International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 6 (2014) pp. 139 - 149
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Viral Marketing in an Online Discussion Forum

Shrihari A. Hudli, Aditi A. Hudli and Anand V. Hudli

Computer Science Department
MS Ramaiah Institute of Technology
Bangalore, India
shrihari@hudli.com

Computer Science Department
ObjectOrbTechnologies
MS Ramaiah Institute of Technology
Bangalore, India
anand.hudli@objectorb.com, aditi@hudli.com

Abstract—**Online opinion leaders play an important role in the dissemination of information in discussion forums. They are a high-priority target group for viral marketing campaigns. On an average, an opinion leader will tell about his or her experience with a product or company to 14 other people. It is important to identify such opinion leaders from data derived from online activity of users.**

We present an approach to modeling an online discussion forum using a two-mode social network called an affiliation network. Studying structural properties of the social network is a useful first step. In order to gain insight into other attributes of online users, it is necessary to follow a data mining approach. These observations lead to the representation of the online profile of each user as a set of attributes based on the online behavior of a user and that of other users as well. We present an approach to identification of opinion leaders using the K-means clustering algorithm. This approach does not require prior knowledge of the user's opinions or membership in other forums.

Keywords- online opinion leaders, social network analysis, affiliation networks, online discussion forum, data mining, clustering, supervised machine learning.

I. INTRODUCTION

Marketing studies have shown that certain online users exert an extra-ordinary impact on online and offline content and commerce [1]. Such users, called opinion leaders, generate most of the buzz about brands, products, and companies. Their spheres of influence expand as their peers pass along the messages about a product or company. Their standing and communication skills in the online community enable them to influence the opinion of others. These online opinion leaders are also more active users of e-mail, newsgroups, listservs, and express their opinions through other channels such as blogs and micro-blogs. They also forward news and website information to other people, send e-mail to companies, and post messages on discussion forums at least several times a month. While they are influential online, they are also approached offline for opinions on products. For example, according to a survey [1], more than 40% of opinion leaders say they offer advice to their peers about companies, businesses, or new technologies, hobbies, and family related issues.

Since opinion leaders are such an influential force, it is critical for companies to establish brand recognition and win the approval of these leaders to expand the customer base. It is important for companies to consider these opinion leaders, who are experts in collecting and spreading information online, in viral marketing campaigns.

While it is essential to communicate with all users of a company-run discussion forum regarding a product or service, it is imperative to identify those users in the discussion forum who are likely to be opinion leaders. These opinion leaders can then constitute a high-priority target group for viral marketing campaigns. Identifying the opinion leaders is an important first step in such a campaign.

Research in social media marketing [2] suggests a seven step process for a successful marketing campaign. In the first step, conversations about a product are monitored. This gives businesses access to valuable information, influential individuals, and relevant conversations. In the second step, influential individuals who can spread the marketing message are identified. In the third step, factors that are shared by the influential individuals are identified. This involves finding commonalities among these individuals and creating profiles of typical influencers. This enables the companies to locate all possible influencers relevant to their campaign and design methods to encourage those influencers to speak favorably about the company's products and services. Researchers have found that influencers in social media are highly engaged in three aspects, namely message spread, influence, and social impact. Message spread is related to the number of times a message is forwarded with or without modification. Influence is related to the number of times a message is forwarded to friends and family. Social impact is related to the number of replies and comments received for each message.

In the fourth step, potential influencers who have interests relevant to the specific marketing campaign are identified. It is not sufficient to identical influential individuals in the social network. These individuals must be particularly interested in the products and services offered by the company. In the fifth step, those influential individuals identified in the previous step are actually recruited to talk about the company's products and services. The sixth step consists of incentivizing the influencers recruited in the previous step to spread the marketing message across the social network. The incentives offered to the influencers could tangible, such as discounts and gifts, or intangible, such as recognition in the social network, or a combination of both. The final step is to enjoy the benefits of a successful social media campaign.

Diffusion of Innovation theory [3] recommends two techniques for identifying opinion leaders: the self-designation method and the sociometry method by means of questionnaires and interviews. In an Internet discussion forum with possibly thousands of users, these methods are expensive and difficult to administer and execute [4]. Further, adding semantic meaning to users' posts requires Natural Language Processing and name entity disambiguation [5].

Previous work in the area of identifying opinion leaders has focused on social network analysis [6] and also the user's interest space [7]. In social network analysis, not only each user's stated opinions are analyzed but also communication relationships among users are taken into account. User interest space analysis can be done using the knowledge of the user's membership in various discussion forums, where the discussion forums have a specific area of interest. Another way of finding user interests is to analyze article chains, where the area of interest of each article chain is known. Recent work has studied the effect of incorporating user semantic profile derived from past user behavior and preferences on the accuracy of a recommender system [8]. These approaches require collection of data that may or may not be feasible in all cases. Further, in an online discussion forum, a user does not necessarily have relationships with other users outside the forum and his/her interaction with others may be restricted to just participating in the discussions. In this paper, we focus on the problem of identifying opinion leaders in an online discussion forum when the focus of the discussion forum is clearly known and no knowledge of each user's membership in other forums is available. Further, no prior knowledge of each user's opinions is required.

The approach followed in this paper is as follows. Social network analysis is an important first step. We model a discussion forum using a special kind of social network called an *affiliation network*. Recognition of responses of users as being positive, negative, or neutral is also an important step. By analyzing the online activity of users, it is possible to construct an online profile of each user. We have chosen to represent the online profile by a set of eight attributes. Some of these attributes, such as the degree of positive feedback, require the implementation of a machine learning algorithm. These profiles or observations can then be analyzed using clustering methods used in data mining. The cluster corresponding to online opinion leaders is identified as part of the clustering process.

In Section II, we briefly discuss useful concepts in social network analysis. In Section III, we show how an online discussion forum may be analyzed by representing it as an affiliation network. In Section IV, we describe the relevance of clustering in machine learning and data mining. Section V deals with partition methods and explains the K-means clustering method. In Section VI, we present the application of the K-means clustering algorithm to the problem of identifying opinion leaders in a discussion forum. Section VII contains recommendations which can be followed for identifying leaders. Section VIII summarizes the experimental results and section IX presents conclusions.

II. SOCIAL NETWORK ANALYSIS

We describe a few basic concepts in social network analysis. The social network can be viewed as a graph of relationships and interactions among individuals who represent nodes in the graph. An idea that appears in the social network can either spread to include many nodes or it could die out quickly. Suppose we have information regarding how individuals influence others in the network and we would like to market a product by providing incentives to a few key individuals in the network. Viral marketing is based on the assumption that by selecting a few key individuals and targeting them for marketing the product, there will be a spreading or cascade of recommendations to buy the product. The fundamental problem raised in [9] is this. How do we select the initial set of individuals such that the spread of influence, defined as the number of nodes ultimately influenced to buy the product, is maximized?

Given the directed graph (V, E), where V is the set of nodes and E is the set of edges among the nodes of the network, the marketer chooses a set $S \subseteq V$ of nodes and makes them active. By this we mean that initially the set S of nodes is influenced by the product. Starting from the set S, called the *seed nodes*, the influence spreads when the seed nodes activate some of their neighbors active. These nodes in turn activate some of their neighbors and so on. Once a node turns active, we assume it remains active. A solution to the influence maximization problem aims to select the select the set S such that the number of nodes activated ultimately is maximized.

Diffusion Models: Regarding the spread of information in social networks, there is a well-known theory called the "the strength of weak ties." It is clear that the importance of stong ties is well understood. Often, the closest people (family, friends, colleagues, etc.) of a person have many overlapping contacts. They all interact with each other closely, and as a result, they all tend to have the *same* information on a variety of topics. Information that reaches any one of them is likely to reach all of them. However, they are less likely to be sources of new information from more distant parts of the network. As a result, any information received is likely to be "stale" information, which has already been received from someone else. What is, therefore, important to realize is that new and different information is likely to become available from relatively weak ties of less frequent contacts or “distant” contacts.

Granovetter and Schelling were among those who proposed the *Linear Threshold Model*, which is based on the concept of node-specific thresholds [10]. In this model, a node v is influenced as follows. A weight function $w: V \times V \rightarrow [0,1]$, such that $w(u,v) = 0$ if and only if $(u,v)$is not an edge in E, and further, $\sum_{u \in V} w(u,v) \leq 1$. Given the initial seed set S, the influence cascades as follows. Each node selects a threshold θ_v, uniformly at random in the interval $[0,1]$. Given the initial seed set S, and the thresholds θ_v, the influence cascades in steps $i = 0,1,2,...$, where at each step i a set S_i represents the nodes active, with $S_0 = S$. The set S_i consists of the nodes already in set $S_{i-1}$and those nodes v whose weighted number of its neighbors reaches its threshold θ_v, i.e. $\sum_{u \in S_{i-1}} w(u,v) \geq \theta v$. The value $IL(S)$ represents the expected number of nodes that are active at the end of the process. In the *Independent Cascade Model*, each node v can be influenced by its neighboring node u with a probability $p_{uv} \leq 1$. Once a node becomes active, it has one chance to activate its neighbors with the corresponding edge probabilities. As before, we start with the initial seed set S, whose nodes are active, and repeat the process of activating nodes until no more active nodes. The resulting set of nodes will be the influenced set of nodes. The value $I_C(S)$ represents the expected number of nodes that are active at the end of the process.

The Influence Maximization Problem: The problem is simply to find, for a parameter k, a k-node set S such that the value $I_L(S)$ (or the value $I_C(S)$) is maximized. For both the models considered above, it has been shown in [9] that the problem is NP-hard.

Basic metrics of a social network: In order to analyze connections and interactions among nodes of the network, a few metrics can be defined [11]. Often, such metrics are defined from a perspective of sociology, behavioral science, or psychology. The first metric is obviously *size*, the number of nodes in the network. This gives us an idea of how large the network is. *Inclusiveness* refers to the number of nodes that are connected to other nodes. In other words, inclusiveness can be expressed as the ratio of connected nodes to the total number of nodes, where the number of connected nodes refers to the total number of nodes minus the number of isolated nodes. *Density* is a measure that is expressed as a fraction of the maximum possible edges in a graph. If a graph has n nodes the maximum possible number of edges is $n * (n-1)$, assuming a directed graph. If the actual number of edges is l, then the ratio -

$l/(n * (n-1))$is the density.

Centrality is an indication of the social power of a node based on the degree to which it impacts the network. One way to measure node centrality is to consider the degree of nodes in the graph, where the degree is defined as the number of outgoing (or incoming) edges for a node. A node is called central if it has a high degree. The node is considered to be central because it is “well connected.” The degree based measure of node centrality can be extended to include paths of lengths greater than 1. However, it is found that determination of centrality based on path lengths greater than, say, 4 is not informative because at greater path lengths a large number of nodes in the graph become reachable. For example, there is a theory called “Six degrees of separation”, according to which everyone is six or fewer steps away, by way of introduction, from everyone else in the world. In fact, Facebook released results in November 2011 that show that among all users in the network, the average distance is 4.74 [12]. For these reasons, it turns out that only the path lengths of 1 or 2 are most informative regarding centrality. Since the degree of a node depends on the number of nodes in the graph itself, it could be misleading to compare degrees of nodes in graphs with differing total number of nodes. For example, a node with a degree 25 in a graph of 100 nodes is not as central as a node of degree 25 in a graph of 30 nodes. To overcome this problem, it has been suggested to measure the relative centrality. Thus, a node of degree 25 in a graph of 100 nodes has a relative centrality of 0.25, whereas a node of degree 25 in the graph of 30 nodes has a relative centrality of 0.86.

The degree is a measure of local centrality. *Global centrality*of nodes is measured by considering shortest distances among nodes. A simple measure of global centrality of a node can be computed from the “sum distance” which is simply the sum of all geodesic (shortest) distances to all other nodes in the graph. A node with a low sum distance is more globally central than one with a high sum distance. So closeness is viewed as the reciprocal of the sum distance. If the geodesic distances between nodes are represented as a matrix, then the sum distance (for undirected graphs) for a node is the row or column sum in the matrix.

The closeness measure as described could be misleading in large and complex networks. Consider two nodes A and B. A is close to a small and fairly closed group of nodes within a large network, but distant from other nodes. B is at a moderate distance from all nodes in the network. In this case, the closeness measure based on the sum of geodesic distances could be similar in magnitude for both A and B. However, B is really more central than A because B is able to reach more nodes in the network with the same effort. The *eigenvector* approach attempts to find a closeness measure based more on the “global” or overall structure of the network and less on local structure.

For a graph (V,E), let $\boldsymbol{A} = (a_{vw})$ be the adjacency matrix such that $a_{vw} = 0$, if there is no edge from node v to node w, or 1 if there is one. The eigenvector equation $\boldsymbol{Ax} = \lambda \boldsymbol{x}$when solved may give many different *eigenvalues* λ, in general. The greatest of such eigenvalues is chosen for the closeness measure. The component in the related eigenvector corresponding to the node v gives the desired closeness score for node v.

Another concept of node centrality is that of *betweenness*. This concept is used to measure the extent to which a particular node lies “between” other points. It is possible that a node of low degree may play an important “intermediary” role in connecting two or more connected parts of a graph. Such a node can then be considered to be central to the network. The betweenness of a node in the network is a measure of the extent to which an agent representing the node can play the part of a “broker” or “gatekeeper”. The betweenness proportion of a node Y for a pair of nodes X and Z is defined as the proportion of geodesics connecting X and Z that passes through Y. In other words, this measures the extent to which Y is between X and Z. The pair dependency of node X on node Y is defined as the sum of betweenness proportions of Y for all pairs of nodes that involve X. The overall betweenness of node Y is then calculated as the sum of the pair dependencies of all such nodes X on node Y. In mathematical terms, let g_{ij} denote the number of geodesics from node i to node j, and let g_{ikj} denote the number of geodesic paths from node i to node j that pass through node k. The betweenness measure of node k, b_k is calculated as:

$$b_k = \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}$$

Prestige is a more refined measure of a node's importance in networks where directional links are relevant. A node A that receives more in-links than another node B also has, intuitively speaking, more prestige. Again, a simple prestige measure can be based on in-degree of nodes. Considering the network size also as a factor, we can write: $P(v) = d_i(v)/(n-1)$, where $P(v)$ is the prestige of node v, $d_i(v)$ is the in-degree of node v, and n is the number of nodes in the network. Just as in the case of centrality measures described above, it is possible to extend this concept of prestige to include path lengths greater than 1. This involves two related concepts. The first is that of the influence domain of a node, which is defined as the set of nodes from which the given node can be reached. The second related concept is that of the average distance of such nodes from the given node. A *Proximity Prestige*$P_P(v)$ measure for node $v$can be defined by combining these two factors:

$$P_P(v) = \frac{I_w/(n-1)}{(\sum_{w \in V} d(w,v))/I_w}$$

Here, I_w is the number of nodes that can reach node v and $d(w,v)$ is the distance from node w to node v.

Rank Prestige: The rank prestige takes into consideration the prominence of the nodes that are within the influence domain of the node whose prestige is being measured. It is based on the following observation. If the influence domain of node v consists of prestigious nodes, then the rank prestige of node v should be high. On the other hand, if the influence domain contains mostly marginally important nodes then the rank prestige of node v should be low. Let $P_R(v)$ be the rank prestige of node v, and let $P_R(w)$ be the rank prestige of any other node in the network. Further, let $x_{wv} = 1$, if node w is one of the nodes that vote or choose node v, and 0 otherwise. Then we can write:

$$P_R(v) = \sum_{w \in V} x_{wv} * P_R(w)$$

It is easy to see that there would be one such equation for each node in the network. Since there are n nodes in the network, there would be n equations with n unknowns to be solved.

Structural balance: A group of nodes (people) is structurally balanced if, when two nodes "like" each other they are consistent in their evaluation of all other nodes. Here, an evaluation of another node can be either positive or negative. A cycle in the graph where edges are labeled either positive or negative is said to have a positive sign if the number of negative signs in the cycle is even. Else, the cycle is said to have a negative sign. An important definition is the following. A graph is said to be balanced if and only if all of its cycles have positive signs. An important result that has been proved by Harary [13] is that if a signed graph is balanced, then the nodes of the graph can be partitioned into two subsets such that only positive lines join nodes within a subset and only negative lines join nodes between subsets. Empirical studies [14] have shown that the number of clusters or partitions is often more than two.

Clusterability:A signed graph is clusterable if its nodes can be partitioned into a finite number of subsets such that each positive line joins two nodes in the same subset and each negative line joins two nodes in different subsets. The subsets are called*clusters*.

Some results, called the clustering theorems, have been proved [14]. A signed graph can have a clustering if and only if it contains no cycle with exactly one negative line. Another result shows that the following four statements are equivalent. 1) The graph is clusterable. 2) The graph has a unique clustering. 3) The graph has no cycle with exactly one negative line. 4) The graph has no cycle of length 3 with exactly one negative line.

In *ranked clusters*, nodes in a lower cluster should have positive ties to nodes in a higher ranked cluster and negative ties to nodes in lower ranked clusters. Research has shown that *transitivity* is a very important structural property in social network data [14]. The ideas of partially ordered clusters and generalized rank clustering naturally lead to transitivity. We discuss transitivity briefly next.

Although a social network can be studied statically as a snapshot of nodes and edges at a particular point in time, it is useful to also study how a network evolves over time. One of the most basic principles is that of *triadic closure* which is stated as: If two people have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future. For example, if nodes A and B are connected and so are nodes B and C, then it is likely that at some point in the future, nodes A and C will become connected. One reason why triadic closure operates, apart from the obvious one that it is intuitively natural, is that there is an incentive for B to bring A and C together. There is latent stress in the relationships if A and C are not connected with each other [15]. It is possible to formulate the triadic closure property taking into consideration the strengths of the ties represented by the edges between nodes. An edge could represent a *strong tie*, as between close friends, or a *weak tie*, as between two acquaintances. The Strong Triadic Closure Property requires that an edge exist between A and C whenever strong ties exist between A and B and between B and C. In this case, node B is said to exhibit the Strong Triadic Closure Property.

The *Clustering Coefficient* of a node A is defined as the probability that two randomly selected neighbors of A will themselves share an edge with each other. The clustering coefficient of a node ranges from 0 to 1.

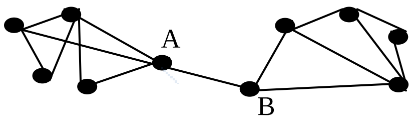

Figure 1: Bridge

In the example shown in Figure 1, nodes A and B are connected to tightly knit groups of neighbors. An edge joining two nodes A and B is a *bridge* if deleting the edge would make A and B lie in two separate components of the network. This would happen if the edge between A and B is the only route between the two end points. We could expect that B plays a role different from A's tightly knit neighbors. This is because A'stightly knit neighbors would be exposed to similar opinions and similar sources of information. Node A's link to B will offer A access to information that are not ordinarily available. However, such edges are not commonly found in real social networks. What is much more commonly found is an instance where removing the edge between A and B would increase the path length between A and B to more than 2. This means there are no common nodes C between A and B, such that there is an edge between A and C and an edge between C and B. In this case, we call the edge between A and B a *local bridge*. The *span* of a local bridge is the distance between its end points if the edge were deleted. Local bridges provide their end points access to parts of the network and sources of information which are otherwise remotely situated. It is possible show in a straightforward way the correctness of the following claim. If a node B in a network satisfies the Strong Triadic Closure Property and is involved in at least 2 strong ties, then any local bridge that involves B must be a weak tie.

Another useful concept from graph theory is that of a *clique*. A clique is a sub-set of the nodes in which every pair of nodes is directly connected by an edge. A *component* is a set of nodes where each pair of nodes is connected by a path. While a component is maximal and connected, a clique is maximal and complete. It turns out that the concept of a maximal complete sub-graph is of limited use in social networks because it is rare to find such tightly knit groups. Therefore, a few extensions of the clique concept are found in the literature. One extension is that of an *n-clique*, a subset of the nodes in the graph such that each pair of nodes is connected by a path not exceeding *n*. Thus, a 1-clique is a maximal complete sub-graph, while a 2-clique is a maximal connected sub-graph where each pair of nodes is connected by a path with length not greater than 2. Here each node is connected to every other node either directly or through an intermediate node. N-cliques can be identified by multiplying the adjacency matrix with itself. For example, the square of the adjacency matrix shows all distance 2 connections, the cube of the adjacency matrix shows all distance 3 connections, and so on. An increase in the value of *n* implies a relaxation in the definition of the clique. It turns out that values of *n* greater than 2 can be difficult to interpret sociologically. As mentioned above, having a path length of 4 or more is not informative. Therefore, it is appropriate to only identify n-cliques with values of n either 1 or 2.

III. REPRESENTING A DISCUSSION FORUM

We now turn to the problem of representing an online discussion forum. The usual kind of social network is not suitable for this task because it presumes an existence of relationships among people. Often, users join a network based on interests they share with the discussion forum's focus. A user in the forum may or may not have relationships with the members of the forum, but yet may participate in discussions. A discussion forum can be modeled as a two-mode social network, often called an *affiliation network*[14]. An affiliation networks describes a set of *actors* or agents$V$, and a set of *events* E, rather than just ties between actors. Connections among members of an affiliation network are based on participation in events. In the case of a discussion forum, apart from the set of members, we also have a set of *discussion topics or threads* which represent events in the affiliation network. Each member may or may not participate in a discussion topic. A bipartite graph may be used to represent an affiliation network. Figure 2 is an example where $V = \{a_1, a_2, a_3, a_4\}$ and $E = \{t_1, t_2, t_3, t_4, t_5\}$ and the edges are as shown.

Mathematically, an affiliation network can be represented by an affiliation matrix that has *n* rows and *m* columns, *n* and *m* being the number of actors and the number of events respectively. So we can write:

$\boldsymbol{A} = \{a_{ij}\}$,where

$$a_{ij} = \begin{cases} 1, & if\ actor\ i\ is\ affiiliated\ with\ event\ j \\ 0, & otherwise \end{cases}$$

The sum of elements in any row *i* is the number of events that actor *i* participates in and the sum of elements in any column *j* is the number of actors that participate in event *j*. If actors *i* and *j* are both affiliated with event *k*, then a_{ik} and a_{jk} will both be 1. We can express a relation x_{ij}^N to stand for the number of events with which both actors *i*and *j* are affiliated. Formally, we can write $x_{ij}^N = \sum_{k=1}^{m} a_{ik} * a_{jk}$. The co-membership frequencies can be summarized in a $n \times n$ matrix, $\boldsymbol{X^N} = \{x_{ij}^N\}$. The relationship between the sociomatrix for co-memberships $\boldsymbol{X^N}$and the affiliation matrix $\boldsymbol{A}$can be expressed concisely as :

$\boldsymbol{X^N} = \boldsymbol{AA^T},$ where $\boldsymbol{A^T}$ is the transpose of $\boldsymbol{A}$.

It is important to note that the diagonal elements of the co-membership matrix $\boldsymbol{X^N}$ stand for the number of events that every actor is affiliated with. Reasoning in a similar fashion, we can find the number of actors affiliated with both events *k* and *l* as:

$x_{kl}^M = \sum_{i=1}^{n} a_{ik} * a_{il}$.

The $m \times m$ sociomatrix $\boldsymbol{X^M} = \{x_{kl}^M\}$ computes the number of actors that are common between any two events, using the affiliation matrix $\boldsymbol{A}$:

$$\boldsymbol{X^M} = \boldsymbol{A^TA}$$

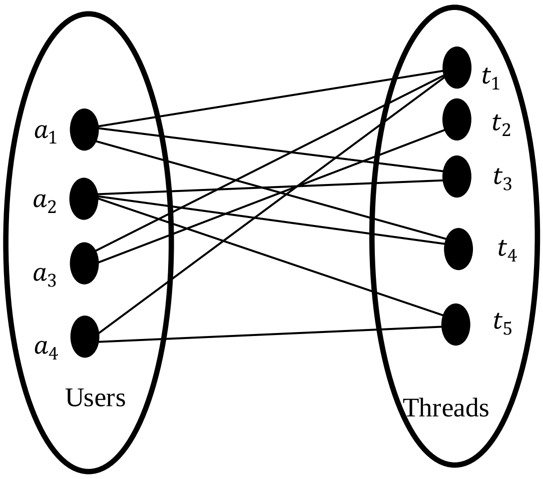

Figure 2: Bipartite Graph representation of a discussion forum

The number of events that an actor i is affiliated with is given by $a_{i+} = \sum_{j=1}^{m} a_{ij} = x_{ii}^{N}$. The number of actors affiliated with an event j is given by $a_{+j} = \sum_{i=1}^{n} a_{ij} = x_{jj}^{M}$.

The density of the one-mode network among actors that is derived from the affiliation network is an indicator of the mean number of events to which pairs of actors belong. This factor, denoted by $\Delta_N$is given by:

$$\Delta_N = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^{N}}{n(n-1)}$$

The values of Δ_N range from 0 to m.

The overlap measure for events is the expressed as the density Δ_M. It is given by:

$$\Delta_M = \frac{\sum_{k=1}^{m}\sum_{l=1}^{m} x_{kl}^{M}}{m(m-1)}$$

*Cohesive subsets of actors and events:*Considering the co-membership relation among actors, a clique at level c is a sub-graph in which all pairs of actors share memberships in at least c events. For the overlap relation among events, a clique at level c is a sub-graph in which all pairs of events share at least c actors.

Centrality: Centrality measures have been defined for ordinary social networks. However, these measures will have to be re-interpreted for affiliation networks and adjusted accordingly, considering affiliation networks as bipartite graphs. The degree measure of centrality in ordinary networks is simply d_i, the number of edges incident on node i. It could be normalized by dividing by $(n-1)$. In the affiliation network case, node i could be a node from the set of actors or it could be from the set of events. The maximum number of edges for a node is always the size of the other set. Thus, we have:

$d_i^* = d_i/(m-1)$ for $i \in V$, and $d_i^* = d_i/(n-1)$ for $i \in E$.

Closeness: For ordinary networks, the reciprocal of the sum of geodesic distances of a given node i, c_i is a measure of closeness. As usual, this can be normalized by considering the term $(n-1)/c_i$ as the closeness measure. In the case of affiliation networks, the underlying graph is bipartite and adjustments must be made. The closest a node i in the set of actors (events) can be from nodes in the same set is a distance of 2 and from nodes in the set of events (actors) a distance of 1. Accordingly, we have:

$c_i^* = \frac{m+2(n-1)}{c_i}$for a node $i \in V$, and

$c_i^* = \frac{n+2(m-1)}{c_i}$for a node $i \in E$.

Remarks on Structural Properties: So far, we have discussed social networks, affiliation networks in particular,by considering their structural properties. It is important to realize that examining the structure of a network can yield useful information but this information may not be sufficient in all cases. For example, structural analysis may identify actors in an affiliation network that are most frequently affiliated with events. But this information does not tell us how effective these actors are and how their opinions are valued by other actors.

Consider the users in an online discussion forum. Key users in a discussion forum are not only likely to participate in more discussion topics than others, but they also tend to receive positive responses from other users. If a user's participation in a large number of discussion topics is considered, based on the representation of the discussion forum as an affiliation network, it may not give us a complete picture of the importance of the user in the forum. For example, even spammers are known to post a large number of messages.

In order to gain insight into attributes other than those based strictly on the structure of the network, it becomes necessary to examine various attributes of actors using data mining techniques. Data mining is used in two ways in this approach. First, using data mining techniques,it is possible to establish a user profile for each user in the discussion forum. Next, using the method clustering, users can be organized into different clusters, such that users in the same cluster have similar attributes. The cluster containing opinion leaders can then be chosen for the purpose of spreading new information in the forum. This will be topic of discussion in following sections.

IV. Clustering

Clustering is a class of unsupervised learning models [16], often used in machine learning and data mining,which make use of notions of distance and similarity between observations. The purpose of clustering methods is to identify homogeneous groups of observations called *clusters*. Observations in the same cluster are close or similar to each other and far from or dissimilar to observations in other clusters.

Given a dataset $\mathcal{D}$, we can represent the *m* observations by means of *n*-dimensional vectors of attributes, so that the dataset is represented by a matrix **X**, with *m* rows and *n*columns.

D = [d_{ij}] be the symmetric m x m matrix of distances between pairs of observations. Here, d_{ij} denotes the distance dist($\mathbf{x_i}$,$\mathbf{x_j}$) between observations $\mathbf{x_i}$ and $\mathbf{x_j}$.

It is possible to transform the distance d_{ij} between two observations into a similarity measure s_{ij}, by using

$s_{ij} = 1/(1+d_{ij})$ or $s_{ij} = (d_{max} - d_{ij})/d_{max}$, where d_{max}denotes the maximum distance between any two observations in the dataset $\mathcal{D}$.

The distance d_{ij} can be the Euclidean distance, the Manhattan distance, or the arccosine distance.

V. partition Methods

Given the dataset Ɗ of m observations, where each observation is represented by a vector in the n-dimensional space, partition methods construct a subdivision of Ɗ into a collection of nonempty subsets C = {C_1,C_2,…,C_K}, where K≤m. The number K of clusters is predetermined and assigned as an input to clustering algorithms. The clusters generated are mutually disjoint in the sense that each observation belongs to one and only one cluster.

Partition methods begin with an initial assignment of the *m* observations to the *K* clusters. A reallocation technique is iteratively applied to place some observations in different clusters in such a manner as to improve the quality of the overall partitioning. The partitioning algorithm stops when the no reallocation happens during an iteration.

The K-means method [17] is one of the best known clustering algorithms. It is an efficient clustering method that effectively produces clusters of spherical shape in the *n*-dimensional space.

The K-means clustering algorithm begins by choosing K observations arbitrarily as the centroids of the clusters. During each iteration, each observation is assigned to the cluster containing the centroid that is most similar to the observation. Here, the most similar centroid is the one whose distance from the observation is the minimum. If no observation is assigned to a new cluster in the current iteration, the algorithm stops. At the end of each iteration, the new centroid of each cluster is computed as the mean of the observations in that cluster.

The first step is to initialize the K clusters with cluster representatives. These cluster representatives will be the initial centroids. At the end of each iteration, the centroid is recomputed and a new representative is found.

VI. Application of K-means algorithm

We now consider the application of the K-means clustering algorithm to the problem of identifying opinion leaders in an online discussion forum. Typically, the discussion forum consists of a number of users and each user has the choice of participating in a discussion topic, also called a discussion thread. Each user can respond to a specific message from another user in a thread or start a new thread. It is not necessarily the case that the user who starts a discussion thread must the main participant in that thread. It is also not necessarily the case that opinion leaders post messages more frequently in the discussion forum. In fact, a user who posts messages too frequently could also be a spammer. Let *U* be the set of all users of the online discussion forum. So we can write $U= \{u_1,u_2,...,u_L\}$. It is necessary to define the attributes of each user that will be used to perform clustering. The attributes must be selected in such a way as to enable the clustering algorithm to distinguish among the various groups of users efficiently. Clustering is done based on the premise that we can identify four groups of users – leaders, intermediate users, newbies, and spammers. The leaders group is the group of interest for marketing purposes.

By studying the dynamics of Internet based discussion forums, it is possible to make a few observations. Opinion leaders tend to spend a significant amount of time in the online discussion forum. They post messages at least a few times a month. As noted above, just the frequency of messages posted does not automatically qualify a user to be an opinion leader. More often than not,the messages of opinion leaders are responded to by others. Their messages are often met with positive feedback from others, with minimal negative response. As an opinion leader gains popularity and prestige in the discussion forum, messages from other users are likely to contain references to the opinion leaders or their messages. Opinion leaders are likely to write fairly detailed messages rather than one-liners. In some discussion forums messages from most users tend to be short. In this case,the messages of opinion leaders may also be short, but these messages will likely contain links to more detailed explanations, such as those contained, for example, in blogs. They get involved in a discussion thread in a significant manner.

These observations lead to the representation of the online profile of each user as a set of attributes based on the online behavior of a user and that of other users as well.

We have chosen to use 8 attributes as defined below:

- ot_i – the time that user u_i spends online
- ft_i – the frequency with which user u_i posts messges in the forum in a given time period, for example a week or month
- fr_i – the degree to which the messages of user u_i are responded to by other users
- pr_i – the degree to which the messages of user $u_i$have positive feedback from other users
- nr_i– the degree to which the messages of user $u_i$have negative feedback from other users
- rr_i– the degree to which user u_i is referred to in messages of other users
- ms_i – the average size of messages sent by user u_i
- it_i – the degree of involvement of user u_i in a discussion thread

Some of the attributes defined above can be evaluated by collecting statistics regarding each user's usage of the discussion forum. However, other attributes of a user, such as pr_i and nr_i, can be found by analyzing the responses of other users. This analysis typically involves implementing a supervised machine learning algorithmthat is able to learn and classify the nature of responses or the *polarity* as positive, negative, or neutral. The automated recognition of positive, negative, and neutral responses is accomplished by the machine learning algorithm.

Classification models are supervised learning methods and are often used for predicting the value of a categorical target attribute [16]. For example, given a set of symptoms for a patient, a classifier predicts the disease that the patient is most likely suffering from. Starting from a set of past observations whose target class is known, classification models are used to generate a scheme by which the target class of future examples can be predicted.

Classification is an important topic in learning theory due to its theoretical implications and the large number of domains where it can be successfully applied. The development of algorithms capable of learning from past experience represents a fundamental step towards emulating the inductive capabilities of humans.

The opportunities presented by classification extends into many different application domains: the selection of target customers for a marketing campaign, fraud detection, image recognition, early diagnosis of diseases, text cataloguing, and spam email detection are just a few examples of real world problems that can be formulated in terms of the classification paradigm.

A. *Classification Models*

In a classification problem, there is a dataset $\mathcal{D}$ consisting of m observations described in terms of n explanatory attributes and a target categorical attribute. The explanatory attributes are also called*predictive variables*. The target attribute is also called a *class* or *label*, and the observations are also termed *examples* or *instances*. The target variable for classification models takes a finite number of values. In particular, the case where the observations belong to one of only two classes is called a *binary* classification problem. The purpose of a classification model is to identify recurring relationships among observations which describe the examples belonging to the same class. These relationships are then converted into classification rules which can be used to predict the class for observations for which only the values of explanatory variables are known. The rules may be of different forms depending on the type of classification model used.

From a mathematical perspective, in a classification problem, *m* known examples are given, consisting of pairs of $(\mathbf{x_i}, y_i)$, i Є *M*, where $\mathbf{x_i}$ is the vector of values taken by the n predictive variables for the i^{th} example and y_i Є *H* = $\{v_1, v_2, \ldots, v_H\}$ denotes the target class. Each component x_{ij} of the vector $\mathbf{x_i}$ is treated as a realization of the random variable X_j representing an attribute $\mathbf{a_j}$ in the dataset $\mathcal{D}$. In a binary classification problem, *H* may be denoted by *H* = {-1,1} without loss of generality. Here 1 may stand for a positive response while -1 may stand for a negative response.

Let $\mathcal{F}$ be a class of functions $f(\mathbf{x_i}) : \mathbb{R}^n \mapsto \mathcal{H}$ called the *hypotheses* that represent possible relationships between y_i and $\mathbf{x_i}$. A *classification problem* consists of defining an appropriate hypothesis space $\mathcal{F}$and an algorithm A_F such that A_F identifies a function f^* Є $\mathcal{F}$that can optimally describe the relationship between the predictive variables and the target class.

There are three components of a classification problem: a *generator* of observations, a *supervisor* of the examples according to an unknown probability distribution $P_{\mathbf{X}}(\mathbf{x})$. For each vector x of examples, the supervisor returns the value of the target class according to a conditional distribution $P_{Y|X}(y|\mathbf{x})$ which is also unknown. A classification algorithm A_F, also called the *classifier* selects a function f* Є $\mathcal{F}$ in the hypothesis space that minimizes a loss function.

The development of a classification model consists of three main phases.

Training phase: During the training phase, the classification algorithm is applied to the examples of the *training set*, which is a subset of the dataset $\mathcal{D}$. This subset consists of observations for which the target class is already known. This allows the classifier to derive classification rules that establish the correspondence between the target class y and each observation **x**.

Test phase: During the test phase, the rules generated in the training phase are used to classify observations of the dataset $\mathcal{D}$ that are not included in the training set and for which the target class is already known. The accuracy of the classification model is assessed by comparing the predicted class for each example in the test set with the actual class of the example. The training set and test set must be disjoint to avoid an overestimate of the model accuracy.

Prediction phase: In the prediction phase, the classifier is used to predict the class of an observation for which the target class is not known. This phase thus represents the use of the classification model to assign the target class to new observations in the future.

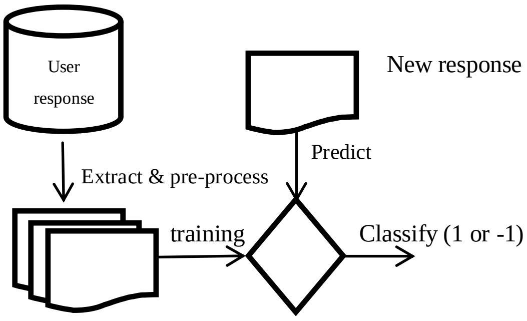

Figure 3: Various steps in predicting response polarity

Figure 3 shows the logical flow of the learning process for a classification algorithm. Classification models may be divided into four broad categories. *Heuristic models* make use of classification algorithms that are simple and intuitive. These include *nearest neighbor* methods and *classification trees*. *Separation models* divide the attribute space $\mathbb{R}^n$ into *H* disjoint regions, $\{S_1, S_2, \ldots, S_H\}$, separating the observations based on the target class, so that observations in region S_H are assigned to class $y_i = v_H$. In general, it is difficult to divide the observations exactly into a set of simple regions. Hence, a loss function is defined to take into account the points that are not classified correctly, and an optimization problem is solved to arrive at a subdivision into regions that minimize the loss. *Discriminant analysis, perceptron methods, neural networks,* and *support vector machines* are some of the most popular separation methods. *Regression models* make an explicit assumption regarding the functional form of the conditional probabilities $P_{y|\mathbf{x}}(y|\mathbf{x})$which correspond to the assignment of the target class by the supervisor. *Linear regression* assumes a linear relationship exists between the dependent variable and the predictors. *Logistic regression* is an extension of linear regression to handle binary classification problems. In probabilistic models, a hypothesis is formulated regarding the functional form of the conditional probabilities $P_{\mathbf{x}|y}(\mathbf{x}|y)$ of the observations, given the target class, known as *class-conditional probabilities*. Next, based on the estimate of the *prior probabilities*$P_y(y)$ and Bayes' Theorem, the *posterior probabilities* of the target class $P_{y|\mathbf{x}}(y|\mathbf{x})$ can be calculated. *Naive Bayes* and Bayesian Networks are popular families of probabilistic methods.

B. Naive Bayes Classification Model

The Bayesian model calculates , the *posterior probability* of a specific target class $P_{y|\mathbf{x}}(y|\mathbf{x})$, given an observation **x**, by means of Bayes' Theorem, using the *prior probability* of class y, P(y) and the *conditional probabilities* $P(\mathbf{x}|y)$, which are computed in the training phase. Consider an observation **x** whose class variable y may take *H* distinct values, $\{v_1, v_2, \ldots, v_H\}$. We can use Bayes' Theorem to calculate the posterior probability P(y|**x**), the probability that the observation **x** belongs to class y:

$$\mathrm{P}(y|\mathbf{x}) = \frac{\mathrm{P}(\mathbf{x}|y)\mathrm{P}(y)}{\sum_{i=1}^{H}\mathrm{P}(\mathbf{x}|y)\mathrm{P}(y)}$$

$$= \mathrm{P}(\mathbf{x}|y)\mathrm{P}(y)/\mathrm{P}(\mathbf{x})$$

To classify an observation **x**, the Bayes' classifier applies the principle of *maximum a posteriori hypothesis* (MAP), which involves calculating the posterior probability P(y|**x**) for all classes y and assigning the observation **x** to the class which has the maximum value P(y|**x**). The prior probabilities P(y) can be estimated using the frequencies $m_h$with which each class appears in the dataset. $\mathrm{P}(y) = m_h/m$.

The sample estimate of the conditional probabilities P(**x**|y) cannot be obtained in practice due to the computational complexity and the huge number of sample observations that it would require. To overcome this difficulty, we use the *Naive Bayes classifier* which we describe below.

Naive Bayes classifiers are based on the assumption that the explanatory variables in the observation **x** are all conditionally independent for a given target class. This assumption allows us to express P(**x**|y) as:

$$\mathrm{P}(\mathbf{x}|y) = \mathrm{P}(x_1|y) * \mathrm{P}(x_2|y) * \ldots * \mathrm{P}(x_n|y) = \prod_{j=1}^{n}\mathrm{P}(x_j|y)$$

The probabilities P(x_j|y) can be estimated using the examples from the training set. P(x_j=v|y) is calculated as the ratio of the number of instances of class y for which the attribute x_j takes the value v to the total number of instances of the class y in the dataset.

Empirical comparisons showing the effectiveness of the Naive Bayes method are found in [18] and a comparative assessment is found in [19].

For the problem under consideration, i.e. prediction of polarity of responses from other users, each term that appears in the text of the response is potentially a dimension in the set of attributes. However, not all terms in the text of the response will represent an attribute. We describe next the preprocessing step where only those terms that are meaningful to be a dimension are extracted from the text.

In the first phase, features of the text are extracted. In the second phase, a learning algorithm is used to identify the polarity of the response.

The first phase is typically simplified by pre-processing each response. Since each term appearing in the response can be considered to be an additional dimension, the textual data can be of very high dimensionality. The pre-processing step partly overcomes this by reducing the number of considered terms. Pre-processing consists of three tasks as described below.

Tokenization: This process consists of dividing a large textual string into a set of tokens where a single token corresponds to a single term. This step also involves filtering out all meaningless symbols like punctuations and commas, since these symbols do not contribute to the classification task. Also, all capitalized characters are converted to lower-case.

Stop-words removal: Natural languages commonly make use of constructive terms like conjunctions, adverbs, prepositions and other language structures to build up sentences. Terms like "the", "in" and "that", also known as stop-words do not carry much specific information in the context of a response. These terms appear frequently in the descriptions of the responses and thus increase the dimensionality of the data which in turn could decrease the accuracy of classification algorithms. This is also calledthe curse of dimensionality. Therefore, it is necessary to remove all stop-words from the set of tokens, based on a list of known stop-words.

Stemming: The stemming step aims at reducing each term appearing in the descriptions into its basic form. Each single term can be expressed in different forms but still carry the same specificinformation. Forexample, the terms "computerized", "computerize" and "computation" all have the same morphological base: "computer". A stemming algorithm such as the Porter stemmer [20] transforms each term to its basic form.

In the second phase, a learning algorithm such as the Naive Bayes classifier,is applied to a training set of responses which have been pre-processed for which the polarity is known. The training set of responses is used to train the Naive Bayes classifier. Once training has completed, the classifier will be able to predict the polarity of new responses.

To facilitate the implementation, each of the eight attributes is discretized and mapped to a scale of 1 to 5. The distance between two users, uand v, with attribute vectors **x** and **y**,is given by:

$\mathrm{dist}(u,v) = (\sum^{8}_{i=1} (x_i - y_i)^2)^{1/2}$, where x_i and $y_i$are the ith attributes of the u and v respectively.

It can be shown that the K-means clustering algorithm minimizes the following cost function:

$\mathrm{Cost} = \sum^{K}_{h=1}\sum_{xi \varepsilon Ch}(\mathrm{dist}(x_{i,}w_h))^2$, where w_h is the centroid of the cluster h.

As initial cluster representative for the leaders we choose the following: (5 5 5 5 1 5 5 5), which represents the ideal leader scoring highest in all attributes except the 5th (negative responses from other users) where the score is the lowest.

Next, we discuss how a company may gather the information regarding the attributes for each user in a discussion forum.

VII. RECOMMENDATIONS FOR IDENTIFYING LEADERS

Any company that is interested in identifying opinion leaders for viral marketing campaigns and other purposes can follow a few steps as outlined below.

The audience group should be identified in terms of demographic characteristics, e.g. women below the age of 45, behaviors, e.g. people who manage their own finances, or attitudes, e.g. people who think technology unites families.

The characteristics of opinion leaders and the classification of the target audience should be formulated. From this information, a list of questions that people who register for a discussion forum should answer can be generated. Alternatively, a survey to be taken by people participating in the discussion forum can be hosted online or the various statistics for each user mentioned in Section IV above can be collected by analyzing data.

Identification of the opinion leaders can be done by analysis of the survey questions or by the analysis of the online data as described in Section IV. Once the opinion leaders have been identified, the company can start communicating with them. Newsletters, white papers on new products and industry trends can be sent to them. The leaders can also be given opportunities to participate in beta testing programs and their feedback can be sought.

The leaders can be tracked whenever they visit the company website. They should be provided any information they are looking for and their questions should be answered promptly. Feedback received from them can help shape future directions of a product.

VIII. EXPERIMENTAL RESULTS

A study of discussion forums with varying numbers of participants was conducted. For each type of forum, a simulation of the types of users likely to participate in discussions was run. For purposes of detecting the polarity of responses of other users to messages from a user, the Naive Bayes classifier as described in Section IV was used. Using the eight attributes, an online profile of each user was constructed. The K-means clustering algorithm was applied to the simulated data in order to identify the opinion leaders.

Consumer product discussion forums are important for companies because they can be a source of valuable feedback. Travel discussion forums help people plan and make the most of their trip. Technology forums are a source of valuable information regarding technology to people and companies. Healthcare discussion forums, for example patient forums,provide both medical information and support. Entertainment forums provide information regarding movies, TV shows, etc.

The results are summarized in the table(Table I) and depicted in the chart (Figure 4) below:

TABLE I. OPINION LEADERS

Discussion Forum	% of users who are Opinion Leaders		
	Number of users	*Leaders*	*Percentage*
Consumer product	450	48	10.66%
Travel	600	68	11.33%
Technology	800	102	12.75%
Healthcare	1500	207	13.8%
Entertainment	2500	658	26.32%

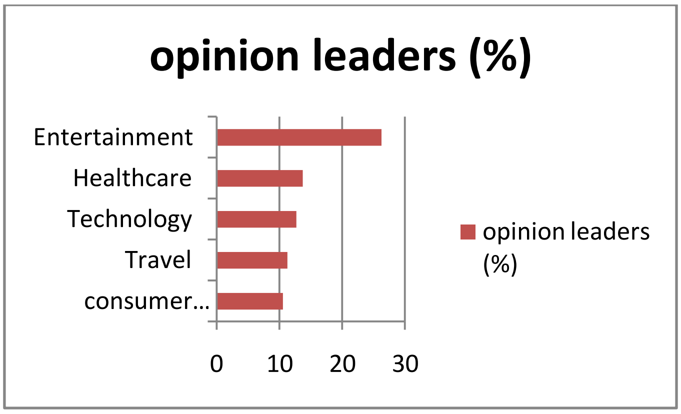

Figure 4: Opinion leaders in various forums

Once the opinion leaders are identified, they will constitute a high-priority target group for viral marketing campaigns, as mentioned in Section I. The orientation of opinion leaders with positive opinions can be used to spread specific information regarding a product.

IX. CONCLUSION

This approach presented in this paper is to identify opinion leaders in an online discussion forum, where each user's membership in other forums and his/her opinions are unknown. It is important for a company to know the opinions of users regarding its products. A manual approach is not scalable when large internet communities and social networks are considered. The opinion leaders play an important role in dissemination of online information. They influence the opinion of others. The study of affiliation networks, a special kind of social network, is useful in revealing the structural properties such as centrality. Other attributes of users in the forum may be found by mining the data available in the form of discussions on various topics.

Using clustering techniques from data mining, opinion leaders may be identified. Some attributes of users can be found in usage statistics of the discussion forum.However, other attributes of a user, such as positive and negative responses to a particular user, can be found by analyzing the responses of other users. This analysis typically involves implementing a supervised machine learning algorithm that is able to learn and classify the nature of responses Knowing the opinion leaders and their opinions, a company will be able to assess the chances and risks of its products. Appropriate measures can be taken to counteract negative opinions, such as product improvements and more effective marketing. Future work will focus on comparing the effectiveness of various clustering algorithms and detecting opinion trends. This can be explored by social network analysis. By analyzing the social network with its opinion leaders, opinion trends may be detected.It is possible more than one opinion trend about a product may exist in the network. Some members may have a positive opinion; some may have a negative opinion, while others may have a neutral or no opinion. It is then important to find out if the positive trend is stronger than the others. As members of the discussion forum interact with each other, they will likely develop social ties with each other as well. It is interesting to study the impact of such social ties on opinion trends.

REFERENCES

[1] I. Cakim, “Online Opinion Leaders: a predictive guide for viral marketing campaigns,” in Connected Marketing The Viral, Buzz and Word of Mouth Revolution, J. Kirby and P. Marsden, Eds, Oxford: Elsevier, 2006, pp107-118.

[2] V. Kumar and R. Mirchandani, “Increasing the ROI of Social Media Marketing”, MIT Sloan Management Review, Vol. 54, No. 1, Fall 2012.

[3] E. M. Rogers, Diffusion of Innovations, The Free Press: New York, 1995.

[4] X. Song, Y. Chi, K. Hino, and B. Tseng, “Identifying opinion leaders in the blogosphere”, Proceedings of the 16th ACM Conference on Information and Knowledge Management, Lisbon, Portugal, 2007, pp 971-974.

[5] T. Steiner, R. Verborgh, J. Gabarro, R. Van de Walle, “Adding Meaning to Social Network Microposts via Multiple Named Entity Disambiguation APIs and Tracking Their Data Provenance’, International Journal of Computer Information Systems and Industrial Management Applications, Vol. (5), 2013, pp. 069-078.

[6] F. Bodendorf and C. Kaiser, “Detecting Opinion Leaders and Trends in Online Communities”, 2010 Fourth International Conference on Digital Society, St. Maarten, Netherlands, pp124-129.

[7] Z. Zhai, H. Xu, and P. Jia, “Identifying Opinion Leaders in BBS”, IEEE International Conference on Web Intelligence and Intelligent Agent Technology, 2008, Sydney, Australia, pp398-401.

[8] H. Kadima and M. Malek, “ Toward Ontology-based personalization of a Recommender System in a social network”, International Journal of Computer Information Systems and Industrial Management Applications, Vol. (5), 2013, pp. 499-508.

[9] D. Kempe, J. M. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network”, KDD 2003, pp 137-146.

[10] M. Granovetter, “Threshold models of collective behavior”, American Journal of Sociology, 83(6), 1420-1443, 1978.

[11] J. Scott, Social Network Analysis, A Handbook, SAGE Publications, 2nd Edition, 2000.

[12] E. Barnett, “Facebook cuts six degrees of separation to four”, Telegraph, November 2011.

[13] F. Harary, “On local balance and N-balance in signed graphs”, Michigan Mathematics Journal, 3, 37-41, 1955.

[14] S. Wasserman and K. Faust, “Social Network Analysis, Methods and Applications”, Cambridge University Press, 1994.

[15] D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World,Cambridge University Press, 2010.

[16] C. Versellis, Business Intelligence: Data Mining and Optimization For Decision Making, Chichester, UK: John Wiley & Sons Ltd., 2009.

[17] X. Wu and V. Kumar, Eds, The Top Ten Algorithms in Data Mining, Boca Raton, Florida: Chapman and Hall/CRC, 2009.

[18] P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss”, *Machine Learning*, 29, pp. 103–130, 1997.

[19] A. Jamain and D. J. Hand, “Mining supervised classification performance studies: A meta-analytic investigation”, *Journal of Classification*, 25, pp. 87–112, 2008.

[20] M. Porter, “An algorithm for suffix stripping,” *Program*,vol. 14, no. 3, pp. 130–137, 1980.

Author Biographies

Shrihari Hudli earned his Bachelor of Engineering degree in Computer Science from the MS Ramaiah Institute of Technology, Bangalore, India in 2012. His research interests lie in the areas of data mining, social networks, software engineering, and parallel computation. He has published several papers in these areas. He has worked on research problems with the faculty of the Indian Institute of Science. He is currently working as a marketing and sales executive in ObjectOrb Technologies, India.

Aditi Hudli is an undergraduate student in the Computer Science Department of the MS Ramiah Institute of Technology, Bangalore, India. She has worked on projects in the areas of data mining, social networks, and software engineering. She has published several papers in these areas.

Anand Hudli, Executive Director and COO of ObjectOrb Technologies received his PhD degree in Computer Science from the Univerity of Nebraska, Lincoln in 1989, MTech degree in Computer Science from the Indian Institute of Technology, Bombay in 1985, and the Bachelor of Engineering degree in Electronics from the University of Mysore in 1982. He taught at the Purdue University, Indianapolis, for several years and later worked in various large companies, including Ameritech, Boehringer-Mannheim, Dow Jones, and IBM.At IBM, he worked on the high performance communication subsystem for the SP2 architecture. He has published papers in several areas, including artificial intelligence, data mining, social networks, parallel and distributed computing, and software engineering. He has been awardedone US patent.