Community and Important Actors Analysis with Different Keywords in Social Network

Nanang Cahyana, S.ST
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
ncrypt652@gmail.com

Dr. Ir. Rinaldi Munir, MT.
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
rinaldi-m@stei.itb.ac.id

***Abstract*—Twitter has hundreds of millions of users around the world. Using the Twitter as a social network analysis material is very much in demand. Social network analysis can analyze groups and actors of a social network so that it can detect early behaviors that will be performed by groups and actors. But social network analysis in general has not shown strong groups and actors because it uses only one keyword. As a result, this method is quite difficult in detecting early events of a group and actors, especially those associated with cyberterrorist. For that, it needs a method of social network analysis so that the group and the actors produced are really strong and can be detected early behavior that will be done group and actors. The method in question is the use of several different keywords but have the same topic. With this method, it can be obtained a network pattern of groups and powerful actors related to the desired topic so that it can detect earlier behavior that will be done groups and actors. The results obtained are different keywords but have a high value of similarity topics can produce groups and actors are getting stronger. It can increase in the value of graph metric. So this method is feasible to search relationships between different keywords to find the powerfull community and important actor in social network.**

Keywords—*community detection; centrality; crawling; fuzzy relation; twitter; different keywords; social network.*

I. INTRODUCTION

Human social in the world of the Internet has made the virtual world into the data dynamic and growing, is called big data. Big data is not a large data but the data is getting bigger with the passage of time because of human behavior contribute to the behavioral data that called dynamic. Big Data trend today is Twitter. Twitter users have hundreds of millions of users around the world so that it can be used for specific purposes, whether academic, business and government.

The programming language used for crawling is Python 3.5. Crawling on stage takes as input a keyword so that the results generated crawling associated with the given keyword. Results crawling is unstructured data and very large so it is necessary to do the process again as desired. One of the methods used to detect a group in a social network, the community detection. Community detection is a method for detecting the groups in a network [1]. The use of community detection method on social networks will generate data that has different groups relate to each other related keywords defined.

The next stage, which needed a method to produce the most important actor in a group. A very large group in a social network that will be sought after actors were very important and influential group. The method of determining the most important actor is called centrality [2]. Centrality is a method for finding the most important nodes in a group. This method can be obtained centrality of the most influential actors in a group. Data Twitter that has been processed by the method of community detection and centrality. The data generated will be visualized in the form of graphs [3]. That can be seen the form of networking the various groups associated with the keyword, large groups and powerful in a social network, and the actors are important and influential in the large and powerful group.

In large groups and actors importantly, it takes a large amount of data to process Twitter community detection and centrality. To generate a large of data Twitter takes a long time in the process of crawling on the social network. Therefore, it needs a simple method to produce data that Twitter is small but produces a large group and the most important actors in social networks a long time so that the problem can be minimized. One of the methods offered, namely the use of different keywords, but the interplay between keywords with one another.

Methods linkages between keywords used offered fuzzy information retrieval methods. It is a method to describe the relationship between objects of different words so that the correlation values between the keywords needed to produce a large group and the most important actors.

This research problem is the uncertainty generated groups and actors with an identical keyword correlated. Social network analysis can analyze groups and actors of a social network so that it can detect early behaviors that will be performed by groups and actors. But social network analysis in general has not shown strong groups and actors because it uses only one keyword. As a result, this method is quite difficult in detecting early events of a group and actors, especially those associated with cyberterrorist. For that, it needs a method of social network analysis so that the group

and the actors produced are really strong and can be detected early behavior that will be done group and actors. The method in question is the use of several different keywords but have the same topic. With this method, it can be obtained a network pattern of groups and powerful actors related to the desired topic so that it can detect earlier behavior that will be done groups and actors. For every change that occurs on group and actors will affect the value of the graph metric so that would be obtained opportunities actors group generated with different keywords.

II. LITERATURE STUDY

Literature study in this research about technology such as programming language, and mathematic theorem such as probability. There can be described:

A. Fuzzy Relation Retrieval

The relation represents the presence or absence of association, interaction or communication between elements over two sets. This concept is known as the degree or strength of association or interaction between elements [4]. The Fuzzy Relation in this research applied on Twitter social network, given types used are keywords to tweets, tweets to tweets, tweets to keywords, and keywords to keywords.

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation [5] in this research applied on Twitter social network, given an N data set twitter periodic of size M. Then the corpus of tweet is summarized on the N occurrence table with $n(k_i\ t_j)$ used storing the number of occurrences of the tweet k_i from the t_j tweet.

Further probability $P(k_i,\ t_j,\ z_k)$ to build the model graph follows equation (1).

$$P(k_i|t_j) = \Sigma P(z_k|t_j)\, P(k_i|z_k) \quad (1)$$

With $P(z_k|t_j)$ showing the probability of topic z_k to tweet t_j, and $P(k_i|z_k)$ showing the probability of keyword k_i to z_k.

C. Community Detection

Community Detection is the key to understanding the complex network structure and ends at extracting useful information [6]. Community in this research uses vertex, edge, average geodesic distance, modularity, and density parameters.

D. Centrality

Centrality related to the potential importance of a node. Some nodes have a stronger influence than others, or more easily accessible to others, or as an intermediary in most relationships node-to-node [2]. Centrality in this research uses degree, betweenness, closeness, and eigenvector centrality.

E. Graph Theory

Graf is a collection of nodes (nodes) are connected to each other through the side/arcs (edges) [7]. The G graph like the set of sets (V, E), is written with the notation G = (V, E), in which case V is the non-empty set of nodes and E is the environmental set of sides [7].

F. Twitter

Twitter is a free social messaging service to send and receive short messages in real time. Commonly called social networking or microblogging service. The messages are limited to 140 characters and are called tweets. Twitter is considered as a social networking service because the user can create a profile and connect with others electronically on services.

G. Cyberterrorism

The cyberterrorism is terrorist activity in cyberspace [8]. The term was first used by Barry Collin, researchers Institute of Security and Intelligence in California in 1997. Activities of terrorism that are well known include propaganda, recruitment, funding, training, planning, and the spread of terror to attack in cyberspace or cyberattack. The internet assists terrorist activity originally was conventional, connected only to individual persons into a modern network with a network of globally connected media [9].

III. METHODOLOGY

The research methodology used was experimental methodology. The experimental research can be regarded as the research methods used to find a specific treatment effect against the other in uncontrolled conditions [10].

The scope of this research is the keywords commonly misused by cyberterrorism. The sample keywords reference to terms commonly abused terrorism, namely *Askariy, Baiat, Hijrah, Idad, Ightiyalat, Istimata, Jamaah, Jihad, Kafir, Khalifah, Khilafah, Qital, Syahid* [9]. Using these keywords on Twitter will create many communities and actors that will be analyzed in community detection and centrality processes. Then use fuzzy relation retrieval to determine the value of the relationship between keywords so as to be a community controller and actors generated by different keywords.

Community detection process will produce a variety of community with varying degrees. Community with the highest degree centrality will feed into the process. In the process would be sought centrality. Degree centrality nodes having the highest among the other nodes so as to produce the most important nodes or actors on the social network Twitter. The output of these two processes will feed into the process of visualization of graphs that can be viewed a network of groups and actors on Twitter. This stage, to determine metric graph of community like value of vertex, edge, average geodesic distance, modularity, and density, then centrality like value of degree, betwenness, closeness, and eigenvector centrality.

Input at this stage of data analysis is groups and actors on the data processing. Groups and actors will be shown in a graph in order to obtain inter-node network view. The next stage is called the analysis of the data by the relation method, which represents a node as a random variable with the value of relationships between nodes as interdependent. With relation,

for every change that occurs on a node will affect the value of the of other nodes so that would be obtained opportunities actors group generated by different keywords.

IV. Result and Discussion

This analysis used a variety of terms are misused by terrorists to launch the action. The term of abuse is taken as a keyword to enter the social network analysis program. It can be analyzed as follows:

A. Keywords

By using all 13 of the term it has been done processing, it can be described:

TABLE I. Result of Crawling Twitter

Index	Keywords	Words	Vertex	Edge
k_1	*Baiat*	9.96 (9)	596 (8)	821 (7)
k_2	*Hijrah*	88.22 (2)	1919 (2)	2680 (2)
k_3	*Idad*	11.67 (8)	772 (7)	764 (8)
k_4	*Jamaah*	64.55 (5)	1081 (6)	1356 (6)
k_5	*Jihad*	74.15 (4)	1265 (5)	1487 (5)
k_6	*Kafir*	76.80 (3)	2165 (1)	2740 (1)
k_7	*Khalifah*	28.32 (7)	313 (9)	380 (9)
k_8	*Khilafah*	89.67 (1)	1453 (3)	2571 (3)
k_9	*Qital*	0.73 (10)	51 (10)	68 (10)
k_{10}	*Syahid*	60.09 (6)	1471 (4)	2305 (4)

1) Trending keyword is *Khilafah*, the total is 89.67 then *Hijrah, Kafir, Jihad, Jamaah, Syahid, Khalifah, Idad, Baiat*, and *Qital* is smallest 0.73 as shown in Table I. So 10 keywords will input to fuzzy relation retrieval.

2) Keywords that do not appear on this research is *Istimata, Ightiyalat,* and *Askariy*.

3) The most discussed key actors are Kafir with a number of actors (vertex) is 2571 and communication relation (edge) is 2740.

B. Fuzzy Relation

The relation between keywords can be seen as shown in Table II.

TABLE II. Result of Fuzzy Relation Keyword to Keyword

R(x,y)	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}
k_1	1,0	0,9	0,9	0,9	1,0	0,9	1,0	0,9	0,9	1,0
k_2	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	0,9	1,0
k_3	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	0,9	1,0
k_4	0,7	0,7	0,7	1,0	0,8	0,8	0,8	0,7	0,9	0,9
k_5	0,9	0,8	0,9	0,9	1,0	0,9	1,0	0,8	0,9	1,0
k_6	0,9	0,9	0,9	1,0	1,0	1,0	1,0	0,9	0,9	1,0
k_7	0,7	0,6	0,6	0,7	0,7	0,7	1,0	0,6	0,7	0,8
k_8	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
k_9	0,7	0,7	0,8	1,0	0,8	0,8	0,9	0,7	1,0	0,9
k_{10}	0,7	0,7	0,7	0,8	0,8	0,8	0,9	0,7	0,8	1,0

1) $R(k_i,k_7)$ has the highest and highest relation value compared to the relation of all ki keywords to one other keyword. This means that all keywords have a strong relation or connection to the caliph's keywords.

2) In contrast $R(k_2,k_i)$ has the highest and highest relation value compared to one keyword relation to all ki keywords. This means that the hijrah keyword has a strong relation or linkage to all ki keywords.

3) The keyword relation that has the highest value with the highest value of the reverse relation is $R(k_5,k_6)$: 0.9561 with $R(k_6,k_5)$: 0.9655. This means that the keyword jihad with kafir has the value of keyword relation mutually reinforcing than other keywords.

TABLE III. Community Metrics

Index	Community	Short Path	Density	Modularity
k_1	*Baiat*	3.2296	0.0016	0.5238
k_2	*Hijrah*	3.3767	0.0005	0.6416
k_3	*Idad*	1.9253	0.0007	0.6956
k_4	*Jamaah*	2.3648	0.0008	0.6479
k_5	*Jihad*	5.0317	0.0008	0.7968
k_6	*Kafir*	6.0762	0.0005	0.6316
k_7	*Khalifah*	1.9642	0.0032	0.7449
k_8	*Khilafah*	3.8054	0.0006	0.3973
k_9	*Qital*	2.3685	0.0161	0.5389
k_{10}	*Syahid*	5.4039	0.0006	0.4437

Based on Table III, it can be explained that:

4) The community that has the average shortest distance between nodes is *Idad* who have the number about 1.9253. Then *Khalifah* has the number 1.9642. This means that the actors in these two communities know each other.

5) Communities that have the smallest density value is *Hijrah* and *Kafir*. There are have value 0.0005. Then Khilafah has value 0.0006. This means that the actors in the community of *Hijrah*, *Kafir* and *Khilafah* are very crowded.

6) The community that has the greatest modularity value is *Jihad*. This means that the *Jihad* community is so easily broken down into parts of the community that it is easy to control or manage other actors.

TABLE IV. Centrality Metrics

Index	In-Degree	Out-Degree	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality
k_1	182	17	80054.578	1	0.067
k_2	249	20	213658.183	1	0.042
k_3	162	7	25760.000	1	0.076
k_4	44	13	4510.167	1	0.075
k_5	52	13	22878.000	1	0.069
k_6	178	37	327689.222	1	0.018
k_7	97	10	9120.000	1	0.097
k_8	381	31	492953.027	1	0.047
k_9	5	6	174.000	1	0.159
k_{10}	347	122	515834.114	1	0.048

Based on Table IV, it can be explained that:

7) The community with the highest degree is c_8 or *Khilafah*. This means that the actor *Khilafah* in this community has a high degree, to communicate with many other actors in the community.

8) Communities with the greatest eigenvector centrality value are c_9 or *Qital*. This means that actors in the *Qital* community have an influence over each other compared to other communities

C. Actor

The actor who has the highest degree in each community and has relationships with other communities is as shown in Table V.

TABLE V. ACTORS OF HIGH DEGREE AND RELATED TO COMMUNITIES

Actor	**k_1**	**k_2**	**k_3**	**k_4**	**k_5**	**k_6**	**k_7**	**k_8**	**k_9**	**k_{10}**
wahhabicc_jabar	g_1			g_{40}		g_1	g_1	g_1		g_3
wahhabicc	g_1			g_1	g_5	g_1	g_4	g_1		
b3811nez	g_1					g_1	g_1	g_1		
networksmanager		g_1			g_{159}					
nblhn		g_2			g_{174}					
drbenirusani		g_2								g_1
zheraomega	g_6	g_{138}		g_1		g_{162}				
sahal_as				g_4		g_1	g_{12}	g_1		
semiaji_w				g_1		g_1				
joxzin_jogja					g_4	g_1		g_1		
felixsiauw		g_{79}		g_{118}		g_1		g_1		
netizentofa						g_1		g_1		g_{45}
sinyovandlahom	g_1				g_4	g_{62}	g_4	g_1		
wakilgubernurkw	g_1			g_1	g_5	g_1		g_1		
legendbytheway						g_2			g_4	
maulinaantika				g_{10}	g_{18}					g_2

Based on Table V, one of the highest-ranking actors and having a relationship with many communities is *wahhabicc*. A graph formed from a *wahhabicc* actor from several related communities can be seen in Figure 1.

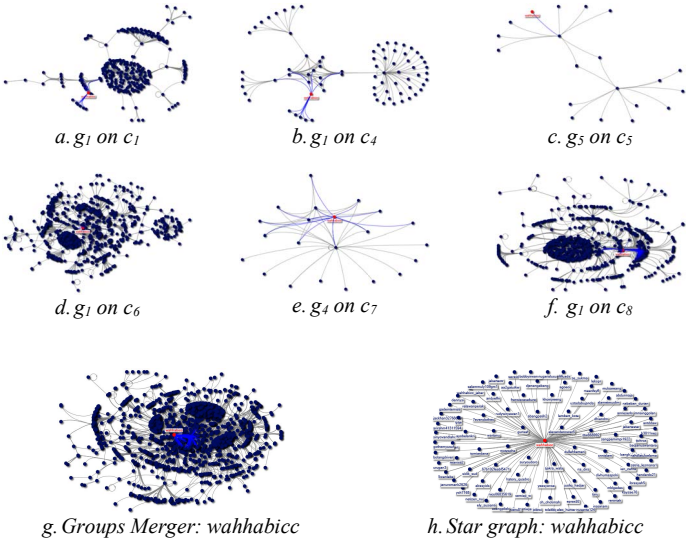

Fig. 1. Graph of actor "*wahhabicc*"

Based on Figure 1, the percentage value of each community size and the centrality actor *wahhabicc* as shown in Table VI and VII is obtained.

TABLE VI. UPDATE COMMUNITY OF ACTOR GRAPH METRIC "WAHHABICC" AFTER GROUP MERGER

g	*v*	*v'%*	*e*	*e'%*	*sp*	*sp'%*	*d*	*d'%*	*m*	*m'%*
a	298	74	431	88	3.3	83	0.005	0.5	0.05	28
b	77	83	142	96	3.1	84	0.023	0.4	0.04	28
c	20	85	20	99	2.5	87	0.053	0.3	0.00	31
d	905	49	1398	60	6.1	69	0.001	0.5	0.11	23
e	22	85	53	98	1.8	90	0.106	0.2	0.09	24
f	791	54	1436	59	3.8	81	0.002	0.5	0.12	22
g	1812	86	3480	100	4.8	99	0.001	0.5	0.13	31
x		71		83		82		0.4		26.

TABLE VII. UPDATE CENTRALITY OF ACTOR GRAPH METRIC "WAHHABICC" AFTER GROUP MERGER

g	*id*	*id'%*	*od*	*od'%*	*bc*	*bc'%*	*cc*	*cc'%*	*ec*	*ec'%*
a	1	93	17	44	3515	41	0.0010	0.33	0.003	1.6
b	11	79	2	56	435	42	0.0051	0.31	0.047	1.2
c	1	93	0	58	0	43	0.0179	0.23	0.033	1.4
d	9	82	10	50	35074	29	0.0003	0.34	0.001	1.7
e	4	89	9	50	33	43	0.0333	0.14	0.099	0.8
f	44	35	31	33	71631	15	0.0004	0.34	0.001	1.7
g	66	94	40	58	260238	43	0.0002	0.34	0.003	1.7
x		79		48		35		0.28		1.4

Based on Table VI and Table VII, the actor *wahhabicc* has an average percentage increase: (1) vertex 71%, (2) edge 83%, (3) average geodesic distance 82%, (4) density 0.4%, (5) modularity 26%, (7) out-degree 48%, (8) betweenness centrality 35%, (9) closeness centrality 0.28%, (10) eigenvector centrality 1.4%. Based on this method, the others actor has value increase of graph metric after actor's group merge by related communities.

D. Time Performance

Time performance on programming Python 3.5 to crawling, pre-processing, community detection, centrality, and graph process, as shown in Table VIII.

TABLE VIII. TIME PERFORMANCE

No	**Process**	**Time Average**
1	Crawling	338.27
2	Pre-Processing	3.075
3	Creating Graph	343.745
TOTAL		685.09

E. Memory Performance

The memory performance in this program with 5,000 tweets has the value of memory average about 23.5 MB, the average processing speed 33.57 kb/s, and get about 7 tweet/s.

V. CONCLUSION

This method can increase graph metric of community and actors. It will increase the value of vertices, edges, average geodesic distance, modularity, degree centrality, betweenness centrality, and eigenvector centrality but will decrease the value of density and closeness centrality. So it can be community and actors getting powefull and used as early detection of behavior community and actors.

FUTURE RESEARCH

The next research is validation of actors suspected of having an effect on group related to selected topics, and develop crawling twitter real time to solve data collecting limited 7 days on twitter.

REFERENCES

[1] S. Fortunato, "Community Detection in Graphs, Complex Networks, and Systems", Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino, I-Italy, 2010.

[2] F. Bloch, M. O. Jackson, and P. Tebaldi. "Centrality Measures in Networks", 2016.

[3] C. Borgelt, "Data mining with graphical models". Dissertation: Otto-von-Guericke-Universit¨at Magdeburg. 2010.

[4] B. A. Darmadi, Aplikasi Search Engine Paper/Karya Ilmiah Berbasis Web Dengan Metode Fuzzy Relation, Jurusan Teknik Informatika, Fakultas Teknologi Industri – Universitas Kristen Petra, Surabaya, 2006.

[5] Blei, D. M., dan Jordan, M. I., (2003): *Latent dirichlet allocation*, Journal of Machine Learning Research 3 (2003) 993-1022.

[6] L. Tang, and H. Liu, "Community Detection And Mining In Social Media," Morgan & Claypool Publishers, 2010.

[7] R. Munir, "*Matematika Diskrit Edisi Ketiga*", Informatika, Bandung.

[8] A. S. Bakti, Deradicalisation Cyberspace, Symbiosis Preventing Terrorism and Media, Penerbit Daulat Press, Jakarta, 2016.

[9] P. R. Golose, Terrorism invasion into Cyberspace, Yayasan Pengembangan Kajian Ilmu Kepolisian, Jakarta, 2015.

[10] Dr. Sugiyono, Qualitative and Quantitative Research Methods R&D, Penerbit Alfabeta, 2010.