observed that it is a problem of data mistyping. Such cases should be handled in different way which has been discussed in future work.

Phase-II Analysis:

Phase-II Clustering using following the same algorithm but the features used in this are external and configurable. The rules are also external to algorithm which includes data type of features and matching conditions and formula and threshold values.

Example of some additional features:

1] f4: email address (String)
2] f5: Father Name (String)
3] f6: Mother Name (String)
4] f7: School Name (String)
5] f8: School Address/city name (String)

7. CONCLUSION AND FUTURE WORK

This paper represents simple and straightforward approach of person clustering. It is using basic concept of CURE (Cluster using Representative) algorithm. Clustering of person information represented here is using simple and basic matching rules like exact matching of strings in name, address fields etc. Here, variation of strings is not considered for example it possible that person surname "Sharma" might be given as "Sharama" or city name "Solapur" is given as "Sholapur" in some data. In such cases, the algorithm will misguide and add this item to wrong cluster or it will create it as new cluster itself. In this system we have used very small set of dataset as test data.

The items given in Table4 using red color is best example of what we are discussing. In those items, all features are matching except first name that is "Amit" and "Amita". The case in is that the first characters are matching and only the last character does not match, meaning in AMITA, a is additional so we need to go ahead to check whether they belong same cluster with all old features as they are. This would be our future work of finding mismatches and recognising in a deeper way.

As a future work and to make result more accurate, there is need to consider different variation in data items and to collect external large set of datasets.

8. REFERENCES

[1] Jain A.K., Murty M.N., and Flynn P.J.: Data Clustering: A Review (ACM Computing Surveys, Vol. 31, No. 3, September 1999)

[2] Margaret H. Dunham Data Mining Introduction & Advanced Topics (Pearson Education, 2006)

[3] Guha Sudipto, Rastogi Rajeev, and Kyuseok Shim: CURE: An Efficient Clustering Algorithm for Large Databases

[4] Teevan Jaime, Morri Meredith Ringel, and Bush Steve: Discovering and Using Groups to Improve Personalized Search

[5] Culotta Aron, Bekkerman Ron, and McCallum Andrew: Extracting social networks and contact information from email and the Web (2004)

[6] Matsuo Yutaka, Mori Junichiro, Hamasaki Masahiro, Nishimura Takuichi, Takeda Hideaki, Hasida Koiti, and Ishizuka Mitsuru: An advanced social network extraction system from the Web

[7] Kautz H., Selman B., and Shah M.: The hidden Web. AI magazine, Vol. 18, No. 2, pp. 27–35, 1997

[8] Kurki Jussi: Finding People and Organizations on the Semantic Web

[9] Fisher Douglas H.: Knowledge Acquisition via Incremental Conceptual Clustering (Machine Learning 2: 139-172, 1987)

[10] Chan Stefani, Pon Raymond K., and C´ardenas Alfonso F.: Visualization and Clustering of Author Social Networks