A Self-Organizing Map for Identifying Influential Communities in Speech-based Networks

Sameen Mansha
Information Technology
University of Punjab, Lahore
sameen.mansha@itu.edu.pk

Faisal Kamiran
Information Technology
University of Punjab, Lahore
faisal.kamiran@itu.edu.pk

Asim Karim
Lahore University of
Management Sciences, Lahore
akarim@lums.edu.pk

Aizaz Anwar
Information Technology
University of Punjab, Lahore
aa364@itu.edu.pk

ABSTRACT

Low-literate people are unable to use many mainstream social networks due to their text-based interfaces even though they constitute a major portion of the world population. Specialized speech-based networks (SBNs) are more accessible to low-literate users through their simple speech-based interfaces. While SBNs have the potential for providing value-adding services to a large segment of society they have been hampered by the need to operate in low-income segments on low budgets. The knowledge of influential users and communities in such networks can help in optimizing their operations. In this paper, we present a self-organizing map (SOM) for discovering and visualizing influential communities of users in SBNs. We demonstrate how a friendship graph is formed from call data records and present a method for estimating influences between users. Subsequently, we develop a SOM to cluster users based on their influence, thus identifying community-level influences and their roles in information propagation. We test our approach on Polly, a SBN developed for job ads dissemination among low-literate users. For comparison, we identify influential users with the benchmark greedy algorithm and relate them to the discovered communities. The results show that influential users are concentrated in influential communities and community-level information propagation provides a ready summary of influential users.

Keywords

Speech-based social networks, Influential users and communities, self organizing maps.

1. INTRODUCTION

With the inception of high speed internet and fast computing resources, online social networking has become an essential tool for connecting various offline communities all across the globe. Despite the popularity of social networks a large portion of the world's population is still unable to be part of online social networks. The prime reasons for this situation are rooted in low-literacy, poverty, and in some places, lack of internet connectivity. Recently, it has been demonstrated that speech-based and non-textual systems are favored by low-literate users over text-based ones [11]. For this reason, some speech-based networks (SBNs) have been introduced to cater to the needs of low-literate users by providing speech-based interfaces for communication [16, 13, 14].

SBNs often face scalability and sustainability challenges in practice. A major reason for this is that most users of SBNs have low buying power and they cannot pay charges for the services. This hurdle in the widespread use of SBNs can be overcome by devising mechanisms for optimizing the utilization of resources. One way to maximize information propagation and hence reduce operation costs is by taking advantage of influential users and communities in the network. However, identification of influential communities has not been reported for SBNs, and methods developed for other social networks are not directly applicable to SBNs because of differences in characteristics.

In this work, we develop a self-organizing map (SOM) for discovering influential communities in SBNs. A SOM is a powerful information clustering and visualization technique that not only finds groups of related objects but also displays the relationships among the groups in two-dimensions for easy interpretation. We adapt SOM to the clustering of users based on their inter-user influences. The resulting communities (clusters) are related to others thus exposing the community-level structure of information propagation in the network. We demonstrate how a social network graph is constructed from an SBN's call data records (CDR). We also present a method for estimating the influence of a user on another based on the information in the CDR. We apply our SOM on a real-world SBN named Polly. The results are validated by identifying influential users using the general greedy algorithm [8] in each community.

2. MOTIVATION AND RELATED WORK

Typical speech-based social networks provide telephone-based voice manipulation and forwarding system for communication and dissemination of development-related infor-

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983885

mation, e.g., provision of job ads, group messaging facility, etc. For example, Polly is a telephone-based entertainment service that allows any caller to record a short message, modify it using a choice of funny sound effects, and forward the modified recording to their friends by phone numbers. Users can also browse a list of audio recorded job ads that can be forwarded to their friends [1]. Polly was first tested in 2011 among low skilled office workers in Lahore, Pakistan. Within 3 weeks, Polly spread to 2,032 users, engaging them in 10,629 calls. However, it was shutdown due to unsustainable cellular airtime cost and insufficient telephone capacity [14]. After a year of rebuilding and capacity enhancement, Polly was launched again in May 2012 [15]. Keeping in view the demands for sustainability of SBNs such as Polly, social influence of users and communities in the network can be used as a criterion for the allocation of resources. Influence maximization aims to identify users in social networks that can lead to greatest influence (e.g., spread of information). This task has been studied extensively for conventional social networks for applications like viral marketing. The greedy algorithm, which is considered a benchmark algorithm for this task, finds top-k influential users [8]. Influence propagation patterns have been studied in Polly [10] and the task of influence maximization has been addressed [3]. More recently, the idea of influential communities in social networks has been introduced to expose the community-level structure of users and their roles in information propagation [4, 12]. We focus on community-level influence but target SBNs with their unique features encoded in call data records. In the area of social network analysis, some works have reported characterization of individual users by applying self-organizing maps (SOM) [5, 6] but none have applied it for identification of influential communities in SBNs.

3. METHODOLOGY

First, we discuss the construction of social network graph from a SBN's call data records (CDR), including a method for estimating influence between users. Then, we develop a SOM for finding and visualizing influential communities.

3.1 Social Network Graph

We process the CDR of the SBN to construct the social network graph. Typically, the CDR contain information about users, their interactions (Short Message Service (SMS) and audio/voice messages), and usage of the network.

A social network graph $N_G = \{V, E\}$ is constructed from the CDR where V is the set of vertices, E is the set of edges and N_G is a weighted directed graph. Each vertex $\mu \in V$ in N_G corresponds a unique user in the CDR. An edge $(\mu, \nu) \in E$ from vertex $\mu \in V$ to vertex $\nu \in V$ exists in N_G when user μ sends voice/text messages to user ν. Every edge is associated with a weight that is the influence of the source user upon the target user (calculation discussed below). The in-degree of vertex μ is the number of other users (vertices) that send messages to (link to) user μ, while the out-degree of vertex μ is the number of other users that user μ sends messages to. In general, N_G can contain disconnected components when a subset of vertices in N_G have no edges to other vertices.

Calculating Influence: Determining influence probability between users in a social network accurately is crucial to the influence maximization problem. Although previous work (e.g., [7]) has discussed estimation of influence probabilities in social networks. We introduce two different parameters to assign edge weights (influence) based on available information in the CDR of SBN.

1. Interaction Indicator ($r(\mu, \nu) \in \{0, 1\}$): This interaction indicator is equal to 1 when user μ sends one or more voice messages to user ν; otherwise, it is equal to zero.

2. Spread Factor ($s(\mu, \nu)$): Equation 1 gives the influence spread of user μ over user ν where F is the number of voice messages that user μ has forwarded to user ν and M is the number of voice messages from user μ that user ν has received and forwarded to other users.

$$s(\mu, \nu) = \frac{M}{F} \tag{1}$$

Suppose a user μ forwards 3 voice messages (m_1, m_2, m_3), such as job ads, to user ν who then forwards m_3 to other users, then $s(\mu, \nu) = 1/3 = 0.33$.

The influence of user μ on user ν in N_G is defined as

$$INF(\mu, \nu) = s(\mu, \nu) + r(\mu, \nu) \tag{2}$$

$INF(\mu, \nu)$ can range in the interval [0,2] with higher values signifying greater influence (i.e., capability to spread information) in the network. Notice that $INF(\mu, \nu) \geq 1$ whenever user μ interacts with user ν but this influence will increase with the number of messages of μ forwarded by ν. Moreover, in general, $INF(\mu, \nu) \neq INF(\nu, \mu)$ as each user can behave differently. We set the influence of a user on itself equal to 2, i.e., $INF(\mu, \mu) = 2$.

All inter-user influences in N_G can be captured in an adjacency matrix X of size $N \times N$ where $N = |V|$ is the number of vertices in N_G. That is, each element of X , $x_{\mu\nu}$, is set equal to $INF(\mu, \nu)$. Note that X will not be a symmetric matrix in general. This matrix is used for identifying influential communities as discussed next.

3.2 SOM for Identifying Influential Communities

A self-organizing map (SOM) is a powerful unsupervised information processing and pattern recognition method [9]. It is a type of artificial neural network used to convert complex nonlinear statistical relationships between high dimensional data into simple geometric relationships on a low-dimensional display. Generally SOM consists of a layer of units or neurons that adapt themselves to input patterns such they are activated for similar patterns only. The adaption is done through an iterative procedure which processes input patterns repeatedly. Upon presentation of each input pattern, neuron closer to it and all its neighbors are moved closer to input pattern. After a significant number of iterations all neurons move into the area of high concentration of input patterns.

To find influential communities and the relationships among them in N_G, we cluster $X = [x_1, x_2, \ldots, x_N]$ using a SOM. The ith user is represented by a feature vector x_i comprising of its influences (INF) on other users and corresponding to the ith row of X. Two users that influence the same set of users strongly are likely to fall in one cluster, and a collection of such users will form a community of users that influence each other strongly. In a SOM, each community is represented by a neuron, and the geometric closeness of

Algorithm 1 SOM for Influential Community Discovery in SBNs

Input: Adjacency matrix $X = [x_1, \ldots, x_N]$ of N_G
Output: A topology of influential communities, $Y = \{y_1, \ldots, y_K\}$
begin
  Initialize weight matrix randomly, $W = \{w_1, \ldots, w_K\}$
  repeat select $x \in X$ randomly
    determine winning neuron y' such that
    $y = \arg\min_{y \in Y} d(x, w_y)$
    for all $y \in h(y') \subset Y$
      $w_y = w_y + \eta(t)(x - w_y)g(y, y)$
    decrease η by a small amount
  until termination condition is true
end

two neurons show the relative strength of influence between the respective communities. Let $Y = \{y_1, y_2, \ldots, y_K\}$ be the set of neurons corresponding to the communities to be discovered. We consider a 2-D $\sqrt{K} \times \sqrt{K}$ arrangement for the neurons; thus $\sqrt{K}$ has to be an integer and each neuron $y_i = (a, b)$ is identified by its coordinates in the 2-D grid where a and b can be integers from 0 to $\sqrt{K} - 1$. The weight matrix of the SOM is given by $W = [w_1, w_2, \ldots, w_K]$ where w_i is the weight vector of size N of the ith neuron.

Algorithm 1 shows the steps required in training a SOM for identifying influential communities in an SBN (N_G). The SOM is initialized by randomly assigning weights in the interval $[-1, +1]$. Subsequently, features vectors $x \in X$ are fed randomly and repeatedly through the SOM. For each input x the 'winning' neuron y is found. This is the neuron whose weight w_y is closest to input x. This is determined via the distance function $d(x, w_y)$. We use the Euclidean distance for this purpose. The neighborhood function $h(y)$ returns the neurons neighboring y. In our case, the immediate neighbors of y in the 2-D grid are considered. For each neighboring neuron y, the corresponding weight w_y is updated such that it is brought closer to x. The learning rate parameter $\eta > 0$ controls the amount of update in each iteration. Furthermore, the update for the winning neuron is stronger than that for its neighbors. This is defined by the Gaussian function $g(y, y)$ centered on y. The learning rate is reduced with the number of iterations to smooth out the convergence of the SOM. Convergence occurs when change in the weight matrix W from one iteration to the next becomes acceptably small.

After the SOM is trained, each user is assigned to the neuron whose weight is closest to the feature vector of the user. In this way, a clustering of users is obtained based on influence. Each cluster is actually a community of users and the relative position of the communities in the 2-D grid indicates their relationship w.r.t. influence and information propagation.

4. EXPERIMENTAL RESULTS

In this section, we discuss the results of our experiments on a real-world speech-based network.

Dataset Description: We evaluate our methodology on Polly which is a telephone-based social network developed for providing services to low-literate users in South Asia [1]. The CDR that we use contains 213,196 users with 2,522,981 interactions among them recorded in more than 700 MB of log data and 200 GB of audio data.From this information we construct the social network graph N_G whose vertices represent unique users and its edges are labeled with influence values (INF). Influences are calculated by considering number of calls, job ads listened, and job ads forwarded. Table 1 shows key characteristics of CDR and (N_G) used.

Table 1: Description of Call Data Records (CDR) and Social Network Graph (N_G)

Key Statistics of Polly's CDR	
No. of users	213196
No. of calls made	2522981
No. of requests for calls	1051669
No. of requests for job delivery	33597
No. of times job ads listened	386199
Key Characteristics of N_G	
No. of edges	13694
No. of connected components	3272
Vertices in biggest connected component	2407
Edges in biggest connected component	3332
Diameter of biggest connected component	10

Community-level Social Influence Analysis: Figure 1 shows the output of our SOM when applied to Polly. We discover 25 communities (clusters of users) that are arranged in a 5×5 grid of cells. Each cell in the grid is identified by its 2-D coordinates with values starting from 0. For example, the bottom left and top right cells (communities) are identified as (0, 0) and (4, 4) respectively. The background color of each cell indicates the normalized sum of the distances between the corresponding neuron's weight and those of its neighbors. A darker background color of a cell signifies that the corresponding community is closer to its neighbors w.r.t. influence, i.e., their influence potential are similar. We also find influential users in the network by using greedy algorithm [8]. The top 50 influential users are shown by yellow dots in their respective communities. Our implementation of SOM for finding influential communities is based on the MiniSom tool [2].

It is observed from Figure 1 that the greedy algorithm finds 5 users from the community (4,4), 4 users each from (2,1), (4,0) and (4,1), and 3 users each from communities (0,4), (1,3), (2,2), (3,1), (3,3) and (4,2). On the other end, communities (0,1), (1,2), (2,3), (3,0), and (3,2) do not contain any top-50 influential users. Thus, top-50 influential users appear in only a few communities. These are therefore the influential communities in Polly. It is also seen from the background colors that influential communities lie closer to each other. For example, communities (2,1), (2,2), (3,1), (4,0),(4,1), (4,2) and (4,4) contain 26 top-50 influential users (i.e., 52% of influential users lie in 7 communities) and they are placed closer to each other on the grid. On the other hand communities that contain less number of any top-50 influential users are also closer to each other, e.g., (0,0), (1,0), (2,0), (3,0) and (1,2). The closeness of influential communities on the grid provide a clearer idea of community-level information propagation. After spreading within an influential community, information introduced in it is more likely to spread to neighboring communities on the grid than to other communities.

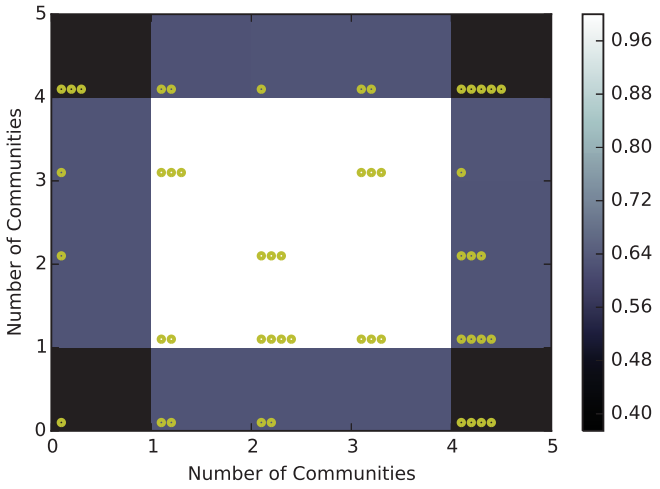

Figure 1: Community-level influence analysis: 5×5 communities with top-50 influential users identified by yellow dots

5. CONCLUSION

In this paper, we present a self-organizing map (SOM) for finding influential communities and exposing the community-level structure of influence in a speech-based network (SBN). A SBN differs from conventional social networks in that it provides voice-based interactions among users for services aimed towards low-literate people in low income regions. We discuss the construction of a social network graph from an SBN's database including the important aspect of determining influence between users. Our SOM not only identifies important communities based on influence but also displays their relationships. As such, the SOM summarizes information propagation in a social network that can help in optimizing its operations for scalability and sustainability. We demonstrate our methodology on a real-world SBN operated for disseminating job ads among low-literate users.

As future work, it would be interesting to evaluate SOM for community-level influence analysis in conventional social networks. For SBNs, further studies are needed to understand its dynamics such as growth, attrition, influence, and societal impact.

Acknowledgment

We would like to show our gratitude to Dr. Agha Ali Raza (Information Technology University, Pakistan) and Prof. Roni Rosenfeld(Carnegie Mellon University, Pittsburgh, USA) for their cooperation and providing us access to Polly's Call Data Records and log files.

6. REFERENCES

[1] https://www.cs.cmu.edu/~./Polly/. [Online; accessed 2016-22-08].

[2] https://github.com/JustGlowing/minisom. [Online; accessed 2016-22-08].

[3] A. Anwar, S. Mansha, F. Kamiran, and A. Karim. Identification of influential users in speech-based networks. In *Pacific Asia Conference on Information Systems (PACIS) Proceedings*, 2016.

[4] N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In *International Conference on Web search and Data Mining Proceedings*, pages 33–42. ACM, 2013.

[5] R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel som and related laplacian methods for social network analysis. *Neurocomputing*, 71(7):1257–1273, 2008.

[6] A. Bruggmann, M. M. Salvini, and S. Fabrikant. Cartograms of self-organizing maps to explore user generated content. In *International Cartographic Conference Proceedings*, pages 25–30, 2013.

[7] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *International Conference on Web search and Data Mining Proceedings*, pages 241–250. ACM, 2010.

[8] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *International Conference on Knowledge Discovery and Data Mining Proceedings*, pages 137–146. ACM, 2003.

[9] T. Kohonen, editor. *Self-organizing Maps*. Springer Verlag New York, Inc., Secaucus, NJ, USA, 1997.

[10] Y. Lin, A. A. Raza, J.-Y. Lee, D. Koutra, R. Rosenfeld, and C. Faloutsos. Influence propagation: Patterns, model and a case study. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining Proceedings*, pages 386–397. Springer, 2014.

[11] I. Medhi, A. Sagar, and K. Toyama. Text-free user interfaces for illiterate and semi-literate users. In *International Conference on Information and Communication Technologies and Development Proceedings*, pages 72–82. IEEE, 2006.

[12] Y. Mehmood, N. Barbieri, F. Bonchi, and A. Ukkonen. Csi: Community-level social influence analysis. In *Machine learning and Knowledge Discovery in Databases Proceedings*, pages 48–63. Springer, 2013.

[13] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh. Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india. In *Human Factors in Computing Systems Conference Proceedings*, pages 733–742. ACM, 2010.

[14] A. A. Raza, M. Pervaiz, C. Milo, S. Razaq, G. Alster, J. Sherwani, U. Saif, and R. Rosenfeld. Viral entertainment as a vehicle for disseminating speech-based services to low-literate users. In *International Conference on Information and Communication Technologies and Development Proceedings*, pages 350–359. ACM, 2012.

[15] A. A. Raza, F. Ul Haq, Z. Tariq, M. Pervaiz, S. Razaq, U. Saif, and R. Rosenfeld. Job opportunities through entertainment: virally spread speech-based services for low-literate users. In *Human Factors in Computing Systems Conference Proceedings*, pages 2803–2812. ACM, 2013.

[16] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. Healthline: Speech-based access to health information by low-literate users. In *ICTD Conference Proceedings*, pages 1–9. IEEE, 2007.