INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 2)
Available online at: www.ijariit.com

Social champion identification

Abhi Savaliya
abhisavaliya01@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Prince Kumar Maurya
princesenpai@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Aditi Sakhare
aditisakhare21@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Rohit Jadhav
jadhavrohit9634@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Aditi Chhabria
vimlajethani@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

ABSTRACT

NGO are non-profit making agencies that are constituted with a vision by a group of like-minded people, committed for the uplift of the poor, marginalized, unprivileged, underprivileged, impoverished, downtrodden and the needy and they are closer and accessible to the target groups. However, in spite of its achievements in various fields, NGOs are facing different problems which differ from organization to organization, region to region. In this context, an attempt is made in this report to solve some of the common problems faced by the NGOs and to give some remedies to overcome these problems. NGOS are always looking for individuals who will promote and spread awareness for their work and cause. We are looking to create a solution that will help the NGOs use basic information about existing and prospective donors like their name, email id, Facebook and LinkedIn profiles, and mobile numbers to look for any information available in the public domain to identify the ones most likely to champion their cause as social contributors.

Keywords: *Social activist, NGO, Data Mining and Analysis, K-Means, Big Data.*

1. INTRODUCTION

NGOs are the prime medium in current era which can actually promote and execute a social cause. They are the true helpers of the society and the needy. However due to lack of proper technology and platform they lack the workforce who can actually contribute towards their social cause. Social Champion Identification provides a platform for the NGOs to identify and appreciate the hidden workforce that can help achieve their cause. NGOs are powerful tools in poverty alleviation and development. Many NGOs and organizations for social change have mushroomed in India in the previous decade but very few have expanded on the basis of scale and the impact on the community. Most of the NGOs in India are suffering from paucity of funds. They find accessing donors as challenging as dealing with their funding conditions. Also, the basic characteristic of NGO is volunteerism. In early days, youth are making their career in volunteerism but that enthusiasm seems to have faded these days. The extent of volunteerism is declining day by day and turning it into professionalization. Even the young graduates from social work are interested in making their career in professionalism. This leads to lack of efficient volunteers in NGOs. Using conventional methods to solve these problems will be inefficient as well as time consuming. Hence, we aim to develop a system using Social Media Analysis which will help the NGOs to raise funds and gain volunteers by identifying the social champions who will be interested to contribute to and promote their cause.

© 2018, www.IJARIIT.com All Rights Reserved

2. LITERATURE SURVEY

In place system has certain problems which makes it inefficient. These problems are described by Dr. K. Prabhakar Post-Doctoral Fellow, Department of Anthropology, S.V.University, Tirupati in his paper Voluntary Organisation And its Challenges in India. There's a study [2] done on drug usage using data from social media platforms. It presents approach for mining and analysis of data from social media which is based on using Map Reduce model for processing and analysis of big amount. They applied this system for creating characteristics of users who write about drugs and to estimate factors that can be used as part of model for prediction drug usage level in real world. This paper basically proposes to use data from social media as a source of information for analysis and modelling of drug usage among population of certain territory. The context in which the data is to be extracted is also important. People uses various emoticon to express themselves on the social media platform. There's a study which show how to analysis the sentiments [3] of the tweets in posted in real time. This will help us in identifying the key interest of the people and extract the best possible candidate.

3. PROPOSED WORK

The proposed system will help NGO's find volunteers using social data available on the social media platform. This system will make use of the social media profiles of the existing social worker or donors and using their profiles we we'll search for other social champions.

A. Seed Data

Our system requires some basic details about the existing donor, their social media profile ids, or email id is to be given to our system as initial input.

B. Data Acquisition

This consist of the Crawler, Data Fetching, API/Web Scrapper, and Social Media. Using the seed data the crawler will crawl through the profiles of the user, his tweets, his/her friend list (if access is allowed), tweets and comment on the friend's profile. If the social media has provided us with the data end point through it API we'll make use of it for scrapping and crawling the profiles on social media.

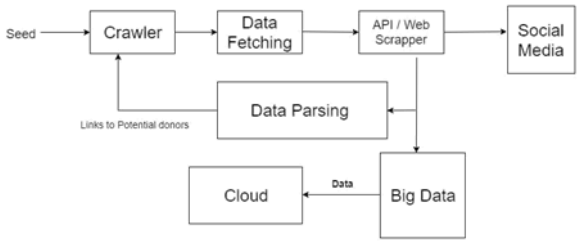


Figure 1. Basic flow of the System

C. Data Processing

As the data obtained from Social Networking Sites is huge and sparse, hence Big Data and Cloud Computing is used to store and process the data efficiently. The data is filtered and aggregated using the Map Reduce model for distributed computations which is implemented using Hadoop framework on a cloud based platform. In the last step, small data sets are obtained which are used in the Data Analysing process.

D. Data Analysis

After the raw data is successfully broken down into chunks of data, the data is analysed using K-Means algorithm. K-Means clustering is a fast, robust, and simple algorithm that gives reliable results when data sets are distinct or well separated from each other in a linear fashion. To identify the prospective donors, volunteers and the social champion, the features considered are relevance, reach and resonance. The features are described below:

- Relevance: The creation of content that is relevant to the NGO, or relevant to a topic that's important to the NGO.
- Reach: The ability to reach an audience that is valuable to the NGO.
- Resonance: The proliferation or engagement with relevant content by an audience that is valuable to the NGO.

The people will be divided into the following clusters:

- Champions
- Mediocre
- Dismal

© 2018, www.IJARIIT.com All Rights Reserved

4. METHODOLOGY

A. Clustering

Cluster analysis is the process of grouping objects into subsets that have meaning in the context of a particular problem. The objects are thereby organized into an efficient representation that characterizes the population being sampled. Cluster analysis is not on specific algorithm, it can be achieved by different algorithms. Depending on the type of problem we can choose appropriate algorithm.

1) Objective:

- Discover patterns in high dimensional data. Grouping of similar data.
- Easy interpretation of data.

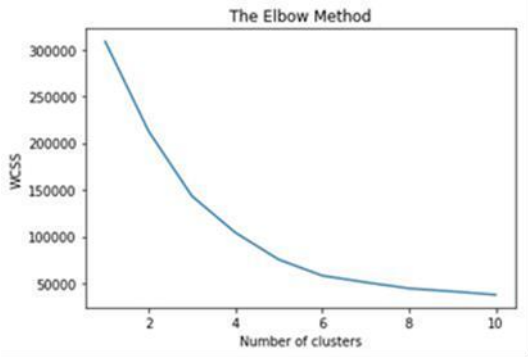
We have used clustering to identify pattern among the donors and categories them based on their contribution.

B. K-Means Clustering with Elbow Method

K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). This clustering algorithm separates data into the best suited group based on the information the algorithm already has. In general, the k-means method will produce exactly k different clusters of greatest possible distinction. Data is separated in k different clusters, which are usually chosen to be far enough apart from each other spatially, in Euclidean Distance, to be able to produce effective data mining results. Each cluster has a centre, called the centroid, and a data point is clustered into a certain cluster based on how close the features are to the centroid. K-means algorithm iteratively minimizes the distances between every data point and its centroid in order to find the most optimal solution for all the data points.

1) Elbow Method:

The Elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. This method looks at the percentage of variance explained as a function of the number of clusters. Number of clusters should be chosen in such a way that adding of another clustering point should not give a better modelling.



2) K-Means Algorithm:

Select k points as initial centroids. Repeat until convergence. Form k cluster by assigning each point to its closed centroid.

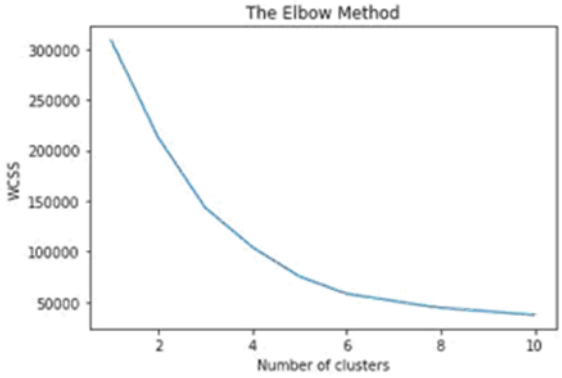


Figure 2: Elbow method for k-Means++ Clusters

- Recalculate the centroid of each cluster.

Computationally, the program will start with k random clusters, and then move objects between those clusters with the goal to

- Minimize variability within clusters.
- Maximize variability between clusters.

In other words, the similarity rules will apply maximally to the members of one cluster and minimally to members belonging to the rest of the clusters. In k-means clustering, the program tries to move objects in and out of groups (clusters) to get the most significant results.

© 2018, www.IJARIIT.com All Rights Reserved

3) Efficiency

This algorithm is relatively efficient: O(tknd) where n is number objects, k is number of clusters, d is number of dimension of each object, and t is number of iterations. Normally k; t; d << n:

4) Solved Example

Table 1: Dataset

No.	Relevance	Reach	Resonance
1	28.94	30.81	44.22
2	31.56	53.14	74.29
3	48.81	42.91	48.42
4	34.74	60.62	23.41
5	16.96	59.84	32.44
6	37.05	54.72	32.64
7	27.88	59.44	52.74
8	40.28	42.91	39.00
9	29.94	48.03	74.29

Initial Points assumed for cluster K1, K2, K3 are:

Table 2: Initial Points

K.	Relevance	Reach	Resonance
K1	16.96	59.84	32.44
K2	31.56	53.14	74.29
K3	48.81	42.91	48.42

Table 3: First Iteration

K.	Relevance	Reach	Resonance	Points
K1	26.88	50.42	20.02	0, 3, 4
K2	29.79	53.53	67.10	1, 6, 8
K3	42.04	46.84	40.02	2, 5, 7

Table 4: Second Iteration

K.	Relevance	Reach	Resonance	Points
K1	39.38	61.15	38.64	0, 3, 5, 8
K2	44.61	83.05	97.06	1, 6
K3	49.36	64.16	53.23	2,4,7

Table 5: Third Iteration

K.	Relevance	Reach	Resonance	Points
K1	46.03	59.40	49.55	0, 2, 5, 8
K2	53.61	101.55	86.60	3,6
K3	46.05	73.35	66.34	1,4,7

Table 6: Forth Iteration

K.	Relevance	Reach	Resonance	Points
K1	47.69	58.96	52.28	0,2,5,8
K2	58.11	110.80	81.37	3,6
K3	44.95	76.41	70.69	1,4,7

© 2018, www.IJARIIT.com All Rights Reserved

After Fourth iteration we can observe that there is no change in the clusters that are formed.

5) Visualization: Consider the fig 2 initially, there are 4 centroids placed at farthest distance from each other randomly.

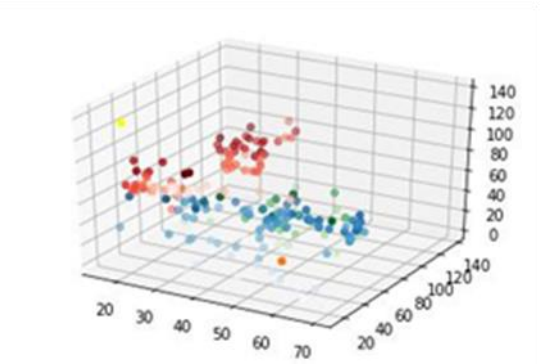


Figure 3: k-Means Clusters (Random Initialization)

C. K-Means++

The initialization problem for the k-means algorithm is an important practical one, and has been discussed extensively. It is desirable to augment the standard k-means clustering procedure with a robust initialization mechanism that guarantees convergence to the optimal solution. K-Means++ is an algorithm for choosing the initial values for the k-means clustering algorithm. This algorithm comes with a theoretical guarantee to find a solution that is O(log k) competitive to the optimal k-means solution.

1) K-Means++ Algorithm:

The first cluster centre is chosen uniformly at random from the data points that are being clustered, after which each subsequent cluster centre is chosen from the remaining data points with probability proportional to its squared distance from the point's closest existing cluster centre. Steps:

- Choose one centre uniformly at random from among the data points.
- For each data point x, compute D(x), the distance between x and the nearest centre that has already been chosen.
- Choose one new data point at random as a new centre, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)^{2.}$
- Repeat Steps 2 and 3 until k centres have been chosen.
- Now that the initial centres have been chosen, proceed using standard k-means clustering.

2) Solved Example

Table 7: Dataset

No.	Relevance	Reach	Resonance
1	28.94	30.81	44.22
2	31.56	53.14	74.29
3	48.81	42.91	48.42
4	34.74	60.62	23.41
5	16.96	59.84	32.44
6	37.05	54.72	32.64
7	27.88	59.44	52.74
8	40.28	42.91	39.00
9	29.94	48.03	74.29

Table 8: Initial Point Calculated

K.	Relevance	Reach	Resonance
1	50.58	53.40	30.81
2	30.75	34.28	28.94
3	74.29	38.1	4.22

© 2018, www.IJARIIT.com All Rights Reserved

Table 9: First Iteration

K.	Relevance	Reach	Resonance	Points
1	36.008	54.166	46.3	1,2,3,5,6
2	29.03	45.3975	37.4875	0,4,7,8
3	74.29	38.1	4.22	none

Table 10: Second Iteration

K.	Relevance	Reach	Resonance	Points
1	467.56	107.30	120.59	1
2	53.39	59.55	45.06375	0,2
3	31.14	54.25	42.41	3,4,5,6,7,8

Table 11: Third Iteration

K.	Relevance	Reach	Resonance	Points
1	65.09	113.68	98.37	3,6
2	47.39	71.81	63.59	1,4,7
3	43.97	57.68	50.49	0,2,5,8

Table 12: Forth Iteration

K.	Relevance	Reach	Resonance	Points
1	63.85	116.87	87.26	3,6
2	45.39	75.90	69.77	1,4,7
3	47.17	58.53	52.51	0,2,5,8

After third iteration we can observe that there is no change in the clusters that are formed.

3) Visualization

Consider the fig 2 initially, there are 4 centroids placed at farthest distance from each other randomly.

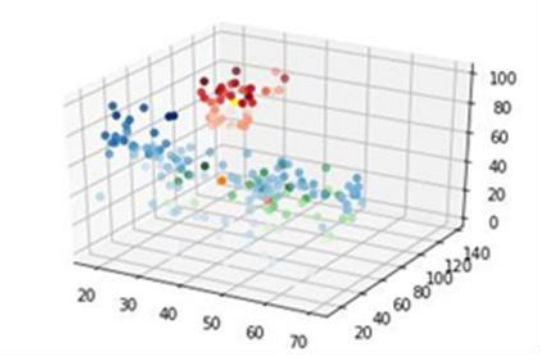


Figure 4: k-Means++ Clusters

D. Comparison

- We can notice the difference in the clusters in fig 4 and fig 5 For K-Means the clusters do not have clear boundary which is can be noticed in K-Means++.
- K-Means++: In order to find the initial points we use the probabilistic approach using which we can find all three centroid. Where in K-Means the initial points are random and as we change the initial points the cluster also change. For us it is important to have constant clusters with minimal change shifting of data points over different clusters. So K-Means++ shows us better results.

© 2018, www.IJARIIT.com All Rights Reserved

5. CONCLUSIONS

Social media contains a lot of personal information and can be used as an additional data source for analysis of social processes in real world. Especially processes that are hardly observable in real world like interest and attitude of society towards social causes. Despite the fact that social media cannot replace traditional sources of information provided by government, they can complement this data. Another advantage of social media is that they provide macro and micro characteristics of users. Macro characteristics are general interest of whole population of users to some topic and micro characteristics are specific to the user. And again social media reveal knowledge about micro level of social processes while other sources of information are more about macro parameters. We proposed to use data from social media as an additional source of information for analysing and finding information about individuals who are interested in promoting social causes and would help the NGOs by donating in kind and in cash or by volunteering for a particular event. Our system will help the NGOs by saving their time and efforts by generating lists of potential donors and volunteers and thus providing financial and organizational stability to the NGO. Hence, the NGOs can focus on helping the disadvantaged and oppressed people.

6. FUTURE WORK

In future, we can add a feature which will create a transparency between the actual donor and the receiver, so that the donors can directly contact the receiver to donate their organs, blood or provide financial support. Also we can create an automated email system which will contact the probable social contributors in the nearby areas to raise or promote a cause which will save a lot of time.

7. REFERENCES

[1] Social networks mining for analysis and modelling drugs usage Andrei Yakushev and Sergey Mityagin ITMO University, Saint-Petersburg, Russia.
[2] Sentiment Analysis on Twitter by Akashi Kumar and Teeja Mary Sebastian Department of Computer Engineering, Delhi Technological University Delhi, India.
[3] Voluntary Organisation and Its Challenges in India Dr.K.Prabhakar Post-Doctoral Fellow, Department of Anthropology, S.V. University, and Tirupati.

© 2018, www.IJARIIT.com All Rights Reserved