ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology , Vol 8(S2), 212–216, January 2015 DOI:.10.17485/ijst/2015/v8iS2/60482

Data Mining: Building Social Network

Sayali Nishikant Chakradeo*, Riya Mary Abraham, Beeda Anusha Rani and R. Manjula

SCSE, Vellore Institute of Technology, Tamil Nadu, India; Sayali31889@gmail.com

Abstract

Data mining is one of the fields of computer science wherein we discover patterns through large databases using some computational process or some other methodology. Extracting such data from social media makes such patterns available which can be helpful for different businesses, markets and for other purposes. In this paper, we have extracted data from Wikipedia and blogs regarding ten different attributes of writers, politicians and athletes and find the similarity between them using some of the pre-existing methodologies and finally build a social network amongst them. We have also conducted some mathematical experiments on them so as to examine the efficiency of our analysis.

Keywords: Data Mining, Social Media, Wikipedia

1. Introduction

In our day to day life we depend mainly on the social media for information gathering and sharing. Internet has become the main source of information exchange. Social media can be considered as the internet oriented applications that allow the generation and sharing of user related content. Social media includes the traditional media like newspaper and non-traditional media. Main online social networking sites which influenced the users all over the world include facebook, twitter etc. Social media provides us easy way of communication than the traditional systems. The popularity of the social media resulted in origin of many social networks. Data generated on social media sites are different from traditional ones. They are vast, unstructured and dynamic. By studying the social media data we have understanding about the personal entities behavioral aspects, their mind set and views etc. For this we have to build a social network out of the data collected from different media sources.

Social Network is built on set of actors and based on their relations. Social Network Analysis emerged years before but is one of the latest and hot research topic now. This analysis would help in the Personal information retrieval process. Various social network mining methods are available. Referral Web, Flink are some of the systems developed social network mining purposes.

In this paper we are going to mine the relations between the personal entities, to build and analyze the social network of these personal entities. In this analysis initially the attribute information of person entities have been collected from social media. We have taken three categories of entities for our study namely, Politicians, Writers and Athletes. Ten attributes have been considered for representing each entity. Next step, person entity relations have been extracted based on the similarity in person entity. In order to extract the information the ten selected attributes are compared. We are doing a performance evaluation on an existing work by adding more number of attributes to represent a person entity and to determine the variations in the result obtained by comparing it with the existing one. Social network is build based on the similarity of person entities. Experiments have been conducted on the analysis and it showed the final social network we build portrays the social relations among the person entities we selected in each category.

In this paper Part II gives the literature survey done on this topic; Part III describes the methodology followed in our research. Results and Conclusions have been presented in the last section.

2. Literature Review

[2]Social network analysis is method of creating the model for communication patterns amongst the entities

*Author for correspondence

and to highlight the importance of each entity. With increasing show of different web related platforms like blogs, wikis, social networking sites, SNA has gained lots of grandness. [3]Building user models plays an important role in adaptive intelligent systems. User requirements, their priorities in various conditions are needed to take into consideration while designing such build. [4]The growing number of individuals is recently writing their own opinions or information freely at the network space on the web such as the blog or Online Cafe and these network spaces are developed toward a new service called social network[5]. The social media has obtained pervasive importance these days. Social media data are huge, noisy and amorphous in nature and change constantly and therefore give rise to new challenges. There are plenteous opportunities in social network mining to discover useful knowledge. Social network is considered as a type of "Big Data" and finds its use in many domains. There is a hopeful future for social network mining to emerge and enhance our research and development in the same. [5]For daily events and news happening in day-today life, social network is considered to be the most crucial source which gives us people's feedback and opinions about the ongoing events. Authors of "Social Network Mining Based on Wikipedia" [1]have proposed a methodology for building a social network for the data regarding people collected from Wikipedia. They have made use of 6 attributes of these people to analyze similarities and differences between these people. They have proved to be capable of building and analyzing social network quite efficiently. For similarity calculations, use of Systematic Similarity Measurement (SSM)[7], Levenshtein distance calculation[8] are the methods being used here. Net draw toolkit[9] helps in building the visual representation of the social network model.

3. Methodology

In this study we are mining the relation between the person entities in the social media and build the social network for them. A social network can considered as a graph which consists of nodes and edges, where nodes represent entities, and edges represent the entity relations. It is a weighted graph with weights represents the strength of the relation. Our Social network building procedure has four steps to be followed. Starting from Data collection, Data Classification, Similarity computing and finally Social network building. The steps can be represented diagrammatically as follows:

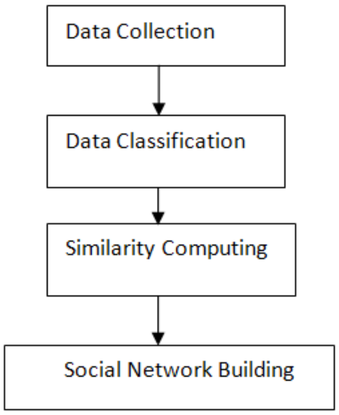

Figure 1. Block diagram for the social network procedure.

The steps involved are explained in detail below:

3.1 Data Collection

The Experimental data have been collected from different online social media. Three categories of person entities have been taken for study. Politicians, Writers and Athletes are the different sets chosen for the social network building.

3.2 Data Pre–Processing

The data needed for study have been collected from various resources and data pre-processing has been done.

3.3 Data Classification

In the existing work four attributes were used to represent the person entity. We have added 6 attributes in addition to the existing 4 attributes, thus forming a total of ten attributes per entity. Ten attributes namely Personal beliefs and Party, Birth Year, Birth Place, Personal Experience, Education, Gender, Language spoken, Hobby, Social Networking, Religion have been selected to compute the person entity similarity.

3.4 Similarity Computing

We have followed the Systematic Similarity Measurement method. It can be explained as follows:

Let (Oa, Ob) be a pair of system

$$Oa = \{A1, A2..Am\}\ m = |Oa|$$

$$Ob = \{B1, B2..Bn\}\ n = |Ob|$$

Where Ai is an attribute in Oa and Bj is an attribute in Ob. Attribute Similarity of <A1, B1>, <A2, B2> and <Ap, Bp> is

Sim (Aj, Bj) which represented as ‘s’. The weight of attribute Ai is represented as xi and of Bj is yj. So the formula for Systematic Similarity Measurement method is

$$S(Oa, Ob) = \frac{\sum_{i=1}^{n} sixi^2}{\sqrt{\sum_{i=1}^{n} xi^2 \sqrt{\sum_{i=1}^{n} uixi^2}}}$$

To compute the similarity of the systems the above mentioned formula is used. To calculate the attribute similarity between two person entities the following methods have been used. As we have taken 10 attributes to represent a person entity, different methods have been used to compute the similarities of each attribute.

3.4.1 Personal Beliefs and Party Attribute

String comparison method, the similarity value will be 1 if the strings are the same, otherwise value would be 0.

3.4.2 Birth Year

Let (Ya ,Yb) be the birth year of (Oa, Ob) respectively, then Sim(Ya, Yb) is

$$Sim(Ya, Yb) = 1/(log2\wedge(|Ya - Yb| + 1) + C)$$

Here if the birth years are close, then they have more chance of knowing each other. The value of C is taken as 1.

3.4.3 Birth Place

In order to compute the birth place similarity we have used the Hierarchical method. In this method while computing the similarity of (Oa, Ob) first check the similarity of their province names of birth place. If the province names appeared to be same then only further comparison is done. The more similarity between the person entities as deeper we compare.

3.4.4 Personal Experience

It is the most abundant attribute. It can be a list of strings. These include many strings which are same.

Two person entities will be more similar if it shares more common things. We have used Levenshtein distance to compute the similarity. Let T[1 – m] and S[1 – n] be the strings and d be the edit distance between T and S where 0<=d<=min (m,n). The similarity can be computed using the equation

$$Sim(T, S) = 1 - d/(max(m, n))$$

3.4.5 Gender, Religion, Social Networking

For these 3 attribute also the similarity can be calculated using the String comparison method. The gender attribute can have value either Male or Female. For Social networking attribute the value can be either yes or no. If the person uses social networking sites then the value would be yes else no. If the strings have same value then result will be 1 else 0. For Religion attribute also the same concept can be followed.

3.4.6 Hobby, Language Spoken

For the attributes Hobby and Language Spoken by the person entity, the modified String Comparison method is used. The Hobby/Language by a person is compared with every Hobby/Language by other entity.

3.5 Social Network Building

Consider each person entity as a node; by computing each pair of nodes (person entity) we can build a similarity array. A threshold value k is predefined, select some pair of nodes with person entity similarity value than k from similarity array and connect them to form edges. They are weighted edges with person entity similarity values. Thus a social network is build.

4. Discussion

To compute the similarities we have implemented the java program which will take data from the excel worksheet and will give the similarity output into an excel worksheet. For each similarity computing method different java program have been used. The results obtained have been used to build the social network. Social network build has been depicted using the netdraw software as well.

Below diagram shows the output of social network that we built in netDraw.

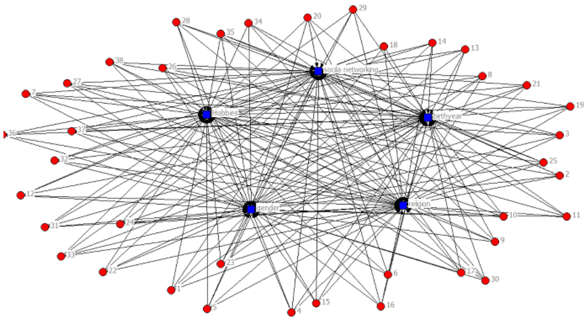

Figure 2. Network built in netDraw.

Below is our analysis for all the 10 attributes of writers, politicians and players.

Here we use below formulae.

Pn = total no of person entities
Rn = total no of person entities we could extract
Cn = total no of entities we could correctly extract
Wn = total no of entities we extracted wrongly

Also in the tables below, P = Cn/Rn and R = Cn/Pn

$$F = 2^*P^*R/(P + R)$$

Average accuracy for each parameter:

Table 1. Accuracy for Writers

Parameter	Pn	Rn	Cn	P	R	F
Gender	50	50	19	38	100	55.07246
Religion	50	50	42	84	100	91.30435
Birthyear	50	50	50	100	100	100
Social	50	50	33	66	100	79.51807
Hobbies	50	50	45	90	100	94.73684
Languauge	50	50	45	90	100	94.73684
Personal Expeience	50	45	45	100	90	94.73684
Personal Beliefs	50	45	45	100	90	94.73684
Birth place	50	50	38	76	100	86.36364
Education	50	50	43	86	100	92.47312

Table 2. Accuracy for Politicians

Parameter	Pn	Rn	Cn	P	R	F
Gender	50	50	38	76	100	86.36364
Religion	50	50	38	76	100	86.36364
Birthyear	50	50	50	100	100	100
Social	50	50	35	70	100	82.35294
Hobbies	50	50	48	96	100	97.95918
Languauge	50	50	42	84	100	91.30435
Personal Expeience	50	45	45	100	90	94.73684
Personal Beliefs	50	45	45	100	90	94.73684
Birth place	50	50	39	78	100	87.64045
Education	50	50	42	84	100	91.30435

Table 3. Accuracy for Players

Parameter	Pn	Rn	Cn	P	R	F
Gender	50	50	50	100	100	100
Religion	50	50	16	32	100	48.48485
Birthyear	50	50	46	92	100	95.83333
Social	50	50	35	70	100	82.35294
Hobbies	50	50	50	100	100	100
Languauge	50	50	45	90	100	94.73684
Personal Expeience	50	45	45	100	90	94.73684
Personal Beliefs	50	45	45	100	90	94.73684
Birth place	50	50	41	82	100	90.10989
Education	50	50	33	66	100	79.51807

Table 4. Average accuracy

Parameter	F Avg
Gender	80.4787
Religion	75.3242
Birthyear	98.6111
Social	81.4879
Hobbies	97.56534
Languauge	93.5926
Personal Expeience	94.7368
Personal Beliefs	94.7368
Birth place	88.0379
Education	87.76518

5. Conclusion

From the above table we can see that, the accuracy is more in case of the parameters Birth Year, Language, hobbies, personal experience and personal beliefs. FangFang Yang and Xu Sheng Li Zhikai Xu[1] had considered 4 parameters for analysing the similarity between people they took into consideration. Here we have made an analysis of similarities among 10 parameters of people hence we can judge which parameters play an important role while building the social network i.e. Birth Year, Language, hobbies, personal experience and personal beliefs.

6. References

1. Yang F, Xu Z, Li S, Xu Z. Social Network Mining Based On Wikipedia.
2. Akhtar N, Javed H, Sengar G. Analysis of Facebook Social Network.
3. Ortigosa A, Quiroga JI, Carro RM. Inferring User Personality in Social Networks: A Case Study in Facebook.
4. Cho KS, Yoon JY, Kim IJ, Lim JY, Kim SK, Kim U-M. Mining Information of Anonymous User on a Social Network Service.
5. Gundecha P, Liu H. Mining Social Media: A Brief Introduction.
6. Rafea A, Mostafa NA. Topic Extraction in Social Media.
7. Guan Y, Wang X, Wang Q. A New Measurement of Systematic Similarity. IEEE Transactions on Systems, Man and Cybernetics–part a: Systems and Humans. 2008; 38(4):743–58.
8. Ristad ES, Yianilos PN. Learning String-edit Distance. IEEE PAMI. 1998; 20(5):522–32.
9. Huisman M, van Duijn MAJ. Software for Social Network Analysis. 2003.