Scalable Learning for Identify and Ranking Prevalent News Topic using Social Media Factor

S. Savitha[1], K. Logeswaran[2],
[1]Assistant Professor, Department of CSE,
K.S.R. College of Engineering, Tiruchengode.,India
[2]Assistant Professor, Department of IT,
Kongu Engineering College, Perundurai.,India

N.Sowmiya[3],S.Suryaprakash[4],M.Tamilselvan[5],
K.Tamilarasu[6]
[3,4,5,6] U G Students, Department of CSE,
K.S.R. College of Engineering, Tiruchengode.

***Abstract*— News media presents professionally verified occurrences or events, whereas social media presents the interests of the audience in these areas, and should therefore give insight into their quality. Social media services like Twitter can also provide additional or supporting information to a particular news media topic. Meanwhile, truly valuable information may be thought of as the area in which these two media sources topically intersect with each other. Unfortunately, even after elimination of unimportant content, there is still information overload in remaining news-related data, which must be prioritized for utilization. To assist in prioritization of news information, news must be ranked in order of estimated importance. At first, preprocessing is carried out. Key terms are extracted and filtered from news and social information admires a selected amount of your time. A graph is made (which is known as Key Term graph) from the antecedently extracted key term set, whose vertices represent the key terms and edges represent the co-occurrence similarity between them. The graph, when process and pruning, contains slightly joint clusters of topics in style in each print media and social media. Then the graph is clustered so as to get well-defined and disjoint sub graphs.The sub graphs from the main graph are selected and ranked based on user attention. Thus the thesis effectively identifies news topics that are prevalent in both social media and the news media, and then ranks them. The Louvain algorithm is used for communication detection for social media. Finally, the results are validated by using the Validation metrics Modularity and Edge density. Clique method is provides the best result, to compare with Louvain algorithm.**

Keywords- Component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Data mining or data discovery is that the computer-assisted method of dig through and analyzing huge sets of information and so extracting the means of the information. Data processing tools predict behaviors and future trends, allowing businesses to create proactive, knowledge-driven choices.

Data mining tools will answer business queries that historically were too time overwhelming to resolve. They scour databases for hidden patterns, finding prophetical data that consultants could miss as a result of it lies outside their expectations. Data processing derives its name from the similarities between looking for valuable data during a giant info and mining a mountain for a vein of valuable ore. Each process needs either winnowing through associate degree quantity of fabric, or showing intelligence searching it to search out wherever the worth resides. though data processing continues to be in its infancy, firms during a big selection of industries - together with retail, finance, health care, producing transportation, and region - square measure already mistreatment data processing tools and techniques to require advantage of historical knowledge.

By mistreatment pattern recognition technologies, applied mathematics and mathematical techniques to shift through warehoused data, data processing helps to analyst acknowledge important facts, relationships, trends, patterns, exceptions and anomalies that may otherwise go unnoticed. For businesses, data processing is employed to find patterns and relationships within the knowledge so as to assist create higher business selections. Data processing will facilitate spot sales trends, develop smarter promoting campaigns, and accurately predict client loyalty.

Specific uses of information mining include:

• Market segmentation - determine the common characteristics of shoppers UN agency purchase constant product from your company.

• Customer churn - Predict that customers square measure seemingly to depart your company and visit a contestant.

• Fraud detection - determine that transactions square measure presumably to be dishonorable.

• Direct promoting - determine that prospects ought to be enclosed during a list to get the very best response rate.

• Interactive promoting - Predict what every individual accessing an internet website is presumably fascinated by seeing.

• Market basket analysis - perceives what product or services square measure unremarkably purchased together; e.g., brew and diapers.

Data Mining is that the method of analyzing unknown patterns of knowledge in line with totally different views for categorization into helpful information, that is collected and assembled in widespread areas, like knowledge warehouses, for economical analysis, data processing algorithms, facilitating business deciding and different data necessities to ultimately cut prices and increase revenue [https://www.techopedia.com/definition/1181/data-mining].
Selecting a Template (Heading 2)

II. RELATED WORKS

In the paper "Toward Collective Behavior Prediction via Social Dimension Extraction" [1] the authors Lei Tang and Huan Liu, Arizona State University within the year of 2010 were expressed that collective behavior refers to however people behave after they area unit exposed during a social network setting. Within the paper, they examined however they may predict on-line behaviors of users during a network, given the behavior data of some actors within the network.

They incontestible several benefits, particularly appropriate for large-scale networks, paving the manner for the study of collective behavior in several real-world applications. Social media like Facebook, MySpace, Twitter, BlogCatalog, Digg, YouTube and Flickr, facilitate folks of all walks of life to specific their thoughts, voice their opinions, and hook up with one another anytime and anyplace. for example, a well-liked content-sharing web site like Delicious, Flickr, and YouTube permits users to transfer, tag and comment differing types of contents (e.g., bookmarks, photos, videos).

One's behavior is often influenced by the behavior of his/her friends. This naturally results in behavior correlation between connected users. Such collective behavior correlation can even be explained by homophily[5].

In this paper "Finding community structure in networks mistreatment the eigenvectors of matrices" [2] the author M. E. J. Newman thought-about the matter within the year of 2006 were detective work communities or modules in networks, teams of vertices with a higher-than-average density of edges connecting them. Previous work indicates that a strong approach to the present drawback is that the maximization of the profit operates referred to as "modularity" over doable divisions of a network. Here the author showed that this maximization method are often written in terms of the eigen-spectrum of a matrix they referred to as the modularity matrix, that plays a job in community detection kind of like that contend by the graph Laplacian in graph partitioning calculations. They expressed that a typical feature of the many networks is "Community Structure", the tendency for vertices to divide into teams, with dense connections inside teams and solely sparser connections between them.

In social networks, as an example, it's long been accepted that people United Nations agency lie on the boundaries of communities, bridging gaps between otherwise unconnected folks, get pleasure from AN uncommon level of influence because the gatekeepers of data flow between teams [6, 7, 8].

In this paper "Yes, there's a Correlation - From Social Networks to private Behavior on the Web" [3] the authors Parag Singla and Matthew Richardson expressed that characterizing the connection that exists between a person's grouping and private behavior has been an extended standing goal of social network [9] analysts. They applied data processing techniques to review this relationship for a population of over ten million folks, by turning to on-line sources of knowledge.

The analysis reveals that folks United Nations agency chat with one another (using instant messaging) area unit a lot of doubtless to share interests (their internet searches area unit a similar or locally similar). The longer they pay talking, stronger their relationship. People who chat with alternative |one another} are a lot of doubtless to share other personal characteristics, like their age and site and, they're doubtless to be of opposite gender. Similar findings hold for those that don't essentially talk over with one another however do have an admirer in common. Their analysis relies on a well-defined mathematical formulation of the matter, and is that the largest such study they were alert to.

In this paper "BIRDS OF A FEATHER: Homophily in Social Networks" [4] the authors Miller revivalist, Lynn Smith-Lovin and James M Cook expressed concerning "Similarity breeds connection". This principle the homophily principle-structures network ties of each kind, together with wedding, friendship, work, advice, support, data transfer, exchange, co-membership, and different varieties of relationship. The result's that people's personal networks area unit unvaried with respect to several socio demographic, behavioral, and intrapersonal characteristics. Homophily limits people's social world during a manner that has powerful implications for the data they receive, the attitudes, and also the interactions they expertise.

Homophily in race and quality creates the strongest divides within the personal environments, with age, religion, education, occupation, and gender following in roughly that order. Geographic proximity, families, organizations, and isomorphous positions in social systems all produce contexts within which homophilous relations type. Ties between nonsimilar people additionally dissolve at a better rate that sets the stage for the formation of niches (localized positions) inside social area.

They argued for a lot of analysis on: (a) the fundamental ecological processes that link organizations, associations, cultural communities, social movements, and plenty of different social forms. (b) The impact of multiplex ties on the patterns of homophily and (c) The dynamics of network amendment over time through that networks and different social entities co-evolve.

III. SYSTEM DESIGN

A. Introduction

Social media and traditional media combined together, they feed off of each other and are mutually beneficial. Together, they create a much stronger, much more effective and successful marketing campaign. The intersection of two media information is used to project the popularity of the news at a particular period of time. Thus the analysis provides a strong opinion about particular news for decision making in future. This chapter states the problem of the community detection in large graph and outlines the overall view of the existing work

B. Existing Work

The concept of detecting the community from a large network in the existing work is given below

a. Girvan-Newman Clustering

The Girvan–Newman algorithm detects communities by increasingly removing edges from the original network. The associated components of the remaining network are the communities. Vertex betweenness

is a pointer of greatly central nodes in network. For any node, vertex betweenness is defined as the number of shortest paths between pairs of nodes that run through it. If there is additional shortest path between a pair of nodes, each path is assigned equal weight such that the total weight of all of the paths is equal to unity. So the edges Note connecting communities will have high edge betweenness (at least one of them). By removing these edges, the groups are divided from one another and so the underlying community structure of the network is exposed.

The algorithm's step for community detection is summarized as follows

Step 1: Find the edge of highest betweenness - or multiple edges of highest betweenness

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(V)}{\sigma_{st}} \quad g(v) \in (0,1) \qquad (1)$$

Where,

σ_{st} isthe total no. of shortest paths starting from beginning node "s" to ending node "t"

$\sigma_{st}(V)$ is number of the shortest path through V.

Step 2: The edge with highest betweennessvalue is removed.

Step 3: Recalculate all betweenness, and again remove the edge or edges of highest betweenness.

Step 4: Proceed in this way as long as edges remain in graph, in each step recalculating all betweenness and removing the edge or edges of highest betweenness.

The betweenness centrality should be recalculated with each step. The reason is that the network adapts itself to the new conditions set after the edge is removed. For example, if two communities are getting connected by more than one edge, then there is no guarantee that all of these edges will have high betweenness. By recalculating betweenness after the removal of each edge, it is ensured that at least one of the remaining edges between two communities will always have a high value [https://en.wikipedia.org/wiki/Girvan%E2%80%93Newman_algorithm].

Drawbacks of Girvan-Newman clustering

- Girvan-Newman clustering is too slow for large networks
- It yields relatively poor result for dense network
- It takes more computation time to partition the large graph
- The clustering approach is not employed in order to obtain overlapping topic clusters

C. PROBLEM DEFINITON

Twitter is an American online news and social networking service on which the users post and interact with messages known as "tweets". Registered user can post tweets, talk about news and share interesting topics via social network services but those who are unregistered can only read them [https://en.wikipedia.org/wiki/Twitter].

The news media (traditional media such as web news crawls, website news forums) contain the professionally verified events. The valuable information is obtained by intersecting two media sources. The community is detected to discover how particular topic is discussed by the user. It can be used to provide the strong opinion on the particular news present in the media. Detecting the communities from the large network is not an easy task.

In the existing system, the community is detected by using the Girvan-Newman clustering method, which detects communities in smaller graphs. Based on the betweenness [https://en.wikipedia.org/wiki/Girvan%E2%80%93Newman_algorithm]

In order to accomplish the detection of communities from the large graph the proposed work is done. The proposed system utilizes two methods namely, CLIQUE (CLusterInQUEst) detection and Louvain method to detect the communities effectively. In CLIQUE detection, it uses the multi-resolution grid data structure. The cluster contains the maximal set of actors in which every actor is connected to each other. It generates the minimal number of description for the clusters. The Louvain method is a greedy optimization method. It allows to efficiently compute the edge ranking in large network in linear time. Finally it discovers the community structure by optimizing the modularity of the network.

D. Implementation Tool

The implementation tool employed in the present work isR.

R could be an artificial language and free code surroundings for applied mathematics computing and graphics supported by the R Foundation for applied mathematics Computing. The R language is wide used among statisticians and information miners for developing applied mathematics code and information analysis. R includes a command interface; there are many graphical user interfaces, like RStudio, AN Integrated development surroundings. R is AN implementation of the S artificial language combined with lexical scoping linguistics, impressed by theme.Some of the R packages employed in current work area unit as follows

Packages	Description
RColorBrewer	Palettes for thematic maps
Tm	Framework for Text mining
TwitterR	Access to the Twitter API
WordCloud	Plot a cloud of words
RoAuth	R open Authentication
NLP	For Natural Language Processing
SnowBallc	For stemming the words
RCurl	Request URL
TextmineR	Create corpus
Textclean	Normalizing and cleaning the text
Igraph	Fast handling of large graphs
Syuzhet	Quickly extract the plot
Plyr	Splitting big data structure, apply function and combine all together
XML	For paring and integrating XML

E. Summary

This chapter describes about the problem definition and the overview of existing algorithm Girvan-Newman clustering. Consider the drawbacks of the existing work; the current work uses the CLIQUE detection and Louvain community detection model to get best community structure than the existing work, which is described in the next chapter 4, System

IV. SYSTEM METHODOLOY

A. *Introduction*

The keywords of the combined sources (Twitter and News media) help to find out the intersection of the words and co-occurrence words, which help to create the news term graph whereas the vertices are the text and the edges are the relationship, exist among vertices. So that the community detection algorithm detects the dense region that are frequently crawled information in news and twitter. CLIQUE and Louvain method is used to detect the better communities than the existing method. This chapter describes the proposed algorithm in the current work.

B. SYSTEM ARCHITECTURE

The modules in the current work are as follows

- Dataset collection
- Preprocessing
- Key term graph construction
- Key term similarity estimation
- Graph clustering : Girvan-Newman clustering
- Content selection : User Attention (UA)
- CLIQUE detection algorithm
- Louvain algorithm
- Performance evaluation

a. DATASET COLLECTION

NEWS DATA:

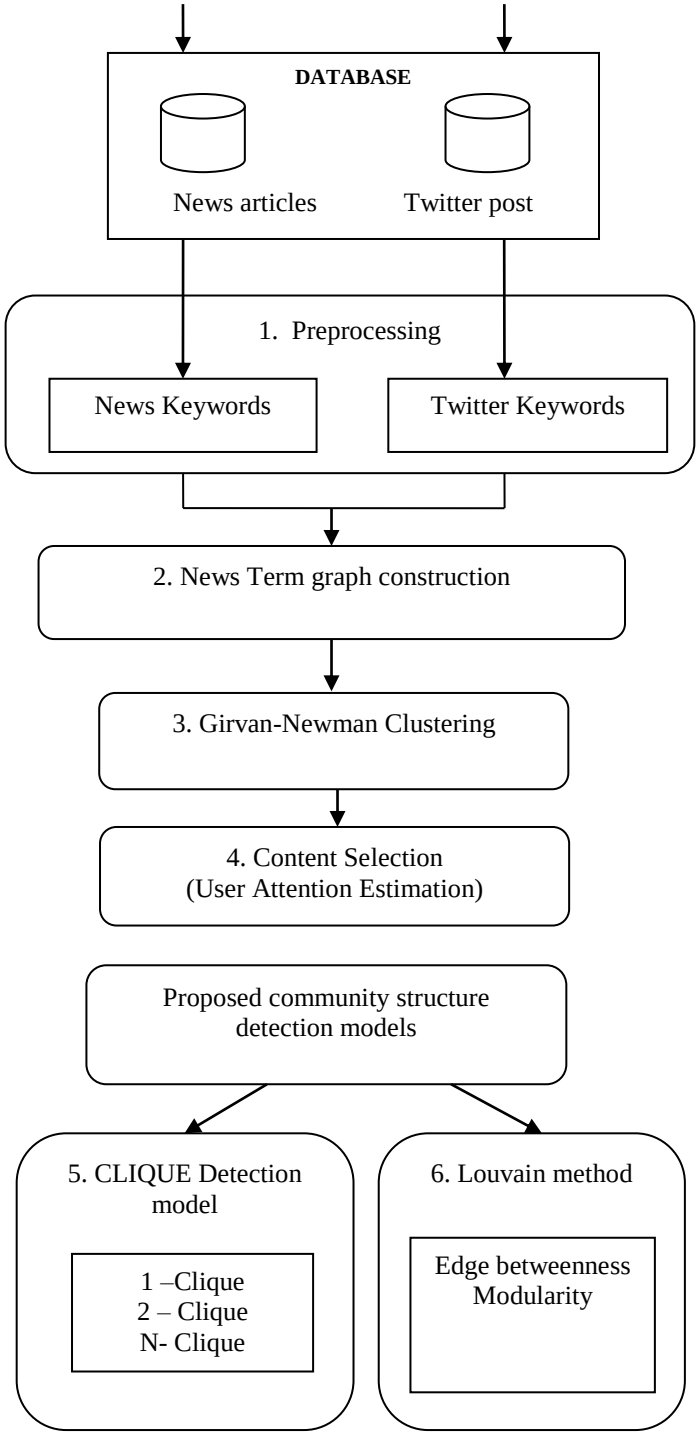
The BBC newswebsite (https://www.bbc.com/news) contains international news coverage, as well as British, entertainment, science, and political news. Many reports are convoyed by audio/video from BBC's television/radio news services. It is providing the interdisciplinary fields of news such as world, port, weather, travel, business, entertainment, health, science, technology and so on. For the current work the sports category news information from the date 1.11.2018 to 30.11.2018 are downloaded from the website (https://www.bbc.com/news/sports).

TWITTER DATA:

Twitter account is used to create the Application Program Interface (API). The API provides the consumer secret key and access token secret key for authenticated retrieval of tweets. The number of tweets related to sports news is collected. The figure 4.1 shows the overall architecture of the current work. The modules in the work are explained below.

b. PREPROCESSING

The collected news articles and the tweets are preprocessed in this step.



The word “data” is plural, not singular.

Figure 4.1. The overall architecture of the current work

c. Key Term Graph Construction

A graph G is generated, whereas the clustered nodes represent the prevalent news topic in both news and social media. The vertices in the graph G is the terms retrieved from N and T and the edges exhibit the relationship among the

nodes. The following methods are used to find out the relationships between the words.

- Term Document Frequency

The document frequency of each term in News and Twitter is calculated accordingly. Here df(n) is the occurrence of term n and df(t) is the occurrence of term t.

- Relevant Key Term Identification

N represents the keywords present in the news article and T represent all relevant term present in the tweets. To extract the topics that are prevalent in both news and social media, the following formula is used.

$$I = N \cap T \quad (2)$$

This intersection of N and T eliminates the terms from T that are not relevant to the news and terms from N that are not mentioned in the social media. I (intersection words) are ranked based on their prevalence in both sources. The prevalence of a term is the combination of it occurrence in both N and T.

$$\forall i \in I : p(i) = \frac{df(n)\times\frac{|T|}{|N|}+df(t)}{2|T|} \quad (3)$$

Where,|T| is the total number of tweets chosen between dates d1 and d2.|N| is the total number of news chosen in the same period of time.

Key Term Similarity Estimation

The perception behind the co-occurrence is the terms that co-occur frequently are related to the same topic and may be used to summarize and represent it when grouped. The co-occurrence for each term pair (i,j) I found, defined as co(i,j). The term-pair co-occurrence is then used to estimate the similarity between terms. A number of similarity measure were tested, namely Jaccard, Dice and Cosine similarity.

The Dice similarity between term I and j is calculated as follows,

$$dice\_QS(i,j) = \begin{cases} 0 & if co(i,j) \le \vartheta \\ \frac{2\times co(i,j)}{df_{top}(i)+df_{top}(j)} & otherwise \end{cases} \quad (4)$$

Where,

df_{top} (i) is the number of tweet that contain term i$\in I_{top}$

df_{top} (j) is the number of tweet that contain term j$\in I_{top}$

co(i,j) is the number of tweets in which terms i and j co-occur in I_{top}

$\vartheta$is a threshold used to discard whose similarity that fall below it

The Jaccard similarity between term I and j is calculated as follows,

$$jac_{QS(i,j)} = \begin{cases} 0 & if co(i,j) \le \vartheta \\ \frac{co(i,j)}{df_{top}(i)+df_{top}(j)-co(i,j)} & otherwise \end{cases} \quad (5)$$

The Cosine similarity between term I and j is calculated as follows,

$$cosine_{QS(i,j)} = \begin{cases} 0 & if co(i,j) \le \vartheta \\ \frac{co(i,j)}{\sqrt{df_{top}(i)\times df_{top}(j)}} & otherwise \end{cases} \quad (6)$$

All of the formerly described similarity measures generate a value between 0 and 1.

d. Graph Clustering: Girvan-Newman

This algorithm used to find out the word clusters. The goal is to identify and separate the well defined sub graphs in the graph.Betweenness

The core idea of Newman clustering is the concept of edge betweenness. The betweenness value of an edge is the number of shortest paths between pairs of nodes that run along it. The betweenness measure of an edge e is calculated a follows,

$$\text{Betweenness (e)} = \sum_{i,j\in V}\frac{\sigma(i,j)|e}{\sigma(i,j)} \quad (7)$$

Where,

V is the set of vertices

$\sigma(i,j)$is the number of shortest path between vertex i and j

$\sigma(i,j)|e$is the number of those paths that pass through edge e.

1. Transitivity

It is a property in a relation between three elements such that if the relation holds between the first and second elements, and between the second and third elements, then it also holds between the first and third elements. The transitivity of a graph G is defined as

$$\text{Transitivity (g)} = \frac{\#\,triangle}{\#triads} \quad (8)$$

Algorithm Girvan - Newman

Improve the Cluster Quality of a Graph

Input: Graph G
Output: Cluster-quality-improved G
B = {} empty set
repeat
for all (edge e ∈ G) do
Calculate betweenness(e) and append to B
end for
if first iteration of loop then
b_{avg} = avg(B)
end if
bmax = max(B)
trans0 = transitivity(G) previous transitivity
Remove edge with b_{max} from G
trans1 = transitivity(G) posterior transitivity
Clear set B
until (trans1 < trans0 or b_{max}<b_{avg})
Add edge with b_{max} to G

Step 1: The betweenness values of all edges in graph G are calculated.

Step 2: The initial average betweenness of graph G is calculated.

Step 3: The high betweenness values are iteratively removed in order to separate clusters.

Step 4: The edge removing process is closed when removing additional edges yields no gain to the clustering quality of the graph. Once the process has been topped, the last detached edge is added back to G.

e. Content Selection : User Attention

The User Attention (UA) represents the number of unique Twitter user related to the selected tweets. The tweets related to that topic are selected and then the number of unique users who created those tweets are counted. The equation for finding the UA is given below.

$$\forall TC \in GUA(TC) = \frac{|U_{TC}|}{\sum_{TC \in G} |U_{TC}|} \quad (9)$$

(OR)

$$UA = \frac{Totalnumberofuniqueusers}{sumoftotalnumberofuniqueusers} \quad (10)$$

Where,

U_{TC} is the number of unique users related to TC

G is the entire graph

This equation produces a value between 0 and 1.

f. Clique Detection Algorithm

The CLIQUE algorithm was one of the first subspace clustering algorithm. It identifies dense clusters in maximum dimensionality's subspaces. The algorithm unites density and grid based clustering. It uses an APRIORI style search technique to detect dense subspaces. Then the algorithm finds adjacent densegridunitsineachofselectedsubspacesusing"depthfirstsearch".Clustersarethen formed by uniting these units with the help of a greedy growth scheme. The algorithm begins with an arbitrary dense unit and then greedily produce sa maximal region in each dimension until the uniono fall there gions covers the entire cluster. Redundant regions are removed by a repeated procedure.

The region growing, density based approach to generating clusters allows CLIQUE to find clusters of arbitrary shape, in any number of dimensions. Clusters are found in same, disjoint or overlapping subspaces. This is often advantageous in subspace clustering since the clusters often exist in different subspaces and thus represent different relationships.

CLIQUE, consists of the following steps:

1. Identification of subspaces that contain clusters
2. Identification of clusters
3. Generation of minimal description for the clusters

Algorithm steps for CLIQUE

1)Identification of subspace that is dense

a) Finding of dense units

Find the set D1 of all one dimensional dense unit

K=1

While $D_k ? = \emptyset$

K=k+1

Find the set D_k which is set of all k-dimensional dense units whose all lower dimension projections (k-1), belong to $D_k - 1$

End while

b) Finding sub spaces of high coverage

2)Identification of clusters

For each high coverage subspace s do

Take the set of all dense units (E in S)

While E! = $\emptyset$

M=1

Select a randomly chosen unit u from E

Assign toC_m, U and all units of E that are connected to U

E = E-C_m

End while

End for

3)Generate minimal cluster descriptions

For each cluster C do

$\underline{1^{st}}$ Stage

C=0

While c! = $\emptyset$

X = X+1

Choose a dense unit in C

For i = 1 to L

Unit proceeds in both the direction along the I^{th} dimension.

End for

Represent the set containing the entire unit covered by the above procedure

C = C-1

End while

$\underline{2^{nd}}$ Stage

Remove all covers from the units covered by another cover

g. *Uvain method*

The Louvain method is simple, effective and easy-to-implement method to identify communities in large networks. The method is used along with success for networks of different types and for sizes ranging upto100 million nodes and billions of links.

The method consists of two phases.

1)It looks for "small" communities by optimizing modularity in a local way.

2) It aggregates nodes of the same community and builds a new network whose nodes are the communities. These steps are iteratively repeated until a maximum of modularity is attained.

The partition found after the first step typically includes many communities of small sizes. At succeeding steps, larger and larger communities are found due to the aggregation mechanism. This process will naturally lead to hierarchical decomposition of the network.This is clearly associate approximate technique and ensures that the world most of modularity is earned, however many tests have confirmed that our algorithmic rule has a wonderful accuracy and sometimes provides a decomposition in communities that encompasses a modularity that's about to optimality.

A graph $G = (V,E)$ is created where V and E are the sets of nodes and edges. Community detection is performed by dividing graph G into clusters $C = \{V1,V2, \dots ,Vx\}$ and each Vi , a set of nodes, is called community. The figure 4.2 shows the large network of nodes and edges which is clustered using Louvain algorithm; the communities are differentiated by using various color nodes.

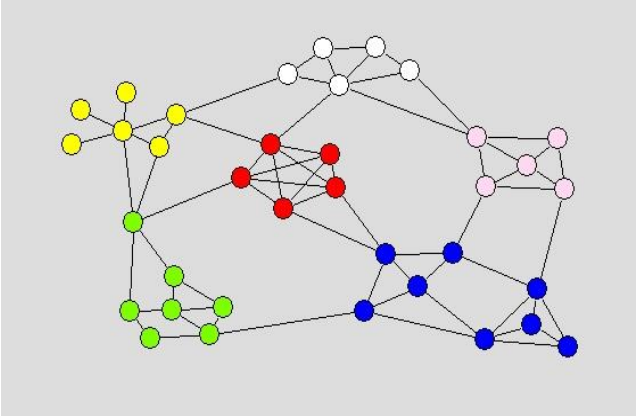


Figure 4.2. Louvain community detection method

Louvain algorithm (Graph G)
G' = G
C the index of community of each nodes of G'
Initialize each node with its own community
q = -∞
while q< Q(G',G) do
q = Q (G', G)
c = MoveNodes (G', G) // Phase 1
G' = Aggregate (G', G) // Phase 2
C = put each node G' in its own community
End while
Return G'
End function

Function MoveNodes (Graph G)
C the index of communities of each nodes of G
While one or more nodes are moved do
for random v ∈ V (G) do
best_q = -∞
best_c = community of v
for all neighboring nodes n of v do
gain_q = ΔQ between v and n
ifbest_q<gain_q< then
best_q = gain_q
best_c = community of n
end if
end for
C = place v in the best_q
end while
return c
end function

Function Aggregate (Graph G, Partition C)

G' = aggregate nodes which are in same community based on C
Return G'
End function

This is an iterative algorithm repeating till there is no additional modularity improvement. It begins with initialization ofall nodes with its own community. In Phase 1, for every node in a graph, it computes modularity gain ΔQ for all neighboring communities if the node found to be moving.

Q indicates gain of modularity and is defined by

$$\Delta Q = \left[\frac{\sum_{in} k_{i,in}}{2m} - \left[\frac{\sum_{tot}+k_i}{2m}\right]^2\right] - \left[\frac{\sum_{in}}{2m} - \left[\frac{\sum_{tot}}{2m}\right]^2 - \left[\frac{k_i}{2m}\right]^2\right] \quad (11)$$

where $\sum_{in}$ is the sum of the weights of the links inside the community to which the node i is assigned, $\sum_{tot}$ is the sum of weights of the links incident to community nodes, and $k_{i,in}$ is the sum of the weights of the links from i to nodes in the community which is same with the community of node i.

In the Phase 2, all communities are collapsed to the vertices to create a new graph internal community edges are collapsed into a single self-looping edge, and the weight is the sum of edge weights of the entire internal community edges in the community. Multiple edges between every two communities are collapsed to form a single edge, and weight is the sum of edges between them.

C. Summary

This chapter has described about the community detection using CLIQUE and Louvain method. CLIQUE detection model finds the communities of minimizing dimensionality. And the Louvain method finds the communities by maximizing the modularity. The current work determines that the CLIQUE clustering find out the best communities than the Louvain method. The results obtained by using CLIQUE and Louvain method are given in the next Chapter 5, Results and discussion.

V. RESULTS AND DISCUSSIONS

A. Experimental analysis

The existing and proposed work detects the community structure from the news media and twitter media. The Process of existing and proposed work contains the following steps:

Step 1: The input news data is first downloaded from the BBC news portals (http://www.bbc.com/) and tweets are collected by using the Twitter API.

Step 2: The keywords of the news and twitter media is generated separately. The intersection of the two media keywords is found. The frequency of the words is calculated using TF-IDF.

Step 3: The relationship between the keywords can be found by using three similarity measures namely Dice, Jaccard and cosine similarity measures. So the vertices are the text words which are connected by the edges.

Step 4: The vertices and edges forms the clusters that are obtained by using the Girvan-Newman clustering method. And the User Attention (UA) of the resultant cluster is calculated.

Step 5: The resultant graph obtained in step 4 is fed into the CLIQUE community detection and Louvain method as input.

Step 6: At last the dense community graph is produced as output.

Step 7: Finally, the edge density and modularity is calculated to evaluate the quality of the community structure.

B. Comparison result

Based on the number of twitter and news keywords, different community structure has been obtained. The performance is evaluated by increasing the number of news, check whether the methods yield finest community even for large graphs. Among the three methods, the CLIQUE method yields better (strengthen) community structure with the high rate of evaluation metrics, even the news and tweets are increased to some extent. The results for different number of input are explained in graphs.

No. of Twitter and News Post	100 N + 400 T	200 N + 500 T	250 N + 1000 T
ALGORITHM			
Newman Cluster	3	4	5
Louvain Cluster	3	2	3
Clique Cluster	1	24	34

In the table 5.1, the number of clusters created according to the number of tweets and news are given below.

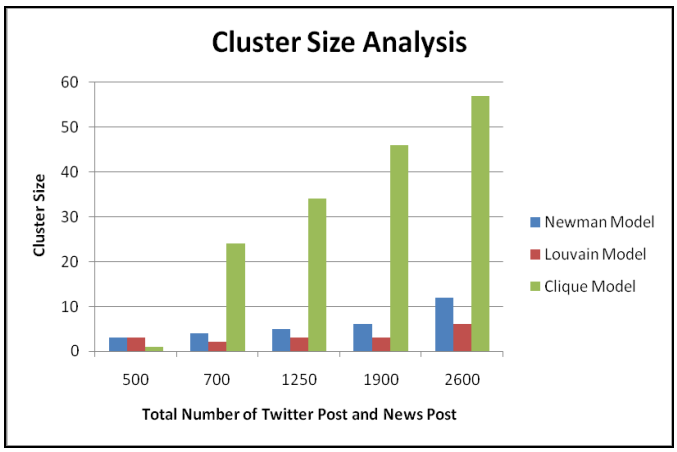


Figure 5.1. Resultant Graph for Cluster size

VI. CONCLUSION

Classification is applied to developed to automatically analyze the emotional polarity of a text, based on which a value for each piece of text is obtained. The absolute value of the text denotes influential power and the sign of text denotes its emotional polarity.

This Graph clustering is applied to develop integrated approach for online sports forums cluster analysis. Clustering algorithm is applied to classify the forums into various clusters, with the middle of each cluster representing a hotspot forum within current time span.

Along with clustering the forums based on data from current time window, conducted forecast is also conducted for the next time window. Empirical studies give strong proof of existence of correlations between post text sentiments and hotspot distributions.

Education Institutions being information seekers benefit from hotspot predicting approaches in various ways. They followed the same rules as academic objectives, and are measurable, quantifiable, and also time specific. However, in real, parents/students behavior is always hard to be capture explored.

Using the hotspot predicting approaches can help the education institutions understand what their specific customer's timely concerns regarding goods and services information. Results generated from these approaches can be combined to competitor analysis to defer comprehensive decision support information.

VII. REFERENCES

[1] L. Tang and H. Liu,"Toward predicting collective behavior via social dimension extraction," IEEE Intelligent Systems, vol. 25, pp. 19–25, 2010.
[2] M. Newman, "Finding community structure in networks using the eigenvectors of matrices" Physical Review (Statistical, Nonlinear and Soft Matter Physics), vol. 74, no.3,2006.
[3] P. Singla and M. Richardson, "Yes, there is a correlation: - from social networks to personal behavior on the web," in WWW '08: Proceeding of the 17th international conference on World Wide Web. New York, NY, USA: ACM, 2008, pp. 655–664.
[4] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual Review of Sociology, vol. 27, pp. 415–444, 2001.
[5] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27:415–444, 2001.
[6] M. Granovetter, The strength of weak ties. Am. J. Sociol. 78, 1360–1380 (1973).
[7] R. S. Burt, Positions in networks. Social Forces 55, 93– 122 (1976).
[8] L. C. Freeman, A set of measures of centrality based upon betweenness. Sociometry 40, 35–41 (1977).
[9] P. Doreian and T. Snijders, editors. Social Networks, 2006.