International Journal of Soft Computing 11 (6): 418-426, 2016
ISSN: 1816-9503

Visualization Model for Viral Marketing

T. Asha, Ravinder Singh, S.K. Chandan, V. Sachin and S. Surya
Department of Computer Science and Eng.,
Bangalore Institute of Technology, Bangalore, India

Abstract: With technology advancing by leaps and bounds over the past couple of decades, communication has now become much more convenient. Social media is an effective means of acquiring, organizing and articulating information. Given the large population of people on social media, an idea or information can easily be spread among the users. Social network analysis deals with the analyzing network of structure and with the propagation of information. This opens up opportunity for advertisers who can advertise using viral marketing. Our system uses a novel method of spreading the awareness of a product or an idea in a social network. The system adopts Principal Component Analysis (PCA) technique of determining the threshold value of a user which will help us in obtaining an accurate seed set from an influence maximization algorithm which is a distinct advantage of our system. In addition, the system also provides statistical data regarding the spread of influence useful to the user in decision making process.

Key words: Social network, social network analysis, viral marketing, influence maximization, visualization

INTRODUCTION

The internet has revolutionized communication to such an extent that the restriction of distance is almost insignificant. It has led the rise of social networks. Social media plays a vital role in spreading influence, making it easier to connect with people across boundaries.

We define social networks as web-based mediums that allow individuals to represent themselves using profiles that provide sufficient information, connect to other users, view and traverse such connections and those made by others within the network form communities and establish rapprochement between such communities. What makes social network sites such as facebook and twitter, unique is that they aid users in articulating and in making their social networks visible. In addition to a wide variety of technical features that the social networks have implemented, user viewable profiles that display an articulated list of friends who are also users of the system form a vital aspect of social networks. The public display of connections is a crucial component of social networking sites.

Social network analysis: In addition to improving communication, social networks hold a great amount of useful information which can be used in the study of social patterns. Social Network Analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs and other connected information/knowledge entities. If a social network is represented as a graph, the nodes would be the people and groups, while the edges would show the relationships or flows between the nodes.

One of the most vital tasks of SNA is to interpret the behavior of the entities in the network, in part or whole, by examining the relationships between the entities. Generally, many aspects of graph theory are used in social network analysis.

Graphical representation of social network: Social networks are usually modeled as digraphs. Digraphs consist of atomic elements called nodes or vertices where a link between them depicts a relationship which also shows the spread as well as clustering. Absence of a link indicates the absence of the relationship. The digraph is shown in Fig. 1.

Viral marketing: One of the many uses of analyzing social networks is to obtain data for the purpose of advertisement. One of the techniques which have gained great interest in the recent years is 'viral marketing'. Viral marketing utilizes social networks to proliferate the publicity of a product or brand, generating more sales. It uses videos, text messages and word of mouth technique. Viral marketing works by targeting a small group of users called seeds and utilize their links to spread the awareness of a productor brand to his neighbors in the social network. The aim of viral marketing is to maximize this diffusion process in order to obtain maximum awareness.

Corresponding Author: T. Asha, Department of Computer Science and Enginring, Bangalore Institute of Technology, Bangalore, India

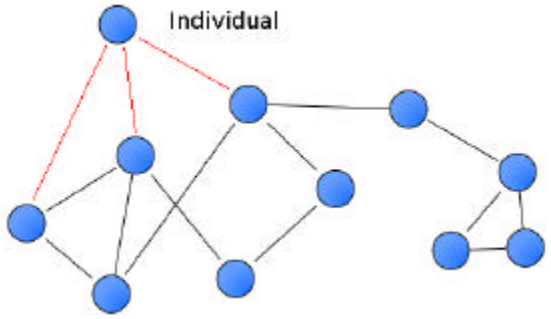

Fig. 1: Example of a digraph representing a small social network

Literature review: In the last few year, viral marketing has proven to be one of the most intriguing techniques and is still being researched nowadays. Due to the impressive results campaigns managed to obtain there has been more and more interest in this method. While there is research about viral marketing and a number of studies have tried defining and coming up with specific strategies to ensuring success there are still doubts when it comes to this technique and some contradictory results.

Influence maximization: Discovering influential nodes in a social network is a key aspect of viral marketing. In this respect, much of the existing work is based on the identification of nodes based on a graphical approach. For example, one of the existing work (Canali *et al*., 2010) illustrates a social network as consisting of users represented as nodes and links represent as lines. The social network is modeled as an adjacency matrix where 1 represents a link and 0 an absence of a link. It states 3 centralities that can be used to 'rank' individual node locations in the network.

Another technique in finding influential users (Chen *et al*., 2009), focuses on finding a way to discover influential individuals who can be classified into three types: one having the propensity like salespeople, opinion leaders and connectors.

Spread of influence: Given a seed set this method depicts how the information can be propagated through a network. Kempe *et al.* (2003), influence is propagated in the network according to a stochastic cascade model. Three cascade models, namely the independent cascade model, the weight cascade model and the linear threshold model are considered.

Kim and Han (2009), laid the foundation for obtaining the spread of influence by introducing two approximate solutions Linear threshold model and independent cascade model. Where linear threshold model concentrated on node weights and independent cascade model concentrated on link weights. Chen *et al.* (2009) and Kempe *et al.* (2003) uses degree discount heuristics with in uence spreads that are signi cantly better than the classic degree and centrality-based heuristics and are close to the influence spread of the greedy algorithm.

Visualization: In social network visualization the major issue is concerning the design of the layout depicting the social network with its users and links. Existing social network visualization methods usually use some topological structure such as tree for this purpose. One of the methods (Lam and Wu, 2009) proposes that a social network can be represented as a graph with individuals and organizations as nodes and relationships and contacts as edges. We can also create social networks where relations are nodes. From the social network that has been constructed, we can use a number of different metrics to find the most important nodes according to different criteria, compute similarities between individuals or find groups within the network.

Another system, vizster is a visualization system for exploring such online social networks. Vizster builds on ethnographic research of online social networking services. It also provides a system that articulates the social network in a customizable manner. The research on viral marketing thus far is mostly based on degree heuristics (Claudia *et al*., 2010; Chen *et al*., 2009; Kim and Han, 2009; Kempe *et al*., 2003; Leskovec *et al*., 2007), i.e., based on finding seeds with the most number of links. The relevance of the seeds to the campaign is not taken into consideration. Once the seeds are selected the spread of influence uses the entire user base and attempts to spread it at multiple levels. This approach has a longer processing time and is not optimal in spreading influence (Heer and Boyd, 2005). Finally, the existing systems use graphs for processing as well as visually depicting the social network (Lam and Wu, 2009). Although, graphs are useful processing tools they are congested and seem complicated to untrained users. Our approach takes into consideration, modern techniques [] and collaborates them to obtain a viral marketing application.

MATERIALS AND METHODS

Design and implementation

Architecture: The architecture of our application is depicted in Fig. 2. The architecture consists of two parts:

- One part of the application collects information from the user and performs data cleaning and stores it in a database

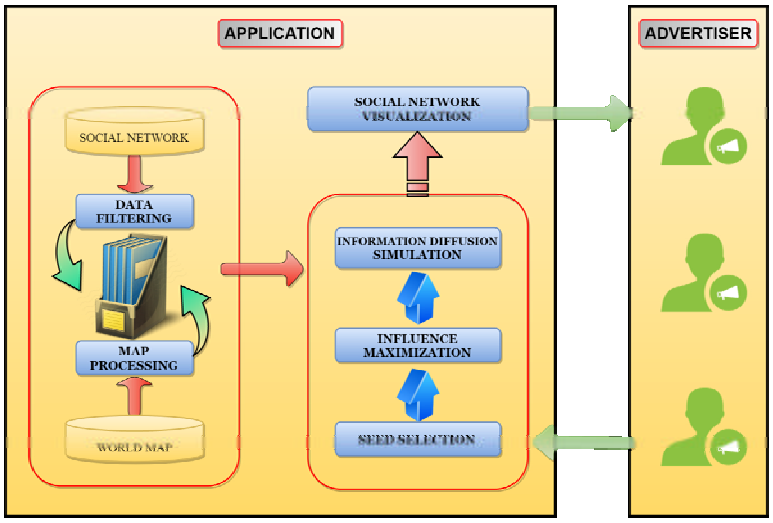

Fig. 2: Architecture

- The second part takes in the user specification from an advertiser and performs the job ofidentifying the influential leaders, diffusion simulation and visualization

Data retrieval and filtering: This module comprises the initial part of our application that is involved in the retrieval of user data and performing data cleaning operation on the data obtained. User data can be obtained from social networking websites. For our application we consider the data obtained from Facebook. This is in accordance to Dinh *et al*. (2014). The user data from Facebook can be accessed using the RESTFUL services offered by the website known as Graph API.

The access to data involves setting up a developer account with Facebook and obtaining an authentication token from facebook. Using the token we can allow users of our application to log in to our application using Facebook credentials. This allows our application to access their profile information that helps us build our dataset. This profile information include filed like, email-id, friend's list, pages or communities the user is associated to location information, etc.

Once the data is obtained we need to clean the data in order to make it more reliable and suitable for our application. The two techniques that are employed for data cleaning are:

Data filtering: It is the process of refining the input data which involves excluding those fields from the input that are not necessary for the application. Here, we employ a parser for filtering. The parser check the required fields for consistency and then includes it into the dataset. If the filed does not match the criteria then a default value is added into the dataset for that field. For example while inducting the email field into the dataset, the parser check to see if the email is of the proper format (e.g., abc@xyz.com). If the email conforms to the format then it is added to the dataset otherwise a null is added to the email field.

Data transformation: It is the task of converting data from one format to another format as required by the destination system. The page information that is obtained from the user data only contains the name of the user page and does not contain any description of the page. In this phase the page name is used to query information about the page from Facebook and retrieve the relevant information about the page.

Map processing: This is another form of data transformation where the location information is transformed into coordinate values for plotting it onto a geographic map for visualization. The data obtained from facebook user data contained only the name of the location and not the geographic coordinates of the map. Hence, we need to transform this location into coordinate values. For this conversion we make us of the RESTFUL services offered by google maps. We query the web service for the coordinate values by passing the location name in the request. The web service sends the data in JSON format.

User selection: User selection is a filtering process that gives us users that are relevant to the advertiser's specification. The advertiser specification relays the constraints for selecting users from the social network and include attributes like location, communities that are relevant to his product age and designation restrictions.

We employ Principle Component Analysis (PCA) (Mislove *et al*., 2007) as it is a baseline technique for social network data.

User information that pertains to location, age and designation can be filtered using a string match function and for information pertaining to communities and links we use PCA.

PCA aggregates the user information and plots them into equal number of dimension using a Euclidian space. After the plot is obtained it normalizes the data points generated into values between {0-1}. The normalized values represent the influence factor of each user in the social network. The influence factor gives us the threshold value for each user using the equation below:

$$\text{Threshold Value} = 1\text{-Influence Factor} \quad (1)$$

Influence maximization: After the network of users is determined, the next step is find a subset of users from the network set that will help propagate the advertisement throughout the network. Influence maximization is the process of finding a set of k users from the network that will have the highest influence in the network called seeds.

The influence maximization problem is a NP-hard problem as it is a combinatorial problem that involves building a set S by adding users represented as vertices in the graph that generate the maximum influence in the network.

For solving the NP-hard problem we use a greedy heuristic method which is based on the linear threshold model (Goyal *et al*., 2011). This algorithm also considers the constraints by Leskovec *et al*. (2007) and Heer and Boyd (2005). The theory proclaims that influence by a user is only to his direct neighbors or the neighbors of his neighbors and usually stops at two hops from his/her social link. Thus, it is often sufficient to consider the propagation within a few hops of the seeding. There are other algorithm for influence maximization (Mislove *et al*., 2007). But, the algorithms must be based of Linear Threshold model (Kim and Han, 2009) as weights are assigns to users.

Information diffusion: Once the seeds set is obtained from influence maximization module the next step is to determine the propagation of information though thenetwork. Hence, information diffusion module determines the spread or propagation of the advertisement across the network. We use the linear threshold model (Kim and Han, 2009) for information diffusion. The linear threshold model considers the probable threshold of a user and the chances of it being influenced based on the number of neighbors that are influenced by the product. The node has high chances on being influenced only if the number of its neighbors influenced have crossed his/her threshold.

Visualization: The visualization module is responsible for depicting the result onto a geographic map and generating graphs to infer results related to the effectiveness of the product in the network.

For geographic map visualization we use google map API which is a web service that takes in the coordinate information and displays the map with the output on a webpage. The visualization consists of markers and polylines, where markers display the location of the node geographically and polylines the influence from one node to another the geographic visualization consists of two parts.

Individual layer: In this map the influence between individuals is portrayed. The seeds are marked differently from the active users and the information boxes display information related to the seed or active user. The polylines showcase the spread of influence throughout the network.

City layer: In this map the influence between cities is portrayed. The markers represent the cities were there existing at least one seed/active user. The polylines depict the influence between cities. Information boxes display the number of seeds and active users in that particular city.

For inferring results related to the effectiveness the advertisement has on the network is depicted by using graphs. Here, we use google charts API which is a web service that takes as input plot values of the graph and displays the graph on a web page.

RESULTS AND DISCUSSION

Results: The result is depicted in the interface the advertiser will interact with. The Window consists of a menu bar, i.e. are links to different results that are useful to infer the effectiveness of the product in the social network. The various results in the menu include:

Individual layer: It depicts the information diffusion result obtained by a selected seed set for a coverage level of 70%.

City layer: It displays the user base, i.e., a combination of seeds and activated users in each city.

Seed coverage ratio: It shows the result of the influence maximization algorithm (Goyal *et al*., 2011) which depicts the number of seeds required to cover a percentage of the social network.

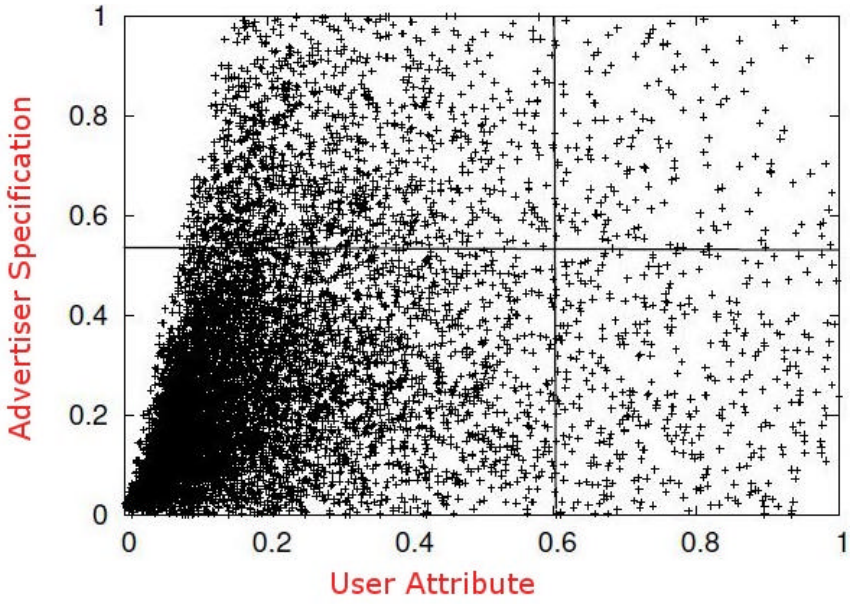

Fig. 3: Scatter plot specifying the user 's influence level

Seed influence measure: It depicts the most influential seeds in the social network, making the advertisers choice of selecting seeds easier.

City influence measure: It depicts a basket analysis measure, where the spread of information in each city is shown by a pie chart.

Demonstration

User selection: The user selection filters users according to Principal Component Analysis (PCA) (Mislove *et al*., 2007). Input to user selection module is the advertiser specification which is matched against the user characteristics. The advertiser specification for a sports product is shown in Table 1 for which our application is executed.

The various attributes assigned by the advertiser are restrictions on the user base and the keyword attributes are words he wants his product to be associated while filtering users. The user selection module gives us a scatter plot shown in Fig. 3 that is in accordance to Goyal *et al*. (2011) which gives us the influence factor of the users.

Influence maximization: The influence maximization algorithm is adapted from Goyal *et al*. (2011) and gives us seeds based on the required social network coverage ratio. The input to the algorithm is the user base obtained from the user selection model. The number of seeds obtained is summarized based on the social network coverage ratio provided in Table 2 and Fig. 4.

Table 1: Advertiser's Specification

Characteristic name	List of attributes
Age	Lower bound = 15; Upper bound = 40
Country	Null
City	Null
Brand Keywords	Nike, Adidas, Puma, Umbro, New Balance, Reebok
Personality Keywords	Ronaldo, Messi, Beckham
Category Keywords	Athlete, Product

Table 2: Seed coverage ratio that determines the number of seeds required for a coverage percentage coverage

Coverage ratio (%)	No. of seeds
50	14
60	23
70	36
80	52
90	76

Information diffusion simulation: Given the seeds from the influence maximization algorithm the information diffusion simulation depicts the information dissemination across the social network. We employ the linear threshold model (Kim and Han, 2009) for information diffusion process as we give weights to our users rather to links in the social network. The spread can be show through the individual layer of our visualization results shown in Fig. 5. Figure 5 shows the various users participating in the diffusion process by the pins in the map. This helps in inferring the dispersion of users in various locations. The green markers represent seeds and the red markers represent active users. And the blue links the diffusion links. Figure 6 shows the information about the user which is obtained by clicking on the markers.

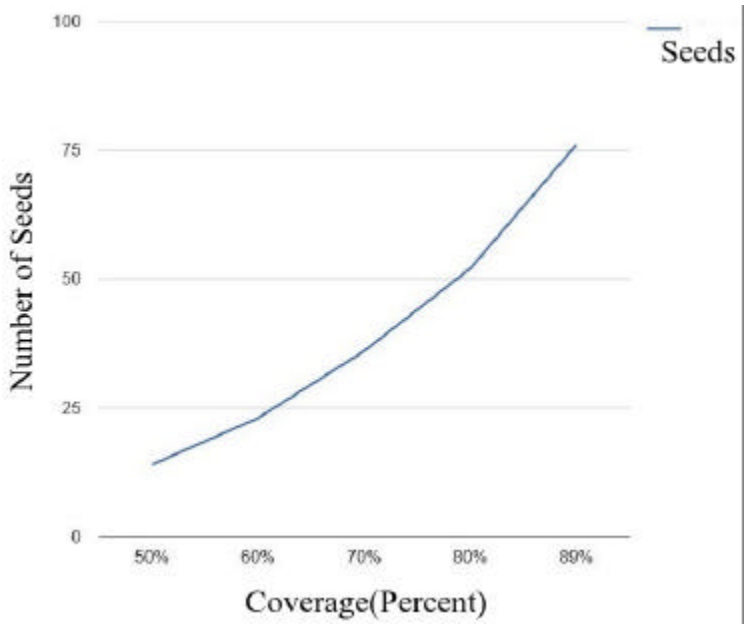

Fig. 4: Coverage ratio depicting the seeds required for a particular coverage percentage

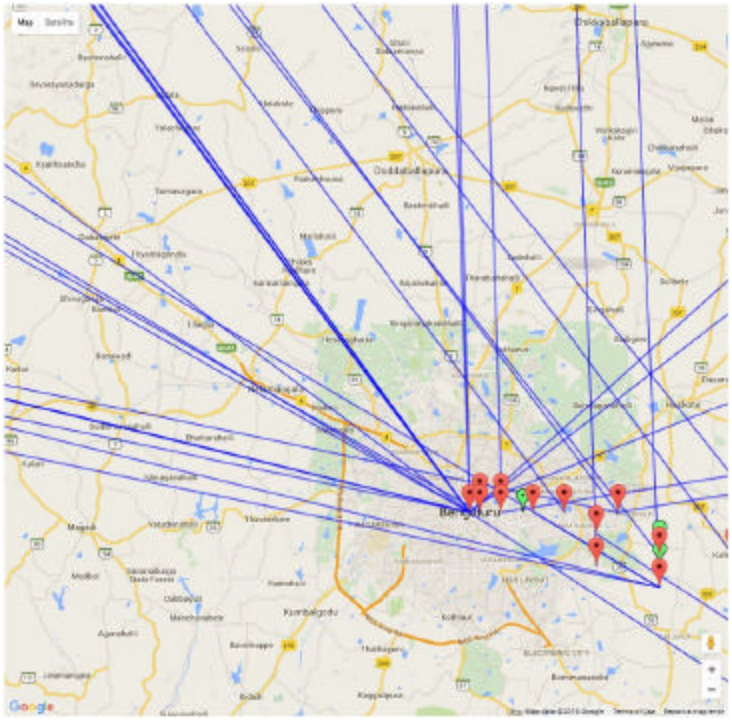

Fig. 5: Information diffusion showing the spread of information in a network

Visualization: This module helps advertisers infer the effectiveness of the product in a particular socialnetwork. It consists of graphs and geographic maps. The significance of these results have been mentioned in results section of V A. Individual layer shown in Fig. 5, city layer shown in Fig. 7, seed coverage ratio shown in Fig. 4, seed influence measure shown in Fig. 8, city influence measure shown in Fig. 9.

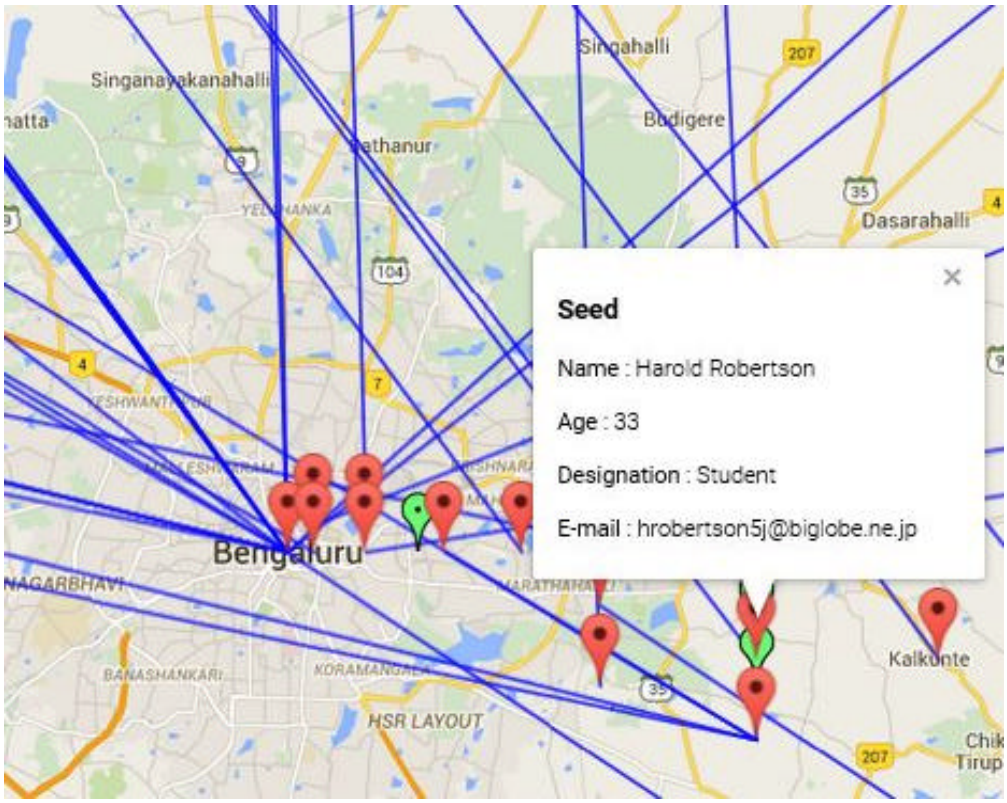

Fig. 6: Information about a user as seen in the Info window

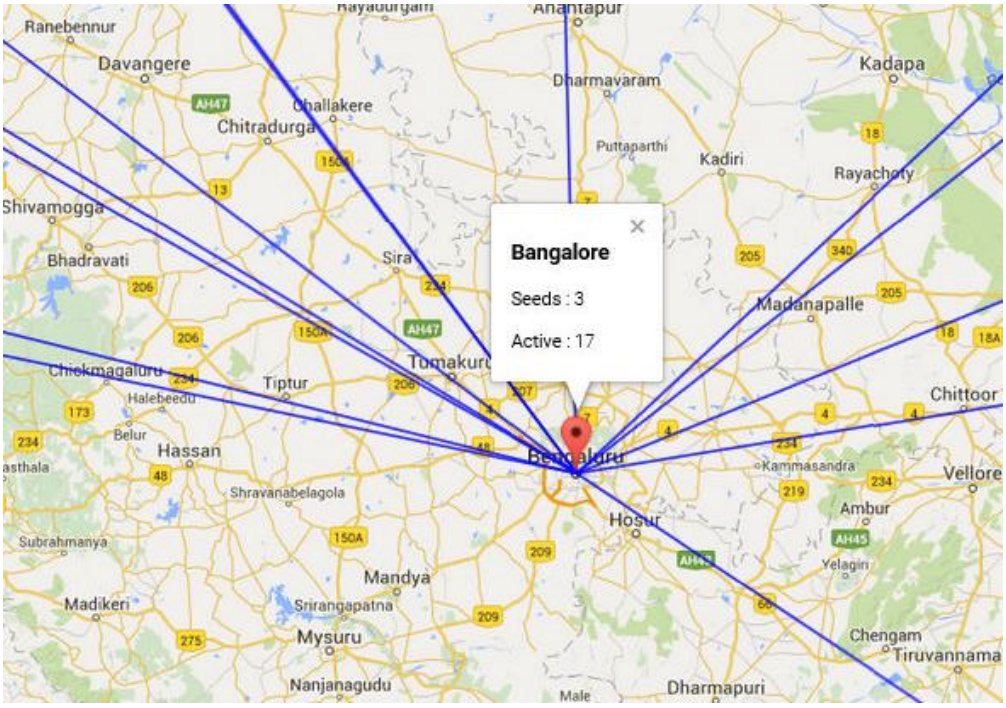

Fig. 7: City layer depicting the number of seeds and active users in the network

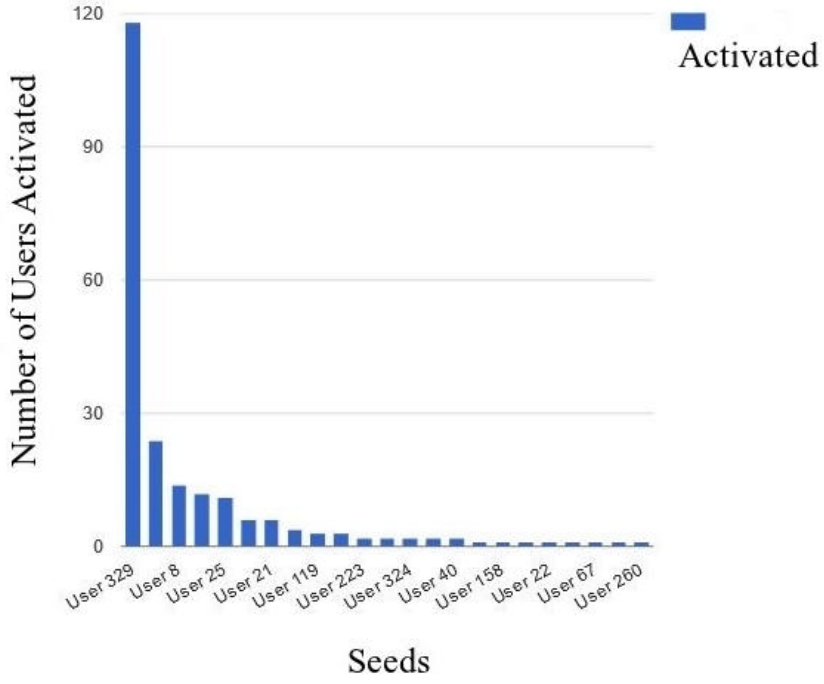

Fig. 8: Top few seeds that have the maximum influence in the network

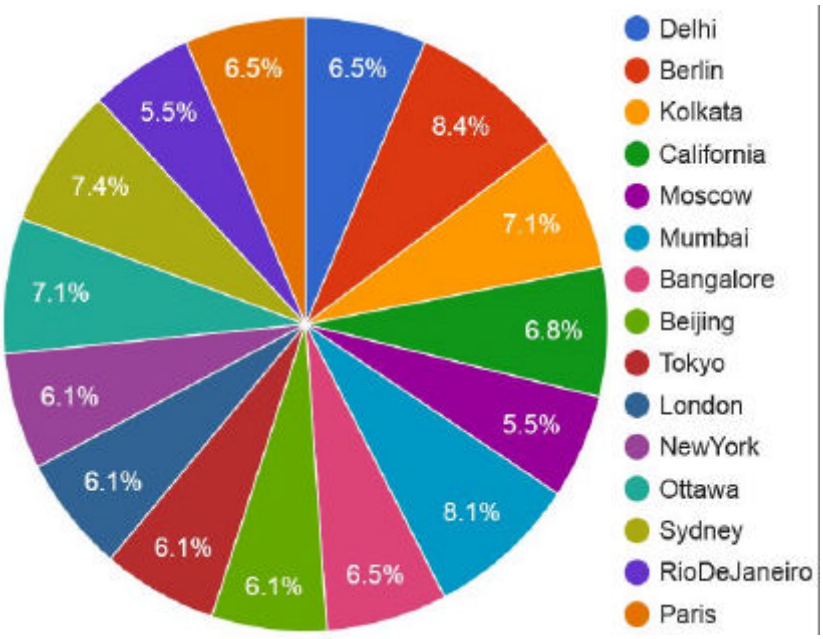

Fig. 9: Pie chart that shows the percentage of users in each city

CONCLUSION

Viral marketing, in recent times has gained popularity because of the growing influence social networks has on our lives. Conventional means for marketing like radio, television and newspapers have started fading. Now social networks have become the pinnacle source for marketing as the cost involved for marketing is less in

comparison to the conventional means. Hence, advertisers are flocking to social media platforms to market their products.

Our application encompasses three techniques vital for a viral marketing application that include user segregation, influence maximization and information diffusion. These concepts help in obtaining the required seeds, determine the influence these seeds have on the network and visualize the result geographically. The application helps an advertiser find the appropriate seeds required to market his product.

REFERENCES

Canali, C., S. Casolari and R. Lancellotti, 2010. A quantitative methodology to identify relevant users in social networks. Proceeding of the 2010 IEEE International Workshop on Business Applications of Social Network Analysis (BASNA), December 15-15, 2010, IEEE, New York, USA., ISBN:978-1-4244-8999-2, pp: 1-8.

Chen, W., Y. Wang and S. Yang, 2009. Efficient influence maximization in social networks. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28-July 01, 2009, ACM, New York, USA., ISBN: 978-1-60558-495-9, pp: 199-208.

Dinh, T.N., H. Zhang, D.T. Nguyen and M.T. Thai, 2014. Cost-effective viral marketing for time-critical campaigns in large-scale social networks. IEEE. Trans. Networking, 22: 2001-2011.

Goyal, A., W. Lu and L.V. Lakshmanan, 2011. Simpath: An efficient algorithm for influence maximization under the linear threshold model. Proceeding of the 2011 IEEE 11th International Conference on Data Mining, December 11-14, 2012, IEEE, British, Columbia, ISBN:978-1-4577-2075-8, pp: 211-220.

Heer, J. and D. Boyd, 2005. Vizster: Visualizing online social networks. Proceedings of the 2005 IEEE Symposium on Information Visualization INFOVIS 2005, October 23-25, 2005, IEEE, Berkeley, California, ISBN:0-7803-9464-X, pp: 32-39.

Kempe, D., J. Kleinberg and E. Tardos, 2003. Maximizing the spread of influence through a social network. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington, DC, USA., pp: 137-146.

Kim, E.S. and S.S. Han, 2009. An analytical way to find influencers on social networks and validate their effects in disseminating social games. Proceeding of the International Conference on Advances in Social Network Analysis and Mining ASONAM'09, July 20-22, 2009, IEEE., Daejeon, South Korea, ISBN:978-0-7695-3689-7, pp: 41-46.

Lam, H.W. and C. Wu, 2009. Finding influential ebay buyers for viral marketing a conceptual model of BuyerRank. Proceeding of the 2009 International Conference on Advanced Information Networking and Applications, May 26-29, 2009, IEEE, Perth, Western Australia, ISBN:978-1-4244-4000-9, pp: 778-785.

Leskovec, J., L.A. Adamic and B.A. Huberman, 2007. The dynamics of viral marketing. ACM. Trans. Web, 1: 5-5.

Mislove, A., M. Marcon, K.P. Gummadi, P. Druschel and B. Bhattacharjee, 2007. Measurement and analysis of online social networks. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, October 23-26, 2007, San Diego, CA., USA., pp: 29-42.