Journal of Engineering Science and Technology
Vol. 16, No. 2 (2021) 1299 - 1310

AN EMPIRICAL STUDY OF VIRAL MARKETING ON ONLINE SOCIAL NETWORKS USING DIMENSION REDUCTION TECHNIQUES

UMASANKAR DAS[1,*], GIRIJA PRASAD MOHAPATRA[2]

[1]MCA Department, Silicon Institute of Technology Bhubaneswar, Odisha, India
[2]Tata Consultancy Services (TCS), Bhubaneswar, Odisha, India
*Corresponding Author: umasankar@silicon.ac.in

Abstract

The study of the viral marketing strategies in online social networking sites that how to reach to the target audience to maximize the positive traffic. We propose a model in which the decision of a buyer group is taken into consideration based on their interest. The recommendation model is quite general based on centrality threshold and interest. In this model, a seller is getting a user space to optimize the business strategy. Finding a seller strategy to maximize the expected revenue with a larger user space is NP-hard. However, we propose a model to consider the centrality threshold to reduce the dimension rather than applying in a larger user space. Social media plays huge opportunities for the seller to reach out to potential consumers and currently is new forms of communication between providers and consumers. As providers and advertising agency search for some ways to reach to target audiences, our model can be useful to advertisers.

Keywords: Centrality, Clustering, Density and influence of nodes, Dimension reduction techniques, Viral marketing.

1. Introduction

Considering the popularity of the Online Social Networks (OSN), companies are adopting viral marketing as one of the easy methods which in other way is being referred as “word-of-mouse”. Viral marketing assorts to the selling strategy in which seller trigger message such as advertisement to others in online social communities and thus the message is circulated with less cost, effort as well as with a much better speed, just like diseases and virus are spread. According to the "small world" idea, pair of nodes in a network can connect in an indicate 2 way through a short path of mutual interest. If each user is estimated to be directly connected to 100 acquaintances, thus using a simple exponential model, 10 000 users are only two steps away from any given node/user and surprisingly, 3 million people are three steps away etc.

Viral marketing is the process of marketing in which the promotional messages about a product or a service are spread virally in the target market. Considering the popularity of OSNs companies are adopting viral marketing, because over there, spreading the message virally is easy like anything. So, the product is promoted in less time with less cost and effort and ultimately the goal is to reduce the cost and increase the sales revenue.

But considering the hugeness of social networks, sending a message to each and every node on a network does not seem to be any good idea either. As a solution, we can choose a set of nodes with greater influence and send the messages to those nodes only. Then the message is likely to be spread quickly in an economically optimized way. To find the most influential nodes we are considering network centrality and dimension reduction techniques. Centrality values can help us identify where to send the message in order get it spread the quickest. Dimension reduction techniques would help us in reducing the size of the network to be processed while finding the influential nodes.

2. Background & Related Work

When two users connected in social network with a direct link, they can have some similarity in interest and can have some trust involved. Social network users connect with each other, by sharing content and broadcasting information. There are a lot of social sites provide links, for users to communicate among each other (i.e., Twitter, Facebook, LinkedIn, MySpace) and platform for sharing content (i.e., YouTube, Flickr). When the consumers/users reach to these platforms, it is important to understand their interest and behaviour for a number of reasons. First, study of user-centric interest, and usage to know what they need and accordingly if advertising agency place or push different ads that can reach to maximum users quickly [1, 2] and widely. Second, understanding how the platform can be a model to handle traffic is equally important.

Graphical analysis

The social network can be represented [3] the structure of relationships between user, organizations, different goals, interests, and other entities within a larger distributed system.

Leadership network

In a highly interconnected network with a good degree of centrality if the platform can collaborate with each other in solving complex problems. It provides resources and support for leaders, and increase the scope and scale of impact, leaders can have individually and collectively to achieve [4].

2.1. Classification of leadership network

There are four different types of leadership networks; "peer leadership networks - socially connected through shared interests and commitments", "organizational leadership networks - structured to increase performance", "field-policy leadership networks - commitment to influence a field of practice or policy", and "collective leadership networks - common cause or focused on a shared goal".

Our concern is finding out the preferred set of nodes to which we can propagate advertisement. Though communities are defined by the structure of the links and individual connections depends on closeness among each node. Our proposed model takes the entire network and find out the influential node based on centrality then apply the dimension reduction technique to get a preferred set of nodes.

2.1.1. Degree centrality

It gives a measure of a node's connectivity in a particular community [4, 5] is in the form number of connection/degrees. In Fig. 1, users differ from one another only in how many connections they have. With directed data, it is important to observe both in-degrees as well as out-degree centrality. If a user having many connections, they seem to be influential within the same community or network. So, if we can direct other users fall in similar interest group to those highly connected users it would be better for the broadcasters to influence that high degree centrally nodes [6, 7] to propagate information.

Degree centrality is the sum of adjacencies for a vertex, *v (m)* over its maximum value, i.e., *N-1*:

$$C_d(v_i) = d_i = \sum_j A_{ij} \quad (1)$$

where A_{ij} is the weight of adjacent edge.

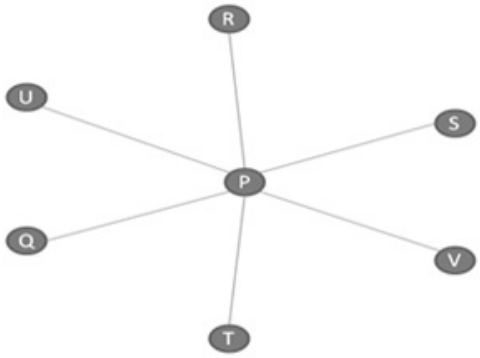

Fig. 1. An undirected graph.

Degree centralities of each node shown in Fig. 1 using simple degree centrality: Node P = 6, Node Q = 1, Node R = 1, Node S = 1, Node T = 1, Node U = 1, Node V = 1, (Here node P has highest degree centrality 6).

Closeness centrality

Closeness centrality is given mean of the shortest path between a vertex and all other connective vertices from it.

$$C_c(i) = \frac{1}{\sum_{j=1}^{n} dis(i,j)} \tag{2}$$

where n is the size of the network's, $dis(i,j)$ is the minimum distance between node i to j. Closeness centralities each node shown in Fig. 2 using distance:

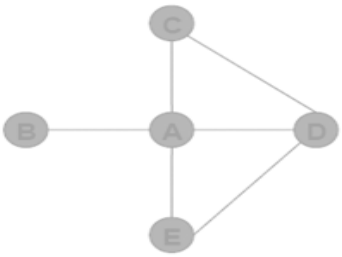

Fig. 2. An undirected graph.

Node A = 1/(1+1+1+1) = 0.25 , Node B = 1/(1+2+2+2) = 0.14 , Node C = 1/ (1+2+1+2) = 0.17 , Node D = 1/ (1+2+1+1) = 0.2 , Node E = 1/(1+2+2+1) = 0.17 (Here node A has higher closeness centrality 0.25).

2.1.2. Betweenness centrality

Betweenness centrality is the number of times a vertex occurs in the shortest paths between other vertices:

$$C_{B-SP}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

where σ_{st} is the number of shortest paths from source(s) to target(t), and $\sigma_{st}(V_i)$ is the number of shortest paths from s to t that pass through a vertex v. Table 1 shows the betweenness centrality of various node and the betweenness centrality of node 4 in the network shown in Fig. 3 using shortest paths.

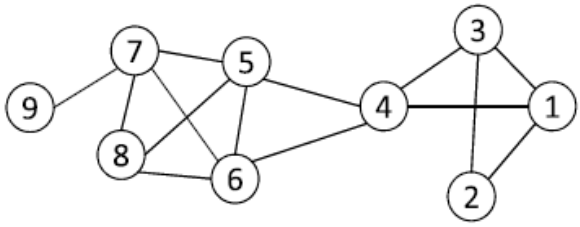

Fig. 3. An undirected graph.

Table 1. Betweenness centrality.

	S=1	**S=2**	**S =3**
T = 5	1/1	2/2	1/1
T = 6	1/1	2/2	1/1
T = 7	2/2	4/4	2/2
T = 8	2/2	4/4	2/2
T = 9	2/2	4/4	2/2

Here betweenness centrality of node 4 is 15.

2.1.3. Dimension reduction techniques

Considering high-dimensional datasets for getting target node set contains many mathematical challenges. The problems with high-dimensional datasets are that, to understand the all the properties and measure the interest. That leads to computationally challenging with a degree of accuracy. It is still of interest in many applications to reduce the dimension of the original data prior to any modelling of the data.

I. Proposed model

The objective of this paper is to predict the desired target user to spread advertisement. Since the users, in a social network having diverse interest and each one can be associated with each other on the basis of different factors, so in order to consider a set of the node we considered degree centrality with a threshold value.

In this section, we explore approaches used to find out semantic nature of recommender systems in various research articles, including design of a social network-based recommender system for the collaborative and participatory platform. In Fig. 4, schematic model of advertisement recommendation, combining degree centrality, social relation, and semantic analysis.

The contribution of the paper

- Design a model to reduce dimension and generate a recommendation using LSI/SVD [8-10]
- Considering influential node based on degree centrality to reduce desired network.

```
Pseudo code - degree centrality
READ the adjacent matrix a[i][j]
for i = 1 to row
        sum=0
        for j=1 to col
                sum =sum + arr[i][j]
        end for
        degcent[i] = sum,
end for
END
```

Then we apply rank-reduced, singular value decomposition to determine patterns in the relationships between the terms and interest contained in the profile of the node.

Indexing algorithm

Collection of n documents can be represented in the vector space model by a term-document matrix.

- An entry in the matrix corresponds to the "weight" of a term in the document.
- zero means the term has no significance in the document or it simply does not exist in the document.

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ \vdots \\ D_n \end{array} \begin{pmatrix} T_1 & T_2 & \ldots & T_t \\ w_{11} & w_{21} & \ldots & w_{t1} \\ w_{12} & w_{22} & \ldots & w_{t2} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ w_{1n} & w_{2n} & \ldots & w_{tn} \end{pmatrix}$$

Steps

tf-idf weighting typical: $w_{ij} = tfij*idf_i = tf_{ij} \log_2 (N/ df_i)$

f_{ij} = frequency of term i in document j

after normalize *term frequency* (*tf*) across the entire corpus:

$tf_{ij} = f_{ij} / \max\{f_{ij}\}$

df_i = document frequency of term i

= number of documents containing term i

idf_i = inverse document frequency of term i,

= $\log_2 (N/ df_i)$, (N: total number of documents)

A similarity measure is a function that computes the degree of similarity between two vectors.

Using a similarity measure between the query and each document:

- It is possible to rank the retrieved documents in the order of presumed relevance.
- It is possible to enforce a certain threshold so that the size of the retrieved set can be controlled.

Similarity between vectors for the document di and query q can be computed as the vector inner product:

$$sim(d_j, q) = d_j . q = w_{ij} . w_{iq}$$

where w_{ij} is the weight of term i in document j and w_{iq} is the weight of term i in the query. The cosine similarity measures the cosine of the angle between two vectors as provided below.

$$CosSim(d_j, q) = \frac{\vec{d_j} . \vec{q}}{\left|\vec{d_j}\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{iq}^{2}}} \quad (4)$$

Pseudo code for information retrieval

1. Convert all documents in collection D to tf-idf weighted vectors, $\boldsymbol{d_j}$, for keyword vocabulary V.
2. Convert query to a tf-idf-weighted vector $\boldsymbol{q}$.
3. For each $\boldsymbol{d_j}$ in D do

   Compute score $s_j = \text{cosSim}(\boldsymbol{d_j}, \boldsymbol{q})$
4. Sort documents by decreasing score.
5. Present top ranked documents to the user.

The issue lies with vector space information Retrieval (IR) is missing semantic information.

Missing syntactic information (e.g., phrase structure, word order, proximity information).

Singular value decomposition (SVD)

SVD [11-13] is a well-known matrix factorization technique that factors an m x n matrix R into three matrices as the following:

SVD: The mathematical formulation
Let X be the M x N matrix of M N-dimensional points
SVD decomposition
X= U x S x VT
U(M x M)
U is orthogonal: UTU = I
columns of U are the orthogonal eigenvectors of XXT
called the left singular vectors of X
V(N x N)
V is orthogonal: VTV = I
columns of V are the orthogonal eigenvectors of XTX
called the right singular vectors of X
S(M x N)
diagonal matrix consisting of r non-zero values in descending order
square root of the eigenvalues of XXT (or XTX)
r is the rank of the symmetric matrices
called the singular values

Pseudo code - SVD

Take matrix A[i][j] as an input.

Find the eigen value using Ann ATnm

For each eigen value find the u

Convert the u into orthogonal format Umm.

Again, find out the eigen value using A TnmAmn

For each eigen value find the v

Convert the v into orthogonal format Vnn.

Take the transpose of the matrix Vnn

Take S as identity matrix

Represent as A mn = Umm SmnV Tnn

Implementation approach

- First express the interest as a matrix. The rows will consist of unique keywords (terms) while the columns will represent the documents from which they were extracted (Node after considering degree centrality).
- Next, the cells in the matrix will hold the counts of the keywords in their documents (i.e., columns)
- The counts are usually modified so that so that rare words are weighted more heavily than common words. A popular weighing technique is the TFIDF (i.e., Term Frequency – Inverse Domain Frequency)
- Next, SVD is applied to the generated matrix using an appropriate dimension.

Benefits of LSI in term similarity measure:

- LSI gives better result in situations i.e., Boolean keyword queries: multiple words that have similar meanings (synonymy) and words that have more than one meaning (polysemy).
- LSI performs automated document categorization.
- Because of its strict mathematical approach, LSI is inherently independent of language.

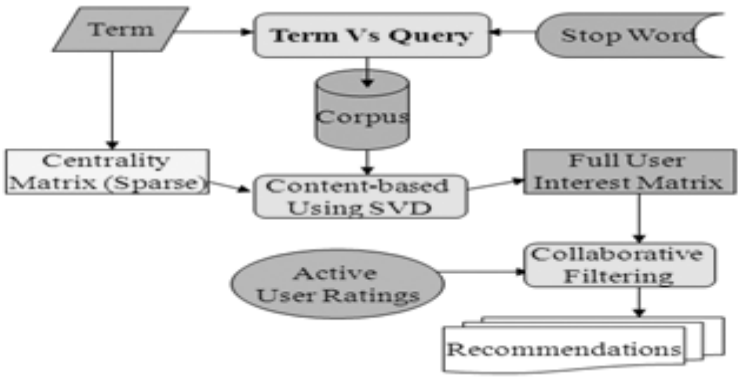

Fig. 4. Proposed system architecture.

II. Results

We evaluate our recommendation model by passing the search query to our preferred network Table 2 shows the simulation result.

Search query: “football a popular sports Shoes”
After stop word pass the search query is “football popular sports”
Football popular sports
no of words = 3
sports

1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2. Simulation result of a search query.

Search Query	Documents	Term Score of football	Term Score of popular	Term Score of sports
Football popular sports	I love cricket as well as football, but cricket is the most popular sports in India	0.77	0.75	1.00
Football popular sports	Ronaldo is a famous football player	0.94	-0.19	0.56
Football popular sports	Hockey is a family of sports in which two teams play against each other	0.84	0.66	1.00
Football popular sports	Hockey is the national game of India	-0.64	0.67	-0.06
Football popular sports	Badminton is a recreational sport played by either two opposing players(singles) or two opposing pairs (doubles) on opposite halves of a court	-0.64	0.67	-0.06
Football popular sports	Saina Nehwal is popular as a badminton player	-0.18	0.95	0.43
Football popular sports	A movie was made based on the story of Indian women hockey team.	-0.64	0.67	-0.06
Football popular sports	Vishwanathan Anand of India is the champion of chess game	0.64	-0.67	0.06
Football popular sports	Swimming keeps a man fit	0.64	-0.67	0.06

Evaluation criteria

The performance of the system is assessed by calculating the precision and recall. In the context of document, appropriate precision is calculated based on retrieved documents sample that is matched (documents that match input query), and recall is based on retrieved relevant documents. Table 3 shows the consideration conditions.

Let precision be denoted by P and recall be denoted by R. Formally, precision and recall are defined as:

$$\text{precision (P)} = \frac{|\{relevantdocuments\} \cap \{retrieveddocuments\}|}{|\{retrieveddocuments\}|} \quad (5)$$

$$\text{recall(R)} = \frac{|\{relevantdocuments\} \cap \{retrieveddocuments\}|}{|\{relevantdocuments\}|} \quad (6)$$

P = TP / (TP + FP)
R = TP / (TP + FN)

Table 3. Truth table.

	Relevant	Not relevant
Appropriate and Retrieved	True Positive	False Positive
Not Appropriate and Not Retrieved	False Negatives	True Negatives

where TP denotes the number of true positives, FP denotes the number of false positives, and FN denotes the number of false negatives.

Precision vs rank comparison

If we map precision against the change in k after koptimal, there is a fall in precision values. From the experimental evaluation we found that the precision saturates at rank (K=100), which we would call the optimal rank, Koptimal. To the left of Koptimal, the precision is low, because of under representation of prominent data from the hyperspace TDM. To the right of K optimal, the precision appears to be decaying slowly because of the over representation of the sub-space with added in noise components.

Accuracy measure

After passing through the LSI the entire retrieved keywords of query set are evaluated for precision. Based on the term co-occurrence pattern we are reducing the dimensions in our model; therefore, we do not require to measure recall. We observed the underfitting and overfitting criteria based on estimated k. In both the cases, the estimated co-occurrence by the model should be lower than the global maximum. Figure 5 shows precision (%) vs query set with respect to $k_{optimal}$.

To select an optimal rank k out of a set of the collection is challenging. The normal way of choosing optimal k by running a set of queries with known relevant documents for different k values, the k that gives the best performance is chosen as the $k_{optimal}$.

With this we can reduce the search space for target advertising for k best case, hence the different advertising agency will be interested in implementing our model to update preference list at the user location and over a time it will help to get better optimal k with better precision value.

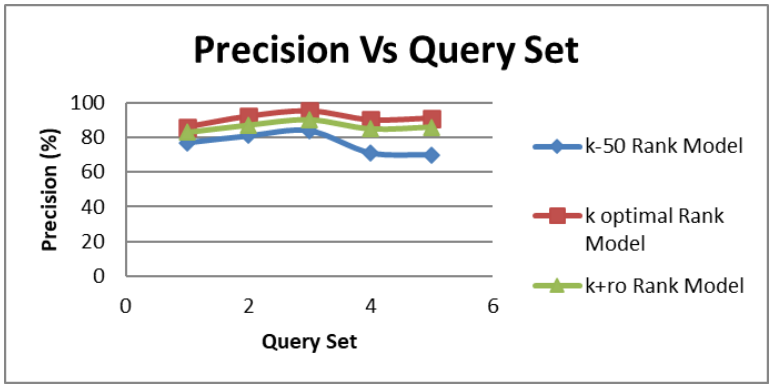

Fig. 5. Average precision scores generated for five different query sets.

The Query-relevance comparison with TF*IDF and LSI is shown in Table 4 and Fig. 6.

Table 4. Query-relevance comparison with TF*IDF and LSI.

Indexing method	Average precision			$K_{optimal}$ (k=100)
	K=50	K=100	K=150	
TF*IDF	.4545	**.4668**	.4599	-
LSI	.5020	**.5132**	.5045	+ 9.940017 (Improvement)

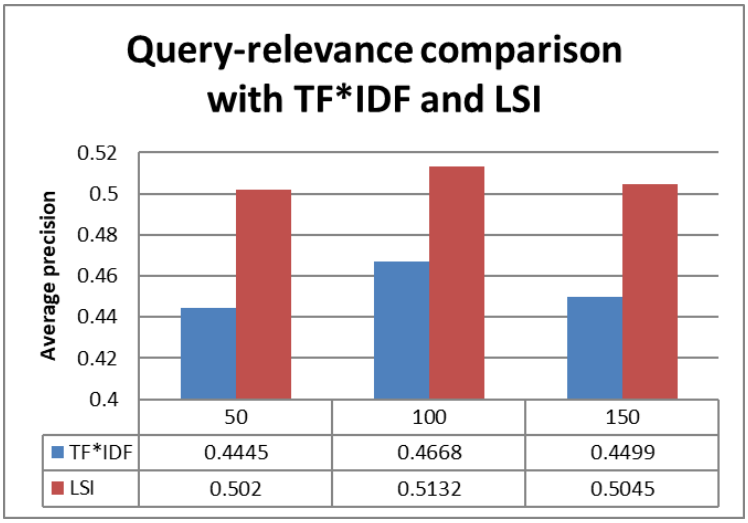

Fig. 6. Query-relevance comparison with TF*IDF and LSI.

3. Conclusion

In traditional marketing the drawback is that we choose a set of nodes for the promotion of our product and send message to those nodes, but we are never sure of what number out of those nodes are actually benefitted by that advertisement and even there are cases where these messages directly go to the spam folder of the users instead of being further spread as we expect it to.

Dimension reduction techniques can be applied to ensure that we have selected a set of nodes that can spread the message virally, so the purpose of information spreading can be achieved efficiently and effortlessly. We are using degree centrality and Singular Value Decomposition (SVD) method to select our target nodes for message propagation.

To achieve the general goal, we implemented one dimension reduction technique (SVD) and integrated it with node centrality to get lesser user space where only we took into consideration the frequency of our search term, however, we did not consider the semantic nature of the search term.

Our experiment has revealed that as part of SVD implementation with large node set is difficult to apply in real time, so in future, we will implement personalization with the localized search to optimize the solution.

References

1. Li, Y.; Zhao, B.Q.; and Lui, J.C.S. (2012). On modeling product advertisement in large-scale online social networks. *IEEE/ACM Transactions on Networking*, 20(5), 1-14.
2. Bulut, A. (2014). TopicMachine: conversion prediction in search advertising using latent topic models. *IEEE Transaction on Knowledge and Data Engineering*, 26(11), 2846-2858.
3. Golbeck, J.; and Mannes, A. (2006). Using trust and provenance for content filtering on the semantic web. *Proceedings of the WWW'06 Workshop on Models of Trust for the Web*. Edinburgh, Scotland, UK.
4. Hoppe, B.; and Reinelt, C. (2010). Social network analysis and the evaluation of leadership networks. *The Leadership Quarterly*, 21(4), 600-619.
5. Izquierdo, L.R.; and Hanneman, R. (2006). Introduction to the formal analysis of social network using mathematica. *Wolfram*.
6. Borgatti, S.P.; and Everett, M.G. (2006). A graph-theoretic perspective on centrality. *Social Networks*, 28(4), 466-484.
7. Lada A.& U(2009) Michigan, Network Centrality, *Computer Networks and ISDN Systems*, 30(1-7).
8. Cline, A.K.; and Dhillon, I.S. (2006). Computation of the singular value decomposition. *The University of Texas at Austin*, 1-13.
9. Kurucz, M. (2011). Data mining applications of singular value decomposition. Doctoral dissertation, *Eötvös Loránd University*, Budapest.
10. Baker, K. (2005). Singular value decomposition tutorial, 1-24.
11. Farooq, J.; and Osadebey, M. (2010). Presentation: dimensionality reduction with singular value decomposition & non-negative matrix factorization.
12. Chou, Y.-F.; Huang, H.-H.; and Cheng, R.-G. (2013). Modeling information dissemination in generalized social networks. *IEEE Communications Letters*, 17(7), 1356-1359.
13. Yang, D.; Zhang, D.; Zheng, V.W.; and Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1), 129-142.