IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Mining Flow of Information Among User Networks in Social Networking Services

[1]Ankishetty Sushma, [2]N.Naveen Kumar

[1]Student, [2]Associate professor

[1]Software Engineering,

[1]School of Information Technology, Hyderabad, India

Abstract: Social networking services is a web based platform which is used by the people to evolve social networks or social relationships with each other who share similar profession or individual interests, activities, events, ideas or real life connections. The popularity of SNS is emerging rapidly and has become a huge date source to analyze social networks. The widespread content by users is very important in social networking services. In this, I proposed a method for finding information diffusion models for the data in social networking sites which can be obtained by using process mining techniques .For this, first we filer the information by expelling inconsistent users. Further, to decrease the complexity nature of SNS filtered information a user-centric clustering approach accomplished utilizing a novel measure called user intimacy value that quantifies the degree of relationship between users ,depending on modularity maximization strategy is carried out. This clustered information is converted to an event log which is utilized as input for process discovery algorithm Infoflow miner. The Infoflow miner utilizes response weight, that can be characterized as, the degree of impact or effect that one user activities have on other user activities. Final outcome of the technique is a graph which represents the flow of information among user networks present in the original information.

Index Terms **- process mining,Information flow, social networking services.**

I. INTRODUCTION

SNS (Social Networking Services) have spread internationally moreover produce massive measures of information consistently. Plenty of individuals throughout the globe associate with companions, family, colleagues, classmates, and associates through these administrations. For instance, Twitter had 3286+ million month to month dynamic clients in June 2017. Facebook revealed a normal of two billion monthly operative users by 30 June, 2017. In the event that we add the above users to the quantity of users of different social networking sites, for example, Google+, LinkedIn, or Instagram, the count increments significantly. Social networking sites provides clients an opportunity to make a profile to share substance, for example, messages, assessments, photographs, and recordings. Besides, SNS permit users to make social connections by associating with different users. The nature or definition of association fluctuates between various social networking sites. For example, Facebook and twitter as models, an association in Facebook is called relationship, among two users consenting to set up the social relationship. Conversely, the association on Twitter is spoken to by an after activity, where followed users don't need to favor the relationship, and the followed users don't need to follow their supporters.When any user posts or distributes content, different users can associate utilizing remarks or spread the substance utilizing various components of the social networking services. for instance, shares on Facebook and again tweets on twitter. At that time when those activities are rehashed ceaselessly, different procedures happen among the users, for example, data spread. Thus, there are huge measures of information produced each second from social networking sites, that may have hidden user cooperation's which can be found.

1.1 EXISTING SYSTEM

Process mining is a analytical discipline that has increased more consideration in the course of the most recent decade. It is based upon information mining standards and manages business execution information to improve today's information systems .process mining, mines a process model using a stored event log files , for example, enterprise execution records .The information in social networking sites, from the start not reasonable for process mining. By using a few preprocessing techniques, the information is adjusted and changed as a process mining input.

1.1.1 DISADVANTAGES OF EXISTING SYSTEM

- The information generated and spread by the users is crucial in SNS, and has noisy data which is not appropriate for mining visually.
- Most of the present solutions are related to static networks, and neglect that networks are dynamic and evolving

1.2 PROPOSED SYSTEM

A method for finding information diffusion models for the data in social networking sites can be obtained using process mining techniques .For this, first we filer the information by expelling inconsistent users. Further, to decrease the complexity nature of SNS filtered information a user-centric clustering approach accomplished utilizing a novel measure called user intimacy value that quantifies degree of relationship between users ,depending on modularity maximization strategy is carried out. This clustered information is converted to a event log which is utilized as input for process discovery algorithm Infoflow miner. The Infoflow miner utilizes response weight, that can be characterized as, the degree of impact or effect that one user activities has on other user activities. Final outcome of the technique is a graph which represents the flow of information among user networks present in the original information

1.2.1 ADVANTAGES OF PROPOSED SYSTEM

- The complexity of noisy SNS data is reduced
- Identify most active users
- Identify the relationship level between users.

II. SYSTEM REQUIREMENTS

2.1 HARDWARE REQUIREMENTS

- Processor : Core - i5
- RAM : 256 MB
- Hard Disk : 20GB

2.2 SOFTWARE REQUIREMENTS

- Coding Language : Java
- IDE : Eclipse
- Operating System : Windows 10

III. RELATED WORK

Finding information disseminators and receptors in online social media [1]

Today, there is critical sharing of data among clients on different internet based life destinations, including Digg, Twitter and Flickr. A fascinating outcome of such rich and broad social collaboration is the developing idea of "jobs" that are obtained by clients after some time, with regards to variegated correspondence exercises, for example, remarking, answering, transferring a media , etc. In this paper, we explore the revelation of two jobs that characterize data dispersal: disseminators and receptors. We propose a computational structure dependent on factorization of stacked portrayal of exercises and test the results on a dataset from Digg. Trials show that our methodology can, strikingly, uncover relationships with client exercises happening at a future point in time.

In this paper, we have proposed a computational structure to find data disseminators and receptors in multi-dimensional action based online life. Our technique defined data dispersal and utilization as double of one another and utilized a non-negative grid factorization based structure to decide scalar proportions of dissemination and utilization for every client in the system, over activities, remarks, answers or transferring comparable data. Examinations on a huge Digg.com dataset show that our found data jobs can yield high relationship with correspondence exercises of the clients after some time.

Fast unfolding of communities in large networks [2]

We propose a straightforward technique to remove the network structure of enormous systems. Our technique is a heuristic strategy that depends on modularity optimization. It is appeared to beat all other known network identification techniques as far as calculation time. Also, the nature of the networks identified is excellent, as estimated by the modularity. This is indicated first by distinguishing language networks in a Belgian cell phone system of 2 million clients and by breaking down a web chart of 118 million hubs and more than one billion connections. The precision of our calculation is likewise checked on specially adhoc modular networks.

We have presented a calculation for improving measured quality that permits us to contemplate systems of exceptional size. The restriction of the technique for the trials that we performed was the capacity of the system in fundamental memory instead of the calculation time. This difference in scales, for example from around 5 millions hubs for past strategies to in excess of 100 million hubs for our situation, opens energizing points of view as the measured structure of complex frameworks, for example, entire nations or enormous pieces of the Internet would now be able to be disentangled. The exactness of our technique has likewise been tried on specially appointed measured systems and is demonstrated to be magnificent in examination with other (much more slow) network discovery strategies. It is intriguing to take note of that the speed of our calculation can in any case be considerably improved by utilizing some straightforward heuristics, for example by halting the primary period of our calculation when the increase of measured quality is underneath a given edge or by expelling the hubs of degree 1 (leaves) from the first system and including them back after the network calculation. The effect of these heuristics on the last

segment of the system ought to be concentrated further, just as the pretended by the requesting of the hubs during the primary period of the calculation.

Workflow mining dicovering process models from Event logs

Contemporary work process the executives frameworks are driven by express procedure models, i.e., a totally determined work process configuration is required so as to order a given work process. Making a work process configuration is an entangled tedious procedure and, normally, there are disparities between the real work process forms and the procedures as saw by the administration. Consequently, we have created procedures for finding work process models. The beginning stage for such strategies is a supposed "work process log" containing data about the work process as it is really being executed. We present another calculation to extricate a procedure model from such a log and speak to it as far as a Petri net. Nonetheless, we will likewise exhibit that it is beyond the realm of imagination to expect to find discretionary work process forms. In this paper, we investigate a class of work process forms that can be found. We show that the -calculation can effectively mine any work process spoke to by a supposed SWF-net.

In this paper, we tended to the work process rediscovery issue. This issue was planned as follows: "Discover a mining calculation ready to rediscover an enormous class of sound WF-nets based on complete work process logs." We introduced the calculation that can rediscover a huge and significant class of work process forms (SWF-nets). Through models, we additionally indicated that the calculation gives fascinating examination results to work process forms outside this class. As of right now, we are improving the mining calculation with the end goal that it can rediscover a much bigger class of WF-nets. We have handled the issue of short circles and are presently concentrating on concealed assignments, copy errands, and progressed directing develops. Nonetheless, given the perception that the class of SWF-nets is near the furthest reaches of what one can do accepting this thought of fulfillment, new outcomes will either give heuristics or require more grounded ideas of culmination (i.e., more perceptions).

IV. PRODUCT DESIGN

A. UML DIAGRAMS:

4.1 UseCase Diagram

Usecase diagram is a UML diagram that indicates the relationships between users, actors and systems to catch the requirements of system, validate system architecture, specifies the context of a system. Use case diagram consists of actors, use-cases, communication link and the boundary of the system. Actor is the one who interacts with the system or usecases , here actor is the user(ourselves) and usecase generally represents the functionality of the system(process- automated or manual), here usecases are upload facebook data, preprocess and clustering, intimacy matrix, information flow miner,response weight graph.Communication link is the solid line connecting an actor to a usecase.

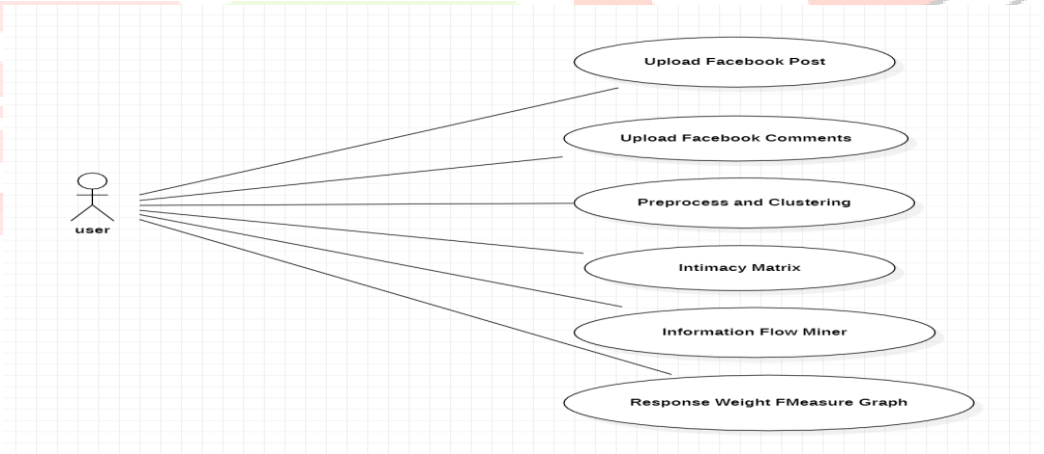


Fig -1: UseCase diagram

4.2 Sequence Diagram

Sequence diagram is a UML diagram that indicates the flow of actions among users here, among user and the system. The user interacts with the system as shown below.First, he/she uploads data consisting of Facebook post, after posts data are successfully loaded, upload Facebook comments data, next comments and posts is pre-processed and clustered into groups by removing noisy users. Then, intimacy matrix is calculated to know the closeness between users. The matrix displays the data of users commenting on posts, then apply information flow miner to get the information between two users and their weights. Lastly, a graph is generated.

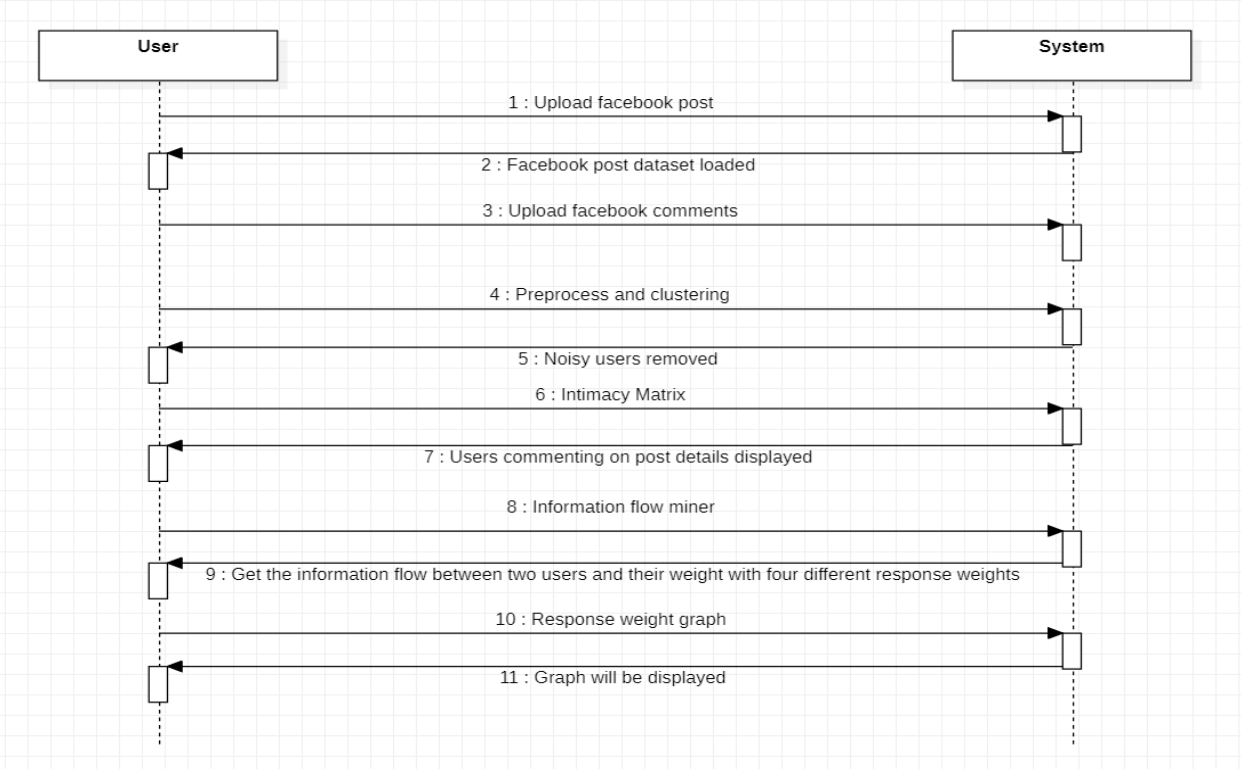


Fig -2: Sequence diagram

4.3 Activity Diagram

Activity diagram is a UML diagram which is essentially an improved form of flow chart representing flow from one action or activity state to another state .Unlike flowchart ,it consists of initial state which is a black filled circle as shown below. Action or activity states which is represented as rectangle with rounded corners, and control flow referred to as paths which represents transition from one action or activity state to another state. Decision node and branching used to make decision before determining the flow of control. The activity diagram for information flow miner is as shown below:

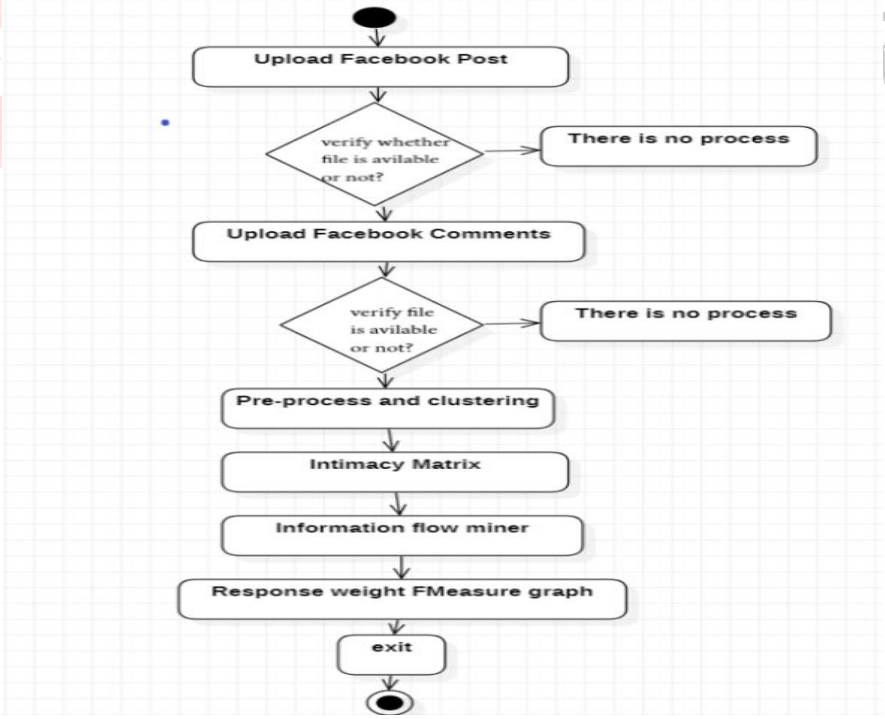


Fig -3: Activity diagram

4.4 DataFlow Diagram

Dataflow diagrams is a graphical form of logical flow of information in the system I.e; between a system and components of a system. It consists of process which receives input and produces output and generally represented in the form of rounded rectangle and, the other one is data-flow which is a path for data to move from one component to other and represented as straight lines with either incoming or outgoing arrows. The data flow diagram for information flow is as follows: user and system are two processes and arrows represent the data-flow between them.

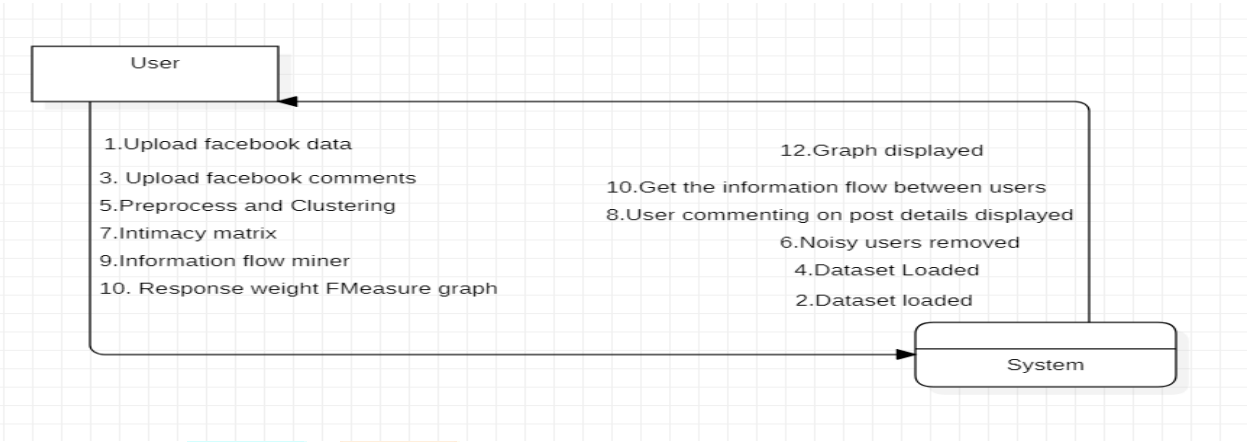


Fig -4: Dataflow diagram

B. SYSTEM ARCHITCTURE

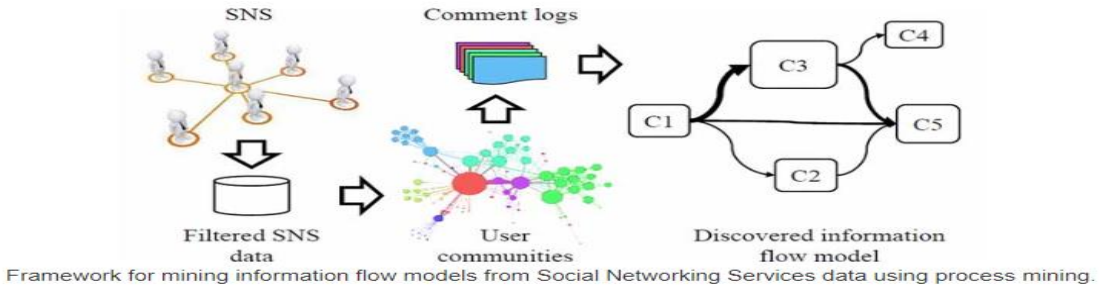


Fig -5: System Architecture

V. IMPLEMENTATION

5.1 Process Mining Techniques

Process mining is a systematic control for discovering, monitoring, and improving real processes by extracting knowledge from event logs readily available in today's information systems.Process mining offers objective, fact-based insights, derived from actual event logs, that help you audit, analyze, and improve your existing business processes by answering both compliance-related and performance-related questions.

The stages of the process mining technique are (1) planning and (2) extraction, during which data is extracted.After the first two stages, one or more analysis iterations are performed, possibly in parallel. In general, each iteration executes the following stages one or more times: (3) data processing, (4) mining & analysis, and (5) evaluation.If the findings are satisfactory then they can be used for (6) process improvement & support.

5.2 Algorithm Description

The methodology is build using "divide and conquer strategy" I.e; in user-centric approach this means to group types of users. For social networking services information, a process comprises of collection of posts by users, where every post consists of user activities like comments and likes. Comment log from SNS data is input to the algorithm and ouput is weighted graph .Information flow miner is used for discovering process models from SNS data.

Assign clustered user or user group names as vertices to the graph I;.e v^f,then empty response weight matrix M is created which is used to measure the closeness between users. Traverse the comment log file L entirely and for every comment in comment sequence sequence in

log file L, calculate response weight using four different formulas and update the matrix with the calculated weights. These values are weighted edges to the users and represented in the graph.

Input: Comment log L
Output: Weighted Graph $G^f = (V^f, E^f)$
1: Assign the cluster of user names to V^f
2: Create an empty response weight matrix $\boldsymbol{M}$
3: **for** every comment sequence σ_n in L **do**
4: **for** every comment a_k in σ_n **do**
5: **for** every antecedent comment a_l in σ_n **do**
6: calculate response weight w_{kl}
7: update M with w_{kl} for comments a_k and a_l
8: **end for**
9: **end for**
10: **end for**
11: Assign response weights in $\boldsymbol{M}$ to E^f
12: **return** $G^f = (V^f, E^f)$

Fig -6:Algorithm

5.3 Methodologies

- Pre-process: using this technique we will remove noisy data such as those users who are not frequent. To discover information diffusion we need to have active users on SSN networks. So infrequent users will be removing out.

- User Intimacy or Community Detection: In this module we will filter out entire dataset to look for users who are commenting to each other. For example to calculate User intimacy suppose, if two users frequently comment in the same post, then their intimacy value increases. Under this assumption high intimacy value will be assign and both users belongs to same community. This step is also called as MODULARITY MAXIMIZATION

- Process Discovery or Information Flow: In this module we can find information flow between two users by generating a matrix. If two users comment on same post then their count will be assign to matrix and these steps continue till all users comments matrix filled up. This matrix can be used to calculate response weight based on commenting on post, if two users are more active and comments on same post then their response weight will be high.

VI. TESTING AND RESULTS

The procedure or strategy for discovering errors or defects in an application or software program with the goal that the application functions as indicated by the end client's prerequisite is called testing. The test case is defined as a set of conditions or factors under which a tester will decide if a system or application under test satisfies requirements and works properly.

Following are the testing strategies followed during the test period:

TEST CASES:

Table -1: Test cases

Test Case Id	Test Case Name	Test Case Desc.	Test Steps			Test Case Status	Test Priority
			Step	Expected	Actual		
01	Upload Facebook Post	Verify the dataset is available or not	If it is available	We can upload	Dataset loaded	High	High
02	Upload Facebook Comments	Verify the dataset is available or not	If it is available	We can upload	Dataset loaded	low	High
03	Pre-process & Clustering	Verify both the datasets loaded are not	If it is loaded	We can process & cluster the data	Noisy users removed	Medium	High
04	Intimacy Matrix	Verify noisy users removed or not	If noisy users removed	We can apply the intimacy matrix on the data	We can see two users id are commenting on same post and count is also give	High	High
05	Information Flow Miner	Verify intimacy matrix applied or not	If applied	We can find out the information flow	we can see information flow between two users and their weight with four different response weights	High	High
06	Response Weight Fmeasure Graph	Verify all the operations completed or not	If it is completed	We can get the graph	Response weight of four formulas was displayed in a graph	High	High

RESULTS:

First, we will upload the file containing facebook posts data as shown below:, where we can see post id, user id, post data after successful upload of a file.

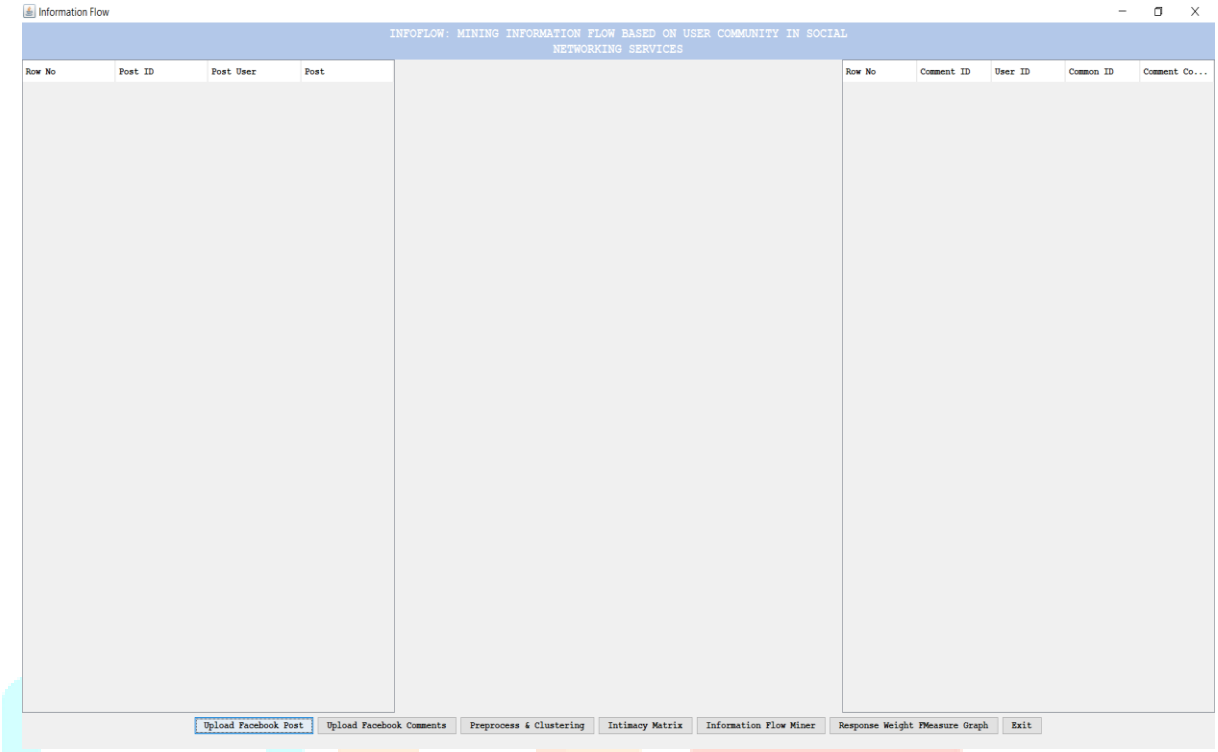


Fig -7:Main screen

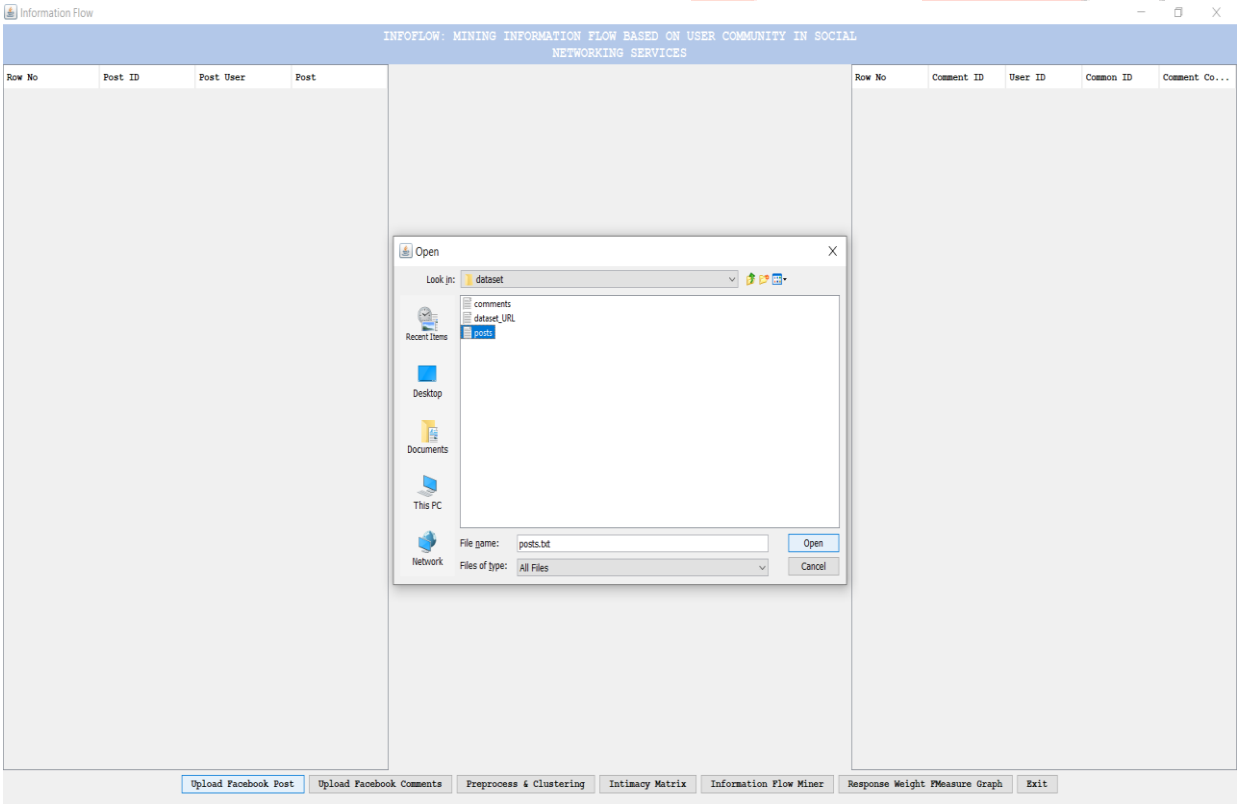


Fig -8:upload facebook post

After uploading the dataset will get the below screen:

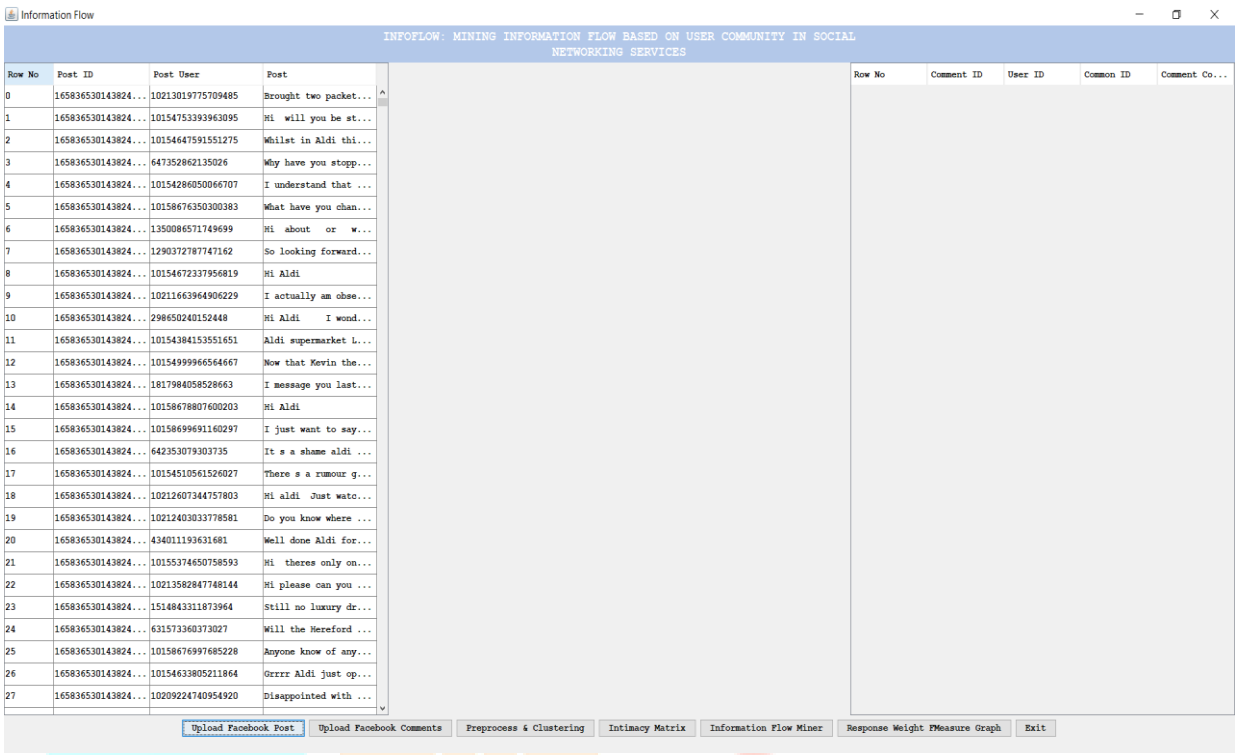


Fig -9: After posts uploaded screen

In above screen we can see post id, user id and post data. Now click on 'Upload Facebook Comments' button to upload comments. By analyzing posts and comments on same id we can detect two users are commenting on same post and can be consider as information flow between them

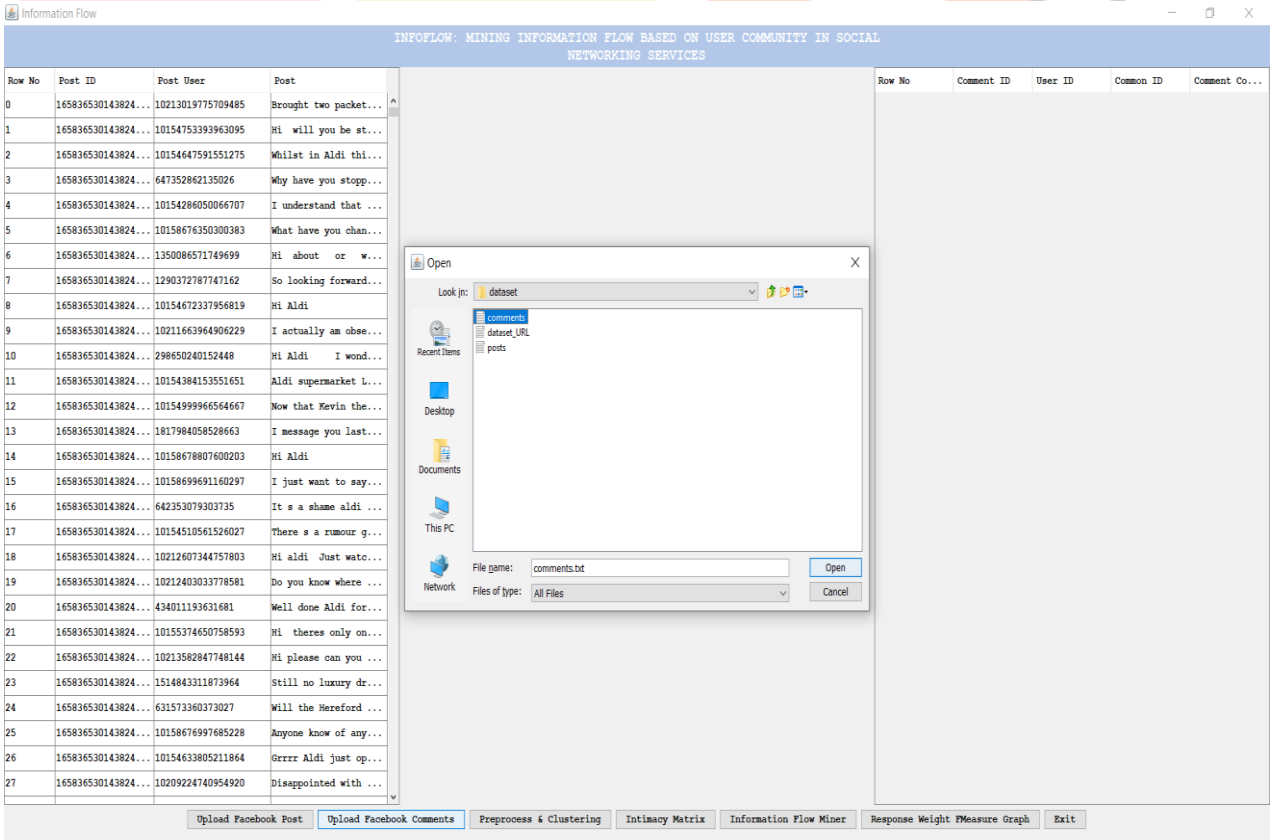


Fig -10: Upload facebook comments

In above screen we are uploading comments dataset and now will get below screen

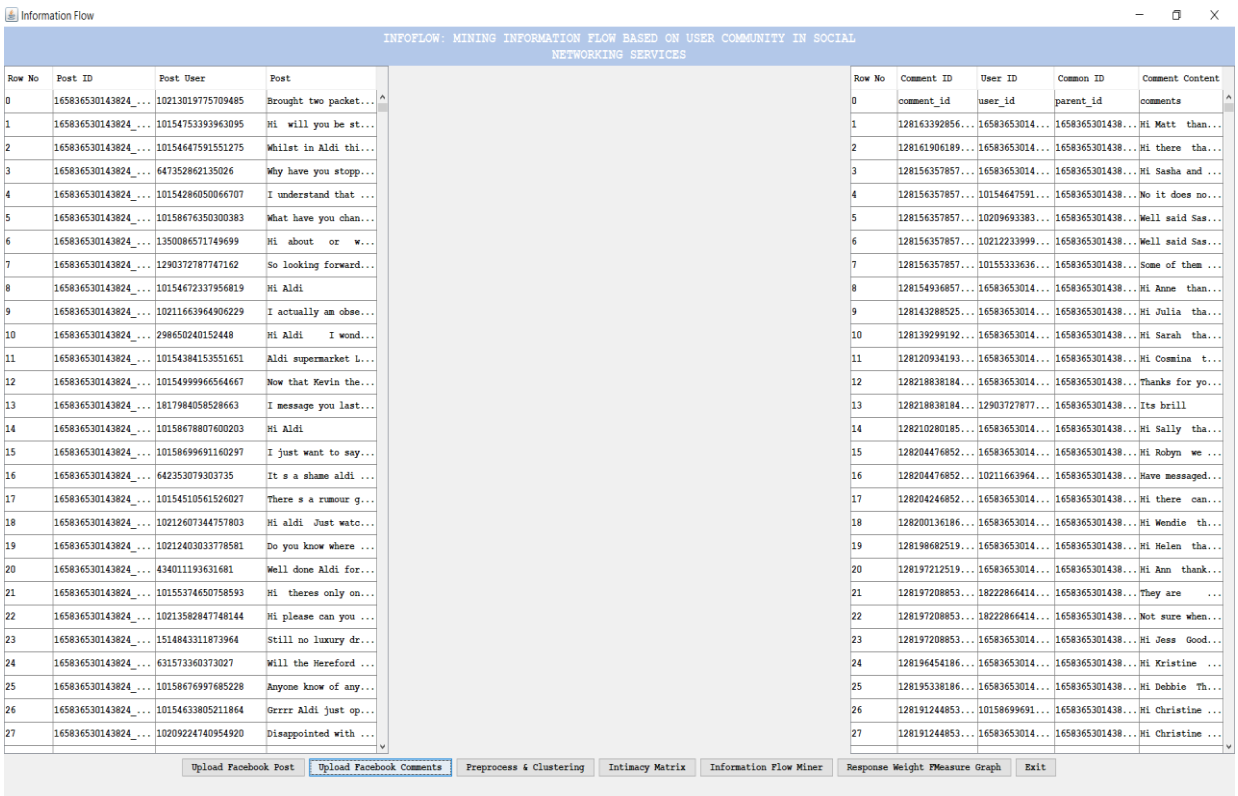


Fig -11: After uploading comments screen

Now click on 'Pre-process & Clustering' button to remove noisy users which means users who are infrequent

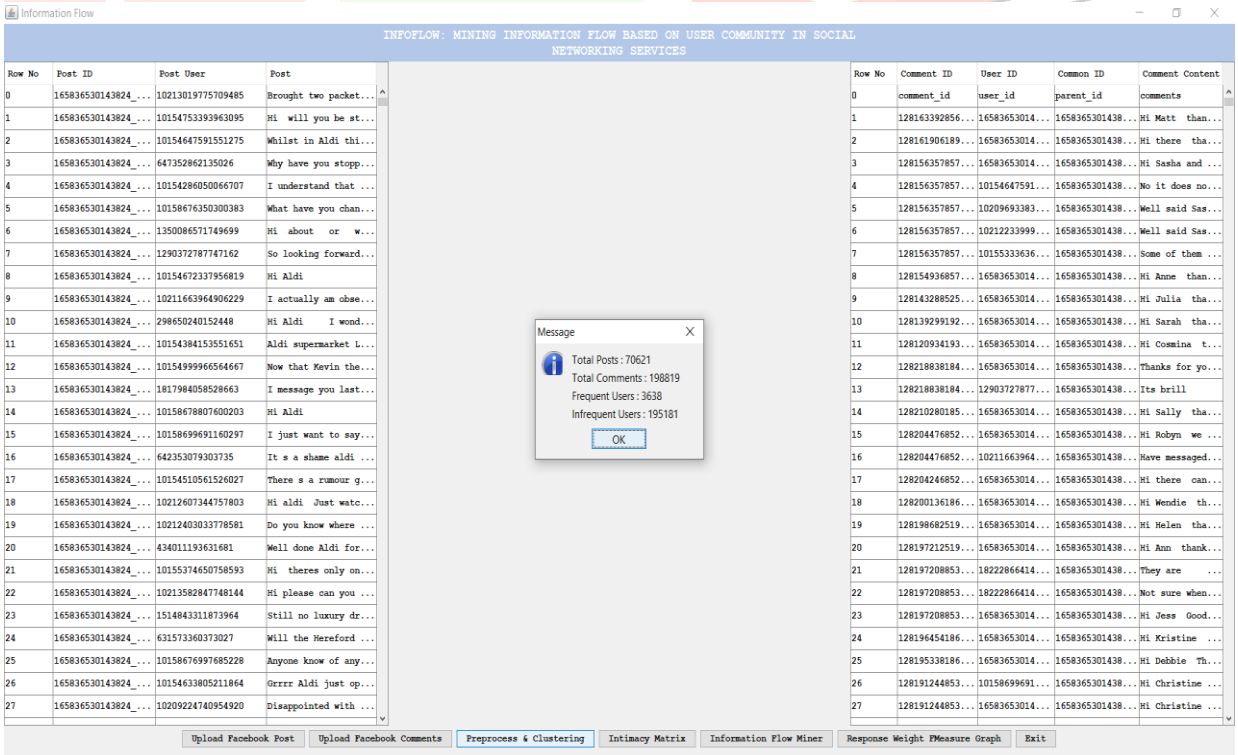
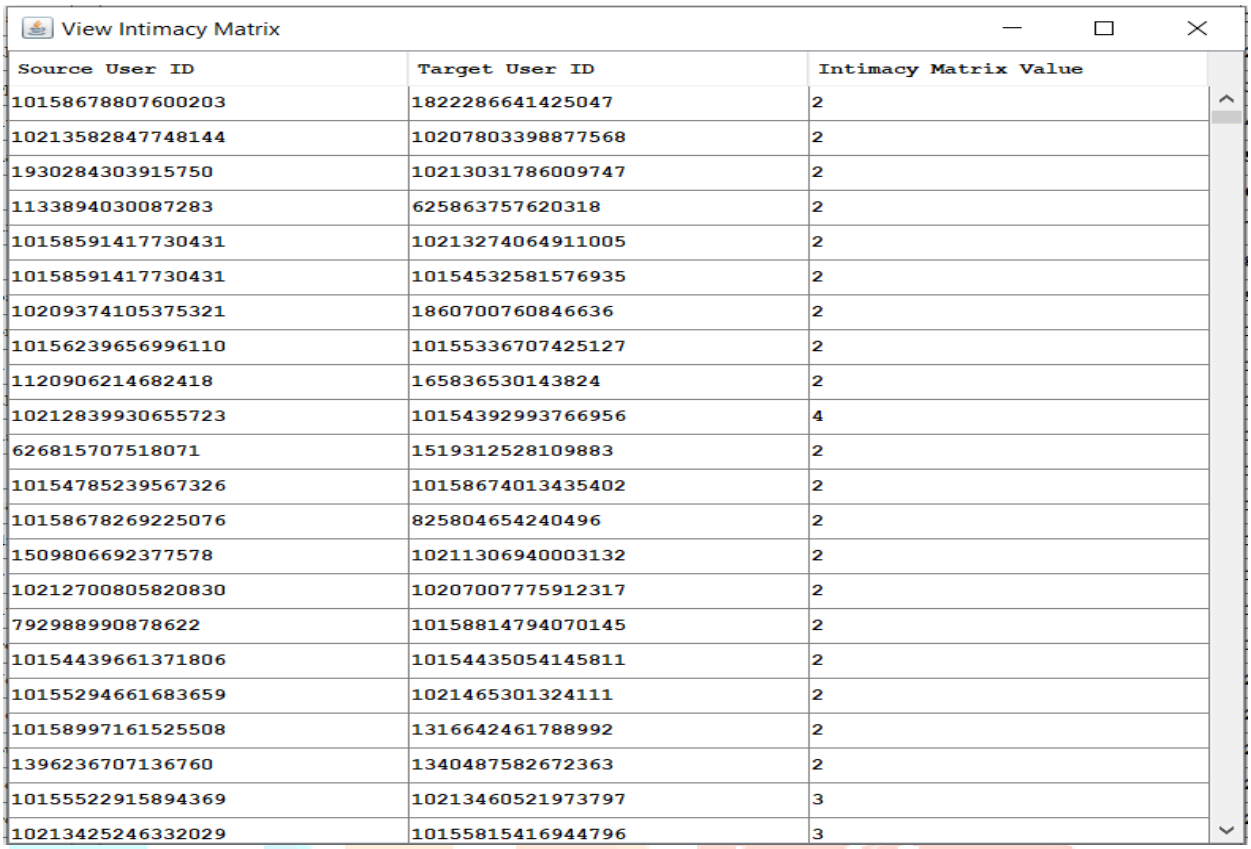


Fig -12: Preprocess and clustering

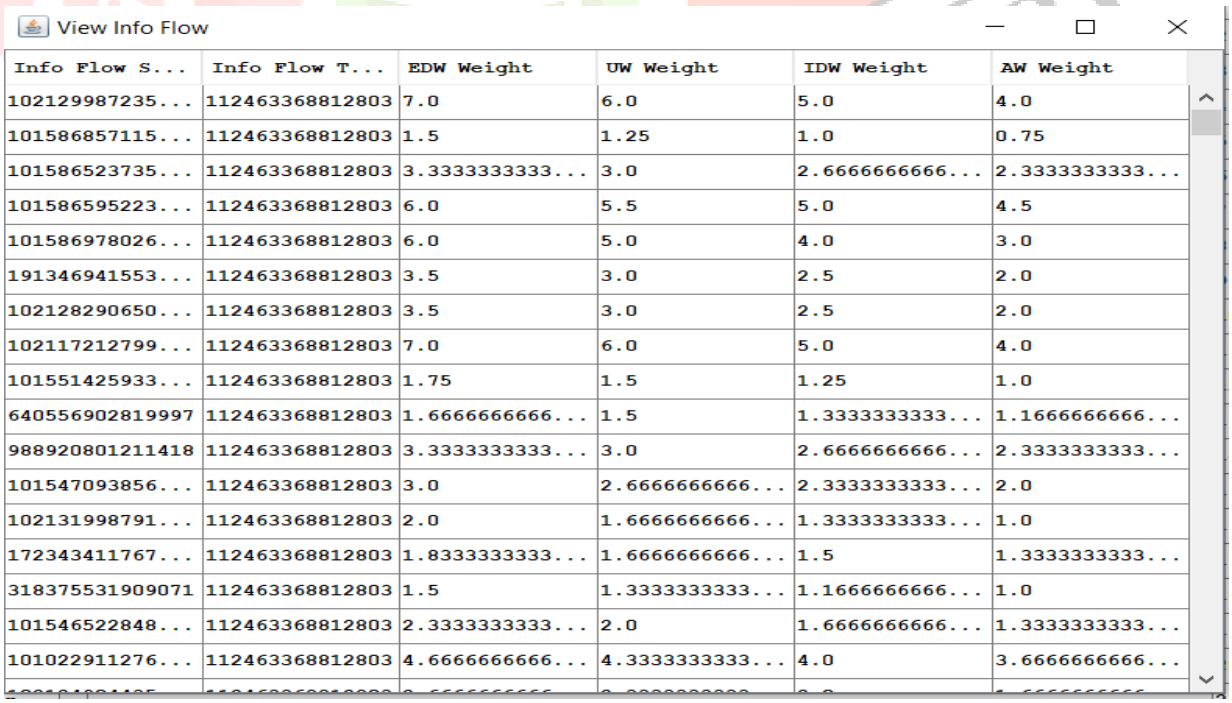
In above screen dialog box we can see total posts and total comments given on that post and the number of comments which as high frequency and then comments are infrequent. All this information we are displaying in above dialog. Now click on 'Intimacy Matrix' to find out two users who are commenting on same post and how many time they commented will show in matrix.



Source User ID	Target User ID	Intimacy Matrix Value
10158678807600203	1822286641425047	2
10213582847748144	10207803398877568	2
1930284303915750	10213031786009747	2
1133894030087283	625863757620318	2
10158591417730431	10213274064911005	2
10158591417730431	10154532581576935	2
10209374105375321	1860700760846636	2
10156239656996110	10155336707425127	2
1120906214682418	165836530143824	2
10212839930655723	10154392993766956	4
626815707518071	1519312528109883	2
10154785239567326	10158674013435402	2
10158678269225076	825804654240496	2
1509806692377578	10211306940003132	2
10212700805820830	10207007775912317	2
792988990878622	10158814794070145	2
10154439661371806	10154435054145811	2
10155294661683659	1021465301324111	2
10158997161525508	1316642461788992	2
1396236707136760	1340487582672363	2
10155522915894369	10213460521973797	3
10213425246332029	10155815416944796	3

Fig -13: View Intimacy matrix

In above screen we can see two users id are commenting on same post and count is also give. Now click on 'Information Flow Miner' button to get all users from whom information is flowing by using four different response weight define in top screens



Info Flow S...	Info Flow T...	EDW Weight	UW Weight	IDW Weight	AW Weight
102129987235...	112463368812803	7.0	6.0	5.0	4.0
101586857115...	112463368812803	1.5	1.25	1.0	0.75
101586523735...	112463368812803	3.3333333333...	3.0	2.6666666666...	2.3333333333...
101586595223...	112463368812803	6.0	5.5	5.0	4.5
101586978026...	112463368812803	6.0	5.0	4.0	3.0
191346941553...	112463368812803	3.5	3.0	2.5	2.0
102128290650...	112463368812803	3.5	3.0	2.5	2.0
102117212799...	112463368812803	7.0	6.0	5.0	4.0
101551425933...	112463368812803	1.75	1.5	1.25	1.0
640556902819997	112463368812803	1.6666666666...	1.5	1.3333333333...	1.1666666666...
988920801211418	112463368812803	3.3333333333...	3.0	2.6666666666...	2.3333333333...
101547093856...	112463368812803	3.0	2.6666666666...	2.3333333333...	2.0
102131998791...	112463368812803	2.0	1.6666666666...	1.3333333333...	1.0
172343411767...	112463368812803	1.8333333333...	1.6666666666...	1.5	1.3333333333...
318375531909071	112463368812803	1.5	1.3333333333...	1.1666666666...	1.0
101546522848...	112463368812803	2.3333333333...	2.0	1.6666666666...	1.3333333333...
101022911276...	112463368812803	4.6666666666...	4.3333333333...	4.0	3.6666666666...

Fig -14: Information flow miner

In above screen we can see information flow between two users and their weight with four different response weights. Now click on 'Response weight FMeasure Graph' to get below graph. Response weight FMeasure graph represents a graph with x-axis as user's id and y-axis as weight values

Fig -15: Response weight graph

VII. Conclusion

In this paper, I have developed a system to find data diffusion models of information from SNS data. The information is utilized to acquire knowledge about data diffusion process among users. This methodology is partitioned into three primary advances. In the first place, information is gathered and separated to distinguish the most dynamic users. Further, to decrease the complexity nature of SNS filtered information a user-centric clustering approach accomplished utilizing User intimacy value which quantifies the degree of relationship between users , based on modularity maximization strategy is carried out. At last,Four distinct approaches to find process models were proposed depending on the Response weight between users using infoflow miner. A quality evaluation for the identified models was performed and talked about.The identified models helps to recognize significant components from the data diffusion procedure, for example, data transmitters and receivers.

VIII. Future Scope

In future, we can conduct additional studies by implementing other methods such as text mining and sentimental analysis to identify the linguistics of the content spread among the users in social networking sites by using data of the users like their language, locations, countries, preferences etc; can be involved for indepth study about information diffusion process.

References

[1] M. De Choudhury, ‘‘Discovery of information disseminators and receptors on online social media,’’ in Proc. 21st ACM Conf. Hypertext Hypermedia, Toronto, ON, Canada, 2010, pp. 279–280.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, ‘‘Fast unfolding of communities in large networks,’’ J. Stat. Mech., Theory Exp., vol. 2008, no. 10, 2008, Art. no. P10008.

[3] W. M. P. Aalst, van der and M. Song, ‘‘Mining social networks: Uncovering interaction patterns in business processes,’’ in *Proc. Int. Conf. Bus. Process Manage. (BPM)*, in Lecture Notes in Computer Science, vol. 3080, J. Desel, B. Pernici, and M. Weske, Eds. Berlin, Germany: Springer, 2004, pp. 244–260

[4] J. L. Moreno, Who Shall Survive?: A New Approach to the Problem of Human Interrelations. Washington, DC, USA: American Sociological Association, 1934.

[5] M. E. J. Newman, Networks: An Introduction. New York, NY, USA: Oxford Univ. Press, 2010.