Uncovering Influential Nodes in Directed Graph Using Graph Centrality

Anjana N Iyyer[1], Mary John[1]
[1]Department of IT, Rajagiri School of Engineering and Technology, India

Abstract: Directed graphs can be used to represent different networks like social network, biological network, urban network etc. Different trends and patterns can be identified and analysed using this graph. Random or manual identification of top users is a tedious task and is less accurate; so graph centrality measures can be used to find top representative nodes in a very large directed graph. Analytics tools that are available are not affordable to all since it is highly expensive. The main objective here is to use different graph centrality measures to identify the top representative nodes in a very large directed graph such as Twitter. The algorithm proposed here can be used to process the graph and find out prominent users in the network. The final prominent nodes uncovered will be categorized based on the location of users.

Keywords: directed graph, fuzzy set, graph centrality, social network, twitter.

I. Introduction

Large data sets have to be computationally analysed so that the associations, patterns and trends can be revealed. Since it is so complex and voluminous traditional applications and data processing systems are not adequate to deal with these. The role of directed graph comes here and it is used to represent the dataset in a different form than it is and can be used in an easier way to manage the data. The big data growth has influenced a lot of fields and social media is one such field which has impacted a lot with this growth. This is a reason for taking social media as an example to analyse the algorithm. Graph centrality is used for gaining knowledge regarding the central nodes and it is used for making predictions about other nodes and thereby their behaviour can be analysed. There is a chance that least influential users (passive users) can provide a barrier to the transmission in the network. So the most influential users need to be found out from a very large network. An example of a directed graph is Twitter social network, where it has more than 300 million users and more than 100 billion connections between them. The algorithm proposed here will be pragmatically verified on Twitter to discover influential users using the fuzzy set theory concept. The web based life, because of its hazardous development has given a huge number of individuals a chance to make and share content on a scale scarcely possible a couple of years prior. Incalculable number of sentiments, news and item audits that are continually posted and talked about in social locales, for example, Facebook, Twitter, Instagram, Pinterest and so forth mirrors the enormous support of the users in interpersonal organization.

Graph Centrality is utilized here to pick up information about the nodes and is utilized to cause expectations about different nodes with the goal that their conduct to can be examined. It recognizes comparable conduct, surprising conduct and increase information about nodes and make forecasts. Utilizing this component the nodes which are focal among different nodes in the system can be resolved. The social media, because of its unstable development has furnished a great many individuals with a chance to make and share content on a scale scarcely possible a couple of years back. Innumerable number of sentiments, news and item surveys that are continually posted and examined in social destinations, for example, Facebook, Twitter, Instagram, Pinterest and so forth mirrors the gigantic cooperation of the users in informal organization. Graph centrality is a measure which can be utilized to recognize the most significant nodes in a graph. Its application can be found in various fields like material science, science, transportation, drug, informal community and bibliometrics and so forth. A portion of the applications incorporate ID of most persuasive users in an informal organization, key foundation nodes in urban systems, super spreaders of infection, assessment rivalry and gossip spreading. In all these, the key thought is to reveal the most persuasive nodes in the graph. Compelling nodes in a graph are utilized to pick up learning about the nodes associated with them and to cause forecasts about these nodes so an investigation on their conduct to can be performed. Section I gives an idea and introduction about big data and graph centrality. Section II is a review on existing works done using Katz centrality, Eigenvector centrality, Degree centrality, Betweenness centrality and Closeness centrality and its applications. Section III discusses about the proposed system analysis and its design. It tells about the various phases in the system and their outputs. It also deals with the proposed system architecture. Results and discussion is the Section IV. It includes the major technologies and algorithms used. Section V is the conclusion part which describes the summary of the whole project and its results. It also includes some future enhancements that can be done to improve the performance of the system. Then some valid reference papers are given.

II. Literarture Survey

In the paper by Mohammad et.al [1], Katz centrality has been investigated alongside proliferation likelihood. It utilizes katz centrality measure to discover top-k powerful users in informal organization. It has been discovered that Katz centrality, a subset of Eigenvector centrality is extremely compelling one and can be utilized alongside proliferation likelihood to discover top-k persuasive users. The primary motivation behind this paper is to choose powerful users relying upon the accessible spending that augments the impact inclusion in the system. Existing work, for the most part center around structuring techniques dependent on centrality measurements because of their low time multifaceted nature and satisfactory impact spread and since methodologies based avaricious calculation experience the ill effects of high time unpredictability. In this paper, another calculation "PrKatz" is proposed dependent on Katz centrality and a proliferation likelihood limit that gives a specific capacity to impact users effectively. The test results on enormous datasets exhibit the presentation of the proposed calculation contrasted and the current calculations in term of impact spread.

Paramita Dey et.al [2] utilizes graph centrality measure to decide the dimensions and profundity of the data spread in the system. Distinctive centrality measures are utilized as a way to deal with select the compelling nodes for data hindering in the informal community. It has been discovered that Betweenness centrality performs superior to anything other centrality measures here. The centrality nodes assume a significant job to decide the dimensions and profundity of the data spread in the system. This paper shows a way to deal with select the compelling nodes for data hindering in the interpersonal organization. It catches the communication proportions of nodes in the informal organization and chooses powerful nodes dependent on three significant system properties, for example degree dispersion, betweenness centrality and closeness centrality. Edge betweenness, utilized to locate the significant edges of the system show better outcome as far as data blocking.

In Ashish Mehrotra et.al [3], it utilizes graph centrality measures to discover and mark counterfeit users. Distinctive centrality measures have been utilized together to discover counterfeit twitter adherents. In this paper, a technique is formulated which can be utilized to distinguish all the phony adherents inside a social graph arrange dependent on highlights identified with the centrality of the considerable number of nodes in the graph and preparing a classifier dependent on a subset of the information. Utilizing just graph based centrality measures, the proposed strategy yielded high precision on fake follower discovery. The proposed strategy is conventional in nature and can be utilized regardless of interpersonal organization stage under thought. While the proposed calculation has indicated promising outcomes amid the testing stage, progressively thorough testing could be performed with the assistance of a lot bigger named datasets alongside increasingly computational assets which could be utilized for further graph representation utilizing open source ventures like Gephi which was utilized in the examination for visual investigation and perceptions. With accessibility of more assets for this trial to be completed further, the calculation proposed in this paper has an extent of progress before it very well may be broadly embraced as a summed up generation level method for identification of phony adherents on any long range interpersonal communication stage.

Prantik Howlader et.al[4] utilizes Degree Eigenvector centrality measures to examine persuasive users in Twitter. Some positive connection between's Degree centrality and Eigenvector centrality has been resolved. It has been discovered that both of these centrality measures must be utilized together for better execution and exact outcomes. In this paper, in light of information gathered from Twitter, an investigation of eigenvector centrality approach of finding the powerful users is performed. The variety in indegree and eigenvector centrality of users taking part in a hashtag in Twitter, as for change in the measure of connections is researched here. The accompanying perceptions were made. Initially, in Twitter, users with high eigenvector centrality need not be compelling users. Second, in a given hashtag, there is an expansion in users with both high indegree and eigenvector centrality when there are more user collaborations. Third, there is a positive relationship among's indegree and eigenvector centrality.

Mahdieh Ghasemi et.al[5] utilizes graph centrality measures to discover significant components set specifically positions. Various systems like PPI arrange, buildup collaboration and quality communication systems are examined here with the centrality measures. It has been discovered that Eigenvector centrality is the best one and can be utilized even on the off chance that where a portion of the information is absent from the system. A normally known actuality in organic and interpersonal organizations' investigation is that in many systems some significant or compelling components are set in some specific positions in a system. These positions have some specific basic properties. Centrality measures evaluate such realities from various perspectives. In light of centrality estimates the graph components, for example, vertices and edges can be positioned from various perspectives. This paper displays a complete audit of existing diverse centrality measures and applications in some natural systems, like Protein-Protein networks, buildup connection and quality collaboration systems.

A study about social media characteristics and the influence of users in social media and the existing algorithms was also done [7], [8], [9], [10], [11], [12]. Some existing data analytics tools for social media which is mostly paid was also referenced [13], [14], [15], [16].

III. System Analysis and Design

A. Proposed Architecture

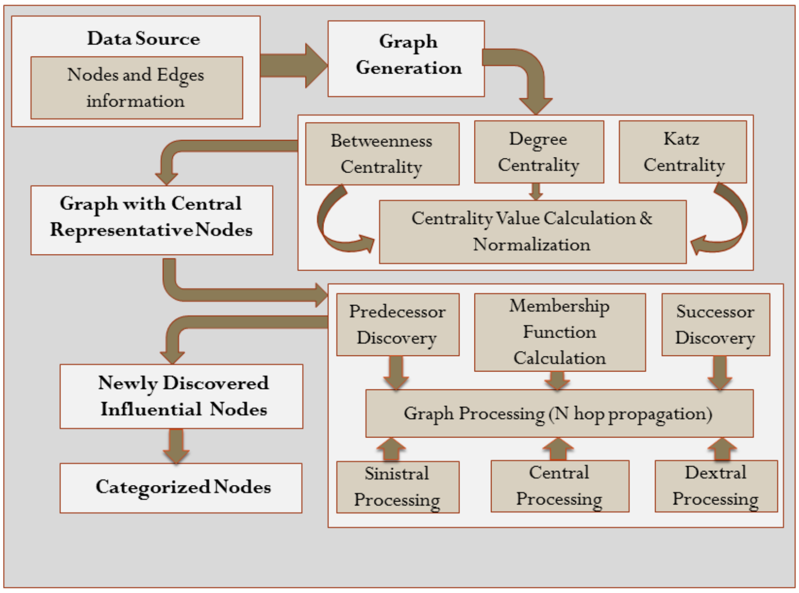
Data Source
Nodes and Edges information
Graph Generation
Betweenness Centrality
Degree Centrality
Katz Centrality
Centrality Value Calculation & Normalization
Graph with Central Representative Nodes
Predecessor Discovery
Membership Function Calculation
Successor Discovery
Graph Processing (N hop propagation)
Sinistral Processing
Central Processing
Dextral Processing
Newly Discovered Influential Nodes
Categorized Nodes

Fig 1 System Architecture

a) Data Source

The information about the directed graph is represented by the data source which includes the successor and predecessor information of the entire nodes participating in the algorithm. The information from the data source can be used to identify the nodes and edges of the graph.

b) Graph Generation

The directed graph which will be used for further processing during the algorithm is generated using the data collected from the above data source. The representative nodes will be also decided in this phase. During the generation of the directed graph, the unique value for the representative nodes called the membership function value is selected and assigned to the nodes.

c) Centrality Calculation and Normalization

Using different graph centrality measures like Katz centrality, Degree centrality and Betweenness Centrality, the centrality value for the nodes are calculated and all the central representative nodes will be uncovered. Finally the values are normalized using min max normalization method so that all the centrality values for the nodes will be in the same range and so that calculations and manipulation will be easier. The concepts used in this module are explained below.

i) Betweenness Centrality

Betweenness centrality estimates the occasions a node lies on the most limited way between different nodes. This measure indicates which nodes go about as 'spans' between nodes in a system. It does this by distinguishing all the briefest ways and afterward checking how frequently every node falls on one. It is utilized for finding the people who impact the stream around a framework. Betweenness is helpful for examining correspondence elements, yet ought to be utilized with consideration. A high betweenness score could show somebody holds expert over, or controls coordinated effort between, different groups in a system; or demonstrate they are on the fringe of the two bunches.

ii) Degree Centrality

Degree centrality allots a significance score dependent on the quantity of connections held by every node. It tells what number of immediate, ‘one hop’ associations every node needs to different nodes inside the system. It is utilized for finding associated people, prominent people, people who are probably going to hold most data or people who can rapidly interface with the more extensive system. Degree centrality is the easiest proportion of node availability. Now and then it's valuable to take a look at in-degree and out-degree as particular measures, for instance when taking a look at value-based information or record movement. A system of psychological militants, over and again separated by degree uncovering bunches of firmly associated nodes is an example.

iii) Katz Centrality

Measures impact by considering the absolute number of strolls between a couple of nodes, dissimilar to common centrality estimates which consider just the most brief way. Used to gauge the general level of impact of a node inside an interpersonal organization. Every way or association between a couple of nodes is relegated a weight controlled by a weakening element α and the separation between nodes as α^{d} . Computes relative impact by estimating the quantity of the quick neighbors and furthermore all different nodes in the system that interface with the node under thought through these prompt neighbors.

iv) Min Max Normalization

Min max normalization is the method used to normalize the data value with a minimum and maximum range. Here it is used to normalize the centrality values between 0 and 1.
It is calculated using the formula as below:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

d) Graph Processing

To analyse the graph, to determine the membership function and to uncover prominent nodes, a number of steps are performed during the graph processing. The graph that is used here is processed up to n hops which can be manually specified during the start of the algorithm.

i) Predecessor Discovery

The set of predecessors of any node can be uncovered in predecessor discovery during any phase of the determination of prominent nodes so that the set of predecessors for a given set of nodes can be determined efficiently.

ii) Successor Discovery

The set of successors of any node can be uncovered in successor discovery during any phase of the determination of prominent nodes so that the set of successors for a given set of nodes can be determined efficiently.

iii) Sinistral Processing

The sinistral processing is done after the process of uncovering the predecessors and successors. Two types of nodes are processed here. The first category is the sinistral nodes which are those nodes that are the predecessors of representative set. The second category is the successors of the sinistral nodes. The process of calculating the intermediate values which is used for the final calculation is also performed here.

iv) Dextral Processing

The dextral processing is done after the process of uncovering the predecessors and successors. Two types of nodes are processed here. The first category is the dextral nodes which are those nodes that are the successors of representative set. The second category is the predecessors of the dextral nodes. The process of calculating the intermediate values which is used for the final calculation is also performed here.

v) Central Processing

After predecessor discovery, successor discovery, sinistral processing and dextral processing, a set of nodes will be obtained which are called central nodes. The central nodes are those nodes which becomes the candidate nodes for uncovering the final prominent nodes. The calculation of intermediate values of the central nodes for the membership function calculation at the end of algorithm is performed. Similarly, the correction coefficient value is also calculated for all the central nodes in the graph. This value is used by the next stage, so the value will be passed for calculating the final membership function value.

vi) Membership Function Calculation

The value of membership function is a significant factor which chooses the similitude of a node with another, since fuzzy set theory is accustomed to deciding the liking of nodes. The membership value is determined with the assistance of correction coefficient value determined in one of the past advances. The arrangement of competitor nodes alongside the determined membership function value is passed on to the following stage. At the point when the calculation begins, it chooses focal delegate nodes (central representative nodes) in graph utilizing the idea of graph centrality. Diverse graph centrality measurements are utilized together and the centrality esteems are determined and focal delegate nodes will be resolved.

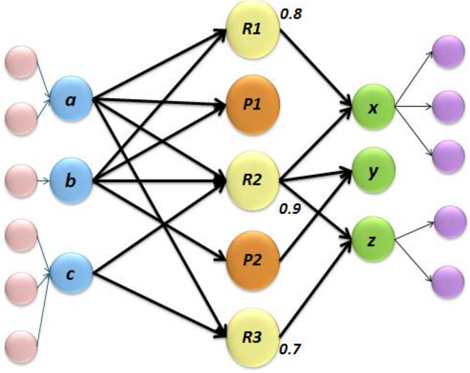

Fig 3 Graph with central representative nodes

A large number of the nodes in delegate set have some regular direct predecessors, which suggest that a considerable lot of these predecessors will have basic direct successors among the nodes in the representative set R(eg: R1, R2, R3 in the fig 3). Essentially, a large number of the nodes in representative set have some regular direct successors, which suggest that a significant number of these successors will have normal direct predecessors among the nodes in representative set R.

Since the immediate successors and direct predecessors in R are subset of every single direct successor and every immediate predecessor separately, two new sets are presented. This crossing point results in just those immediate successors of certain node x that are in representative set R and those immediate predecessors of a specific node x that are in representative set R too separately. For every node in these two sets, a value v is determined by summing membership function values of nodes in R which are immediate successors or predecessors to certain node relying upon in the case of preparing is done on sinistral or dextral side.

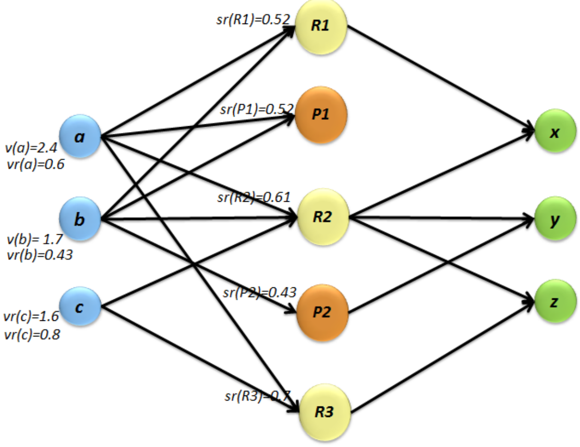

Fig 4 Sinistral Processing

In sinistral processing, the determined values are refined dependent on the outdegree of the nodes. For this the value v is divided with the outdegree of the relating node to get a refined value vr. For every one of the nodes in sinistral set, all its immediate successors are acquired, which are the nodes that are straightforwardly associated with it. The subsequent set is an association of every single direct successor for all nodes in sinistral set. It tends to be noted here this is a superset of R, since it will unquestionably contain every one of the nodes in the representative set, yet in addition different nodes that weren't set apart as representative initially. For every one of the nodes in this superset, a middle of the road value is determined by summing membership function values of nodes in the sinistral set which are immediate predecessors to certain node in the superset. Again these determined values are refined dependent on the indegree of the specific node to get a value sr for the nodes in the superset. The sinistral handling closes with the count of the sr value for the nodes.

Essentially, the determined values are refined dependent on the indegree of the nodes in the dextral processing. For this the value v is isolated with the indegree of the relating node to acquire a refined value vr. For every one of the nodes in dextral set, all its immediate predecessors are gotten, which are the nodes that are straightforwardly associated with it. The subsequent set is an association of every single direct predecessor for all nodes in dextral set. It tends to be noted here this is a superset of R, since it will surely contain every one of the nodes in the representative set, yet in addition different nodes that weren't set apart as representative initially. Likewise, for every one of the nodes in this superset, a middle of the road value is determined by summing membership function values of nodes in the dextral set which are immediate successors to certain node in the superset. Again these determined values are refined dependent on the outdegree of the specific node to get a value tr for the nodes in the superset. The dextral handling closes with estimation of the tr value for the nodes.

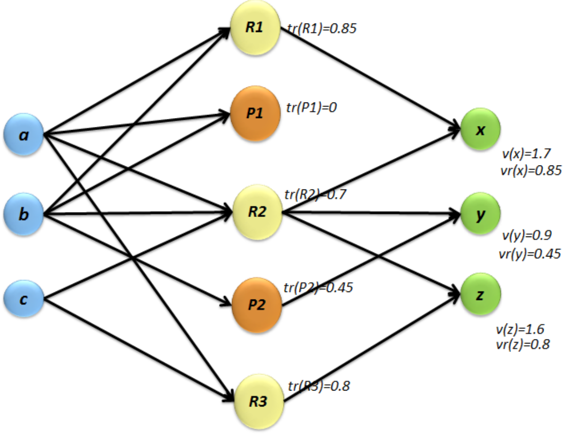

Fig 5 Dextral Processing

Subsequent to getting all the immediate successors of the sinistral set and all the immediate predecessors of the dextral set, and the count of sr and tr values, the central processing is performed. The principle reason for doing this processing is to get a refined and mean value for every one of the nodes who are the successors and predecessors of sinistral set and dextral set separately. For all the central nodes, a value st is determined. This value is the mean of sr and tr values got in the sinistral processing and dextral processing individually. The st value is resolved for nodes that are in representative set and furthermore for some newfound ones.

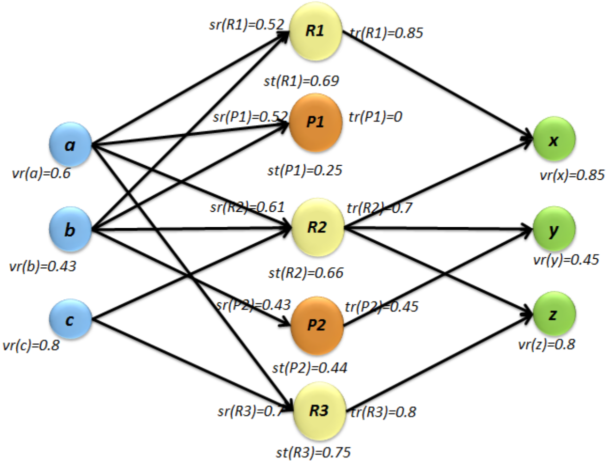

Fig 6 Central Processing

From this, at first the st value is checked for the majority of the nodes in the representative set so as to look at whenever determined similitude values compare to the initially appointed values. However, it was seen that the determined values are littler than the relegated values, which is a result of representative set being littler than normal in-level of the representative nodes. There can be situations where the determined values are only multiple times littler, however in bigger graphs with heaps of associations, this proportion may be an a lot bigger value. So as to address the count, another value called correction coefficient c is presented. This coefficient is really a mean proportion between the doled out value and the determined value.

$$C = \frac{\Sigma_{r\varepsilon R}\ \mu_R(r)\ /\ st(r)}{|R|}$$

The last proportion of similarity or the value of membership function is then determined by increasing the determined value with the correction coefficient c. The subsequent value srf is essentially a membership function value for every node in the central set, where nodes that have the littlest values of this number shouldn't show similitude with the component uncovered among nodes in R, and the other way around. The determined closeness measure is really a value of membership function that speaks to connection with the nodes in the representative set.

$$srf(x) = c * st(x)$$

Presently nodes from the central set with the most astounding calculated similarity can be chosen and analyze if the ideal component is available, as it was in representative set R. The outcomes may be helpful , as a portion of the nodes with the most elevated determined membership function values can be chosen and re-instate the entire calculation by adding these nodes to existing arrangement of the representative nodes. This ought to enable the calculation to repeat until certain criteria are met, for example in two consecutive iterations of the calculation no new representative nodes are found. Additionally, the determined values of membership function can be checked with the assigned values (in the initial step of the algorithm), which can prompt evacuation of a portion of the nodes from the representative set.

e) Discovered Influential Nodes

After the calculation of membership function, the arrangements of persuasive nodes are resolved dependent on the idea of fuzzy set theory. The nodes fulfilling the condition for being powerful are found and are given as the last yield of the whole procedure of influential nodes discovery.

f) Categorized Nodes

The final set of influential nodes discovered from the set of available nodes after N hop propagation is then categorized and the nodes based on category is obtained. The newly discovered influential nodes are used to fetch their information from the corresponding source. Here the example is Twitter, so Twitter user ids will be obtained and the user details are fetched using Twitter API calls. Later the collected details are analysed and the users are categorized based on the location of the users.

IV. Results and Discussion

A. Results

The proposed algorithm used Twitter data which contains nodes and edges information about the Twitter users from Kaggle repository. The algorithm is implemented in Anaconda(Python Distribution) using PyDev. NetworkX is used to study and analyse the graph and Twython is the API used to collect user information from Twitter. Windows 8.1 is the Operating System. The system in which it is implemented is a 64 bit system with Intel Pentium CPU N3520 @ 2.16GHz processor and 3.89 GB memory.

The space and time complexity of the algorithm is as follows.

Space Complexity: For a dataset of size 500 MB, the resulting output file would be of size 300 MB.

Time Complexity: O(n + p + s + r + x + d + c + l) where n : number of nodes, p : number of predecessors, s : number of successors, r : number of representative nodes, x : number of sinistral nodes, d : number of dextral nodes, c : number of central nodes, l : number of newly discovered nodes

The algorithm was utilized and for all intents and purposes confirmed on a subset of users of Twitter social community. Twitter is one of the greatest interpersonal organizations; however it began as a micro blogging administration. Twitter has in excess of 300 million dynamic users every month, which makes it second biggest social network, directly after Facebook. Twitter users tail others or are pursued. Not at all like on most online person to person communication locales, for example, Facebook or MySpace, the relationship of following and being pursued requires no response. A user can pursue some other user, and the user being pursued need not pursue back. Being a supporter on Twitter implies that the user gets every one of the messages (called tweets) from those the user pursues. The proposed algorithm is utilized to find similar influential users in Twitter. The procedure begins with discovering some representative users utilizing graph centrality measures and normalizing them using min max normalization. Influential users connect with their adherents with news and intelligent tweeting in numerous circles pursued by them. For every one of these users, a membership value is determined and decides how much certain user has a place with this representative gathering. It is conceivable that specific user has a place with some other representative gathering too, not only one. The table underneath demonstrates an example of influential representative users.

Table 5.1 Sample representative users

Screen Name	Area of Interest	Location
@parloursaloninc	Fashion	Toronto
@FRVRJT	Music	LA
@LasLad	Music	Rivercity
@kerrycalderon	Techno	Trinidad
@andhesonit	Sports	Skyloft
@tamedcabello	Games	Here
@halsey	Music	NIL
@katlaboriante	Music	Silvermoon
@estradalec	Games	NJ
@jvong1120	Fun	CA
@bestjustinpiccs	Photo	None
@extrafloral	Social	None
@katya_zamo	Business	Hong Kong
@EmileAbouSawan	Fun	Here

It is these set of representative users that are used in the algorithm for the sinistral, dextral and central processing. Also for the final calculation of coefficient and the membership function values, these sets of representative users are used.

The sample representative users can be drawn as a graph as shown below. Since drawing the entire graph is a time consuming process and cannot be plotted on a paper so easily, a small subset of the graph is taken as shown as a sample. This is because the graph file itself would be more than 2 GB in size and it is very difficult to make it to a graph. This is because the graph file not only stores the information regarding nodes and edges, but also many other information about the nodes, which includes the membership function value, their related field of interest, the information about if it is a representative node or not etc. The tabular result is sufficient to understand the algorithm, its implementation and results.

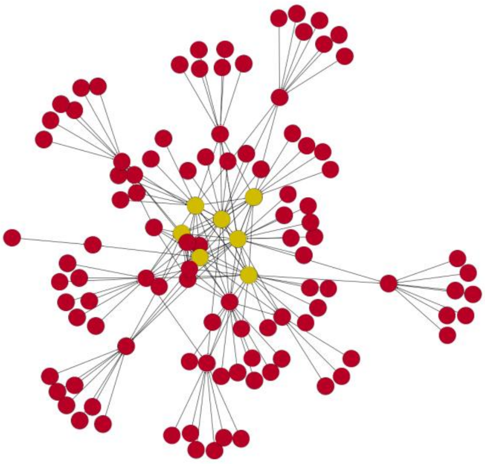

Fig7 Sample graph of representative users

It is important to note that lots of these followers are not unique for certain representative user. It is very likely that two Twitter users that promote a particular area and that have more than 1 million followers will have some (or many) common followers. This implies that a lot of these followers will follow more than just one user from the representative group. After fetching all the representative users' followers, for each of these followers a membership value that determines its affiliation with the representative group is calculated. So, for every follower, all of the users from the representative group are taken and their membership values are summed. Intuitively, this number can show the interest certain user has in this representative group of users.

The required value of membership function according to the algorithm is determined for every one of the nodes including the adherents, their companions, companions, their supporters and every one of the nodes in the representative set. Correction coefficient is likewise decided for the graph. Final similarity value is then determined by multiplying the correction coefficient with the recently determined relative similarity. This final similarity is to respect to Twitter users in the representative gathering, or all the more explicitly influential Twitter users. Here the algorithm stops and the calculated data is examined by looking at the users with the highest calculated value of influence. In this stage, only Twitter users that are not already in the representative group are interesting for examination – by this new influential Twitter users can be discovered. Apart from discovering new users, the results might show that certain Twitter users that were originally put in the representative group are actually not that influential at all. This gives expert an opportunity to revise and review the data as it may be not correct. The table below shows a sample of newly discovered users.

Table 5.2 Newly discovered influential users

Screen Name	Area of Interest	Location
@Andiluv	Fashion	Rivercity
@tam_rapp	Entertain	Trinidad
@bestjustinpiccs	Photo	Skyloft
@KLitzau	General	None
@extrafloral	Social	Here
@Hawkeye_1986	Sports	Silvermoon
@HannahNicoleT01	Music	NJ
@parloursaloninc	Fashion	CA
@ssudhirkumar	Writer	Hyderabad
@andhesonit	Sports	None

The example influential users can be drawn as a graph as demonstrated as follows. Since illustration the whole graph is a tedious procedure and can't be plotted on a paper so effectively, a little subset of the graph is taken as appeared as an example. This is on the grounds that the graph document itself would be in excess of 2 GB in size and it is hard to make it to a graph. This is on the grounds that the graph record not just stores the data in regards to nodes and edges, yet in addition numerous other data about the nodes including the membership function value, their related field of intrigue, the data about on the off chance that it is a representative node or not and so on. The tabular result is adequate to comprehend the algorithm, its execution and results.

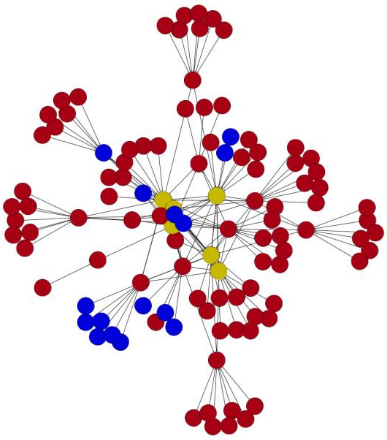

Figure 8 Sample graph of influential users

The principle bit of leeway of this algorithm is that despite the fact that there are various services accessible so as to decide the user impact in social network, a large portion of them are paid services or the services are not totally accessible for the free users. Likewise these services don't demonstrate the algorithms and strategies they have utilized. The algorithms utilized by them were stayed discreet .accordingly the specialists who need to do look into in the field of node similarity needs to build up their own algorithms for their motivation since a significant number of these superior services are not reasonable to them. So by executing this algorithm, individuals can utilize it for finding influential users in a network.

At long last these influential users are ordered dependent on their location. The final arrangement of influential nodes found from the set of accessible nodes after N hop propagation is then ordered and the nodes dependent on location is gotten. The newfound influential nodes are utilized to get their data from the corresponding source. Here the model is Twitter, so Twitter user ids will be gotten and the user details are brought utilizing Twitter API calls. Later the gathered details are dissected and the users are sorted dependent on the location of the users.

Time consumption is a noteworthy issue while managing the algorithm. In any case, in the event that the exactness could be relinquished, at that point time and memory can be spared somewhat. So a lot of new affecting users can be acquired from the bit of downloaded supporters. Likewise, the found users can be added to the representative set and after that the algorithm can run once more, finding much more users. Since numerous different highlights like user movement (number of tweets), adherents' commitment on user action (number of retweets, top picks) and so forth has not been considered in this algorithm, results won't be 100 percent exact.

B. Inferences

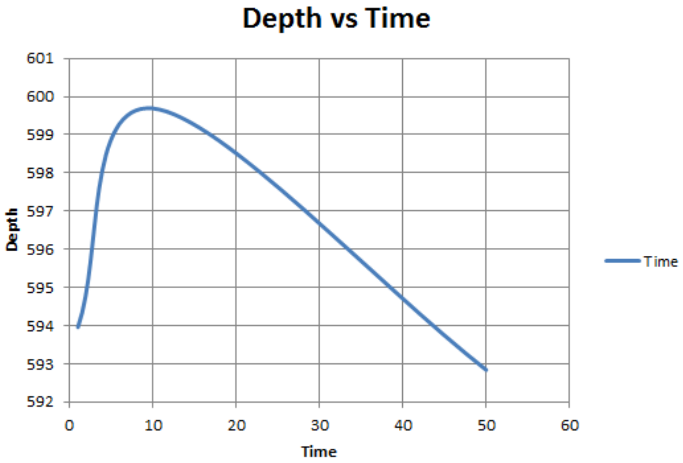

Fig 9 Depth vs. Time

The Depth vs. Time graph appeared in the figure 9 above delineates the time taken to finish the algorithm for expanding depth. The depth implies the quantity of hops to be taken when the successors and predecessors of a node is considered. At first, the time increments when the depth increments since when depth is expanded, the number of nodes to be prepared by the graph likewise increments. In any case, after a specific time, the time taken by the algorithm begins diminishing notwithstanding when the depth is expanding. This is on the grounds that after quite a while, the graph has officially gathered the data about numerous nodes thus it requires less investment in further processing.

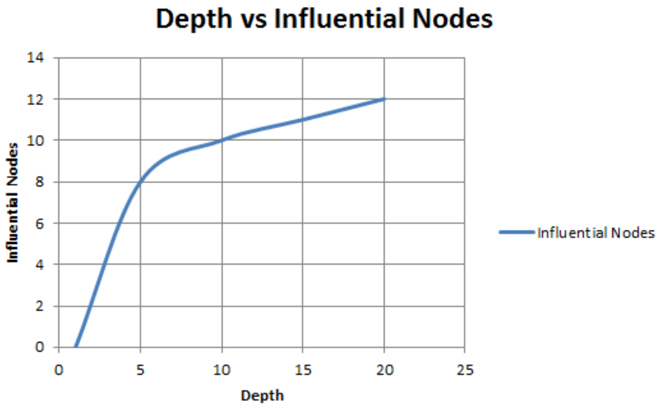

Fig 9 Depth vs. Influential Nodes

The Depth vs. Influential Nodes graph shown in the figure 9 depicts the count of influential nodes being discovered when the depth is increased. The main feature that can be analysed is that there were many nodes that could not be discovered when the depth was very less or when the depth was 1. The depth of 1 means that depth is not considered, it works as just a simple graph. When the depth is taken into consideration, even though the time may be a bit higher, this can be tolerated since more number of resulting output can be obtained.

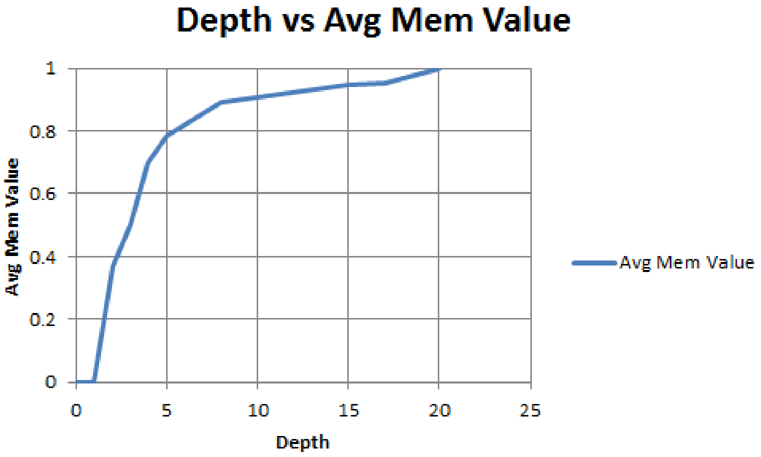

Fig 10 Depth vs. Avg Mem Value

The Depth vs. Avg Mem Value graph shown in the figure 10 depicts the average of the membership function value of the influential nodes being discovered when the depth is increased. The main feature that can be analysed is that there were many nodes that could not be discovered when the depth was very less or when the depth was 1. The depth of 1 means that depth is not considered, it works as just a simple graph. When the depth is taken into consideration, even though the time may be a bit higher, this can be tolerated since more number of resulting output can be obtained. As a result the average value also increases accordingly.

V. Conclusion and Future Scope

A. Conclusion

There are a number of services available in order to determine the user influence in social network, most of them are paid services or the services are not completely available for the free users. Also these services do not show the algorithms and methods they have used. The algorithms used by them were kept secret. As a result the researchers who need to do research in areas related to node similarity has to develop their own algorithms for their purpose since many of these premium services are not affordable to them. In the proposed algorithm, graphs are used to represent structural information in the data. Graph centrality is a proportion of impact of a node in the network which recognizes the most significant vertices of a graph. The vertices of the graph are described dependent on a genuine valued function that gives a positioning to the nodes. Many related works have been broke down and Eigenvector centrality, Katz centrality and Betweenness centrality have been recognized as the significant ones and are utilized to discover central representative nodes.

A directed graph is utilized here to execute the algorithm. Additionally the idea of fuzzy set is utilized to compute a membership function for every node in the graph. The algorithm utilizes N hop propagation in the graph to consolidate more nodes in the outcome. The users in these social networks can be spoken to as nodes in graphs. An algorithm is proposed here for finding similar nodes in directed graphs which uses the thought: graph nodes that are directed towards another with similar neighborhood ought to be similar. The algorithm is checked on a subset of Twitter social network contextual investigation by finding influential Twitter users utilizing Graph Centrality measures and Min Max Normalization. The followers and friends of the representative nodes are considered in the algorithm with the goal that the impact of a specific client can be all the more exactly decided. The influential nodes were found utilizing the algorithm and they have been arranged dependent on the location amid the last stage.

B. Future Enhancements

Remembering that impact in social media is dynamic and unclear idea, it very well may be refined by arrangement of measures: client movement (number of tweets), followers engagement on user activity (number of retweets, favourites),user zone of intrigue and so forth. Likewise, every one of these measures change after some time, and these progressions could be followed excessively so as to give some profitable pattern data. By taking a portion of these measures in record to certain Twitter user's impact, the entire algorithm could yield better outcomes.

Another refinement could be the iterative execution of the algorithm. In the event that a few criteria was presented that could reject users that were initially set apart as influential, however that ended up being false. It could likewise prompt development of the representative set by including new influential users over longer timeframe. Likewise, it could be helpful to see whether there exists any relationship between's the determined value of similarity (membership function) and number of supporters, number of tweets, user commitment, the field of enthusiasm of user and so forth.

The utilization of Natural Language Processing to discover the user region of intrigue would yield precise outcomes if the complexity can likewise be overseen alongside the present algorithm. There are great deals of Natural Language Processing algorithms accessible which are mind boggling in nature. So it would be great if these algorithms can be utilized or changed or joined with the algorithm proposed here. Likewise it would be better if live Twitter information gathering could be consolidated and furthermore devise some approach to deal with as far as possible forced by the Twitter API in gathering enormous sum user data. The algorithm would turn out to be better in the event that in the wake of giving a point or watchword and, at that point dependent on this data; the influential nodes can be found.

References

[1]. Anjana N Iyyer, Mary John, "*A Fuzzy Based Approach for Discovering Similar Nodes in Directed Graph*", 8th International Conference on Advances in Computing & Communications (ICACC), September 2018.

[2]. Mohammad Alshahrani , Zhu Fuxi, Ahmed Sameh, Soufiana Mekouar, Sheng Huang, "*Top-K Influential Users Selection Based on Combined Katz Centrality and Propagation Probability*", 3rd IEEE International Conference on Cloud Computing and Big Data Analysis, April 2018.

[3]. Paramita Dey, Sarbani Roy, "*Centrality based information blocking and influence in social network*", IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), December 2017.

[4]. Ashish Mehrotra, Mallidi Sareddy, Sanjay Singh, "*Detection of Fake Twitter Followers using Graph Centrality Measures*", 2nd International Conference on Contemporary Computing and Informatics (IC3I), May 2017.

[5]. Prantik Howlader, Sudeep KS, "*Degree Centrality, Eigenvector Centrality and the relation between them in Twitter*", IEEE International Conference On Recent Trends In Electronics Information Communication Technology, May 2016.

[6]. Mahdieh Ghasemi, Hossein Seidkhani, Faezeh Tamimi, Maseud Rahgozar, Ali Masoudi-Nejad, "*Centrality Measures in Biological Networks*", Current Bioinformatics Journal, 2016, Vol. 9, No. 4.

[7]. Endre Pap, Marko Jocic, Aniko Szakal, Djordje Obradovic, Zora Konjovic,"*Managing big data by directed graph node similarity*",17th IEEE International Symposium on Computational Intelligence & Informatics,17–19 November, 2016, Budapest.

[8]. L. A. Zadeh, "*Fuzzy sets*" , Inf. Control, vol. 8, pp. 338–353, 1965.

[9]. D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "*Influence and passivity in social media*," in Machine learning and knowledge discovery in databases, Springer, 2011, pp. 18–33.

[10]. A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "*Effects of user similarity in social media*," in Proceedings of the fifth ACM international conference on Web search and data mining, 2012, pp. 703–712.

[11]. J. Leskovec, A. Singh, and J. Kleinberg, "*Patterns of influence in a recommendation network*," in Advances in Knowledge Discovery and Data Mining, Springer, 2006, pp. 380–389.

[12]. M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "*Measuring User Influence in Twitter: The Million Follower Fallacy*," ICWSM, vol. 10, pp. 10–17, 2010.

Website References:

[13]. "Social Media Marketing, Statistics & Monitoring Tools," Socialbakers.com. [Online]. Available: http://www.socialbakers.com/. [Accessed: 28-Oct-2018].

[14]. "Simply Measured | Easy Social Media Measurement & Analytics," Simply Measured. [Online]. Available: http://simplymeasured.com/. [Accessed: 28-Oct-2018].

[15]. "Social Media Monitoring Tools & Sentiment Analysis Software," Trackur. [Online]. Available: http://www.trackur.com/. [Accessed: 28-Oct-2018].

[16]. "Klout | Be Known For What You Love," Klout. [Online]. Available: https://klout.com/home. [Accessed: 28-Oct-2018].