ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 26-33 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp18.pdf

Finding Twitter Communities with Common Interests for Hot Topics in the World

Ala'a Abbadi[1], Ameera Jaradat[2]
[1] Yarmouk University, Jordan, alaamabb@gmail.com
[2] Yarmouk University, Jordan, ameera@yu.edu.jo

Abstract

Community detection problem is one of the most growing problems in computer science due to its need in many applications for large scale networks such as Twitter, LinkedIn, and Facebook. Great efforts have been spent on entire networks' linkages to find communities. But little effort has been spent on finding communities based on common interests of entire networks. Detecting communities based on common interests may be used as a prior step for finding network linkages based on the resulted communities; which increases the efficiency of such algorithms. However, a well-designed technique must be used to find the common interests without adding computational overhead to linkages search algorithms. This work intends to detect communities of people who have common interests on Twitter, based on tweets Hashtags for specific topics in Arabic language. The common interests are detected using the Cosine similarity measure. After that, cliques are extracted from these communities using depth first search algorithm. Generally, the experiment results are reasonable and logical.

Keywords: Clique, Community detection, Cosine similarity, Social Networks, Twitter Hashtags.

1. INTRODUCTION

One of most applications in computer science for extracting cliques is finding the largest communities on twitter. This work intends to detect Twitter communities of people based on common interests using tweets Hashtags for specific topics in Arabic language. Most previous approaches for community detection start with finding communities before determining their interests [5] and using methods that detect communities which do not share any specific interest; this leads to inefficient use of all communities to obtain any needed information .But few approaches have focused on checking interests before finding communities [13,14] The Cosine similarity measure will be used in this work; by which, the detected communities will more likely have higher linkages between the individual users than the communities of the previous approaches. We intend to make more comprehensive research than the work presented by (Lim and Datta, 2012) by choosing more interests. The common interests will be detected using Cosine similarity measure, for finding communities of users and their related tweets according to specific Hashtag. After that, cliques of users will be extracted from these communities using Depth First Search algorithm.

1.1. BACKGROUND

Twitter is a social networking service which let users to send messages with 140 characters. Those messages are called tweets. Twitter is a rapidly growing service, which was one of the ten most visited websites and it was described as the SMS of the internet [17].As of October 2016, Twitter has more than 313 million monthly active users. Moreover, Twitter provides Application Programming Interface (API), which includes the required functions for developers and researchers to link their applications with twitter data [22].Therefore, Twitter is considered a good choice for researchers due to its popularity and data availability [9].

Maximum Clique problem is one of the most popular problems in Graph theory, which stills do not have its polynomial time solution. Many algorithms have been proposed for solving this problem. The idea behind maximum clique problem is to extract the sub graphs with the maximum cardinality. The maximum clique in a random graph is NP-Hard problem, which has been stimulated by many problems like social network, mobile networks, and computer vision [3].

Cosine similarity measure is a document classification formula that is used for text documents. In Cosine similarity, documents (tweets) are represented as term vectors. The similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors which is known as Cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents [2].

2. LITERATURE REVIEW

Twitter is one of the most popular social networks due to its popularity and data availability. (Kwak et al., 2010) provided a study for the topological characteristics of Twitter. This study has found that Twitter is different from other social networks since it has law reciprocity, and the distribution of follower-following topology analysis shows non-power and short effective diameter. This study provided an identification for influential on Twitter. That is, users are ranked based on number of followers and their tweets' popularity, rather than using number of users' followers and pages' popularity; because pages' popularity gives similar results, while tweets' popularity can give more accurate results. And then, the resulted popular tweets of users can be used to extract most popular pages [4].

Behind Twitter Hashtags (Cunha, 2011) studied how Twitter hashtags are created and published based on linguistic inspiration models. The author has analyzed the propagated hashtag terms for groups of people who can influence each other's to spread new terms for a specific hashtag. This study shows that as the length hashtag term decreases, the possibility for this term to be popular increases. Also, this study provided a model to find the newly high spread terms that may be propagated to the public in the future [7].

Wang, (2011) presented a graph model for hashtag classification, based on the sentiment polarity for a hashtag within a period of time. The classification factors that are used for this model are hashtags literal meaning, co-occurrences, and the sentiment polarity of tweets that includes those hashtags. The presented model has resulted in better performance than the baseline, by using the hashtag literal sentiment hint[10].

Santoro et al., (2015) presented a graph model for Hashtag Entity graph (HE) based on hashtags and Wikipedia entities. This model has solved the relatedness and classification of problems in Twitter Hashtags using Cosine similarity measure. It has joined semantic relatedness between entities and co-occurrence data between entities and hashtags. So HE graph has provided a structured representation for tweets and their occurring hashtags. Moreover the authors presented a novel algorithm to improve lexical classifiers by proposing new hashtag classifier that hinges on HE graph. The results shows that error rates is decreased up to 1% [20].

Many community detection algorithms have been designed over the previous years; but few algorithms were designed for interests' detection. (Li et al., 2010) proposed TTR-LDA-Community model which combines the Latent Dirichlet Allocation model (TTR-LDA) and Girvan Newman community detection algorithm with an inference mechanism. The model is applied on data from social tagging system. The authors show that users in the same community tends to be interested in similar topics in all tested time periods. Besides, topics may divide into several subtopics and scatter into different communities over time in order to detect interests. It is suggested that the combined model outperforms the TTR_LDA in tag prediction [5].

Jin et al., (2011) proposed the LikeMiner system that uses likes of users to detect popular topics for user's friends, for detecting user interests. Because LikeMiner system has been conducted on Twitter, and it works on individual users not communities which needs long computations [8].

Lim, (2012) proposed a method for detecting communities based on the highly interactive community members, ignoring members who are rarely interact with their communities. His approach first classify popular celebrities into different interest categories. After that the communities are detected based on the communication and linkages among celebrities' followers [12].

Lim and Datta, (2012b) continued the previous study so that they proposed an identification for communities with common interests, based on celebrities that are representative of interest category. They detected communities using linkages between followers of these celebrities. This study resulted in providing a tool for the implementation of target advertising and viral marketing for products with specialized audience. Moreover the experiments shows that an increasing level of interest in a category correlates with detecting larger communities on average, higher clustering coefficient, and shorter path lengths [14].

Yang et al., (2011) proposed framework, called Friendship and Interest Propagation (FIP) that works individual users, by detecting interests for each users, and then recommend friends based on the detected shared interests. FIP system has been conducted on Yahoo! Pulse1 with predefined list of interests. The experiments demonstrate that coupling friendship with interest, FIP achieved much higher performance on both interest targeting and friendship predicting than systems using only one source of information [11].

Behind solving maximum clique problem [1] presented an innovation for finding all maximum cliques in a complex netwok using a parallel algorithm named PEAMC (Parallel Enumeration of All Maximal Cliques).After that the authors have presented a comparison with existing algorithms for evaluating the run of PEAMC. The results show that PEAMC is more effective and scalable.

Rajawat et al., (2010) presented the genetic algorithm that aims to solve the Maximal Clique Problem. The authors found that the genetic algorithm can solve the problem correctly and the cycle necessary to get the correct solution that is almost a linear function of the number of nodes. The result shows that with this new algorithm, it is possible to get an answer from a very small initial dataset, avoiding counting all candidate answers [6].

Pattabiraman and Patwary, (2013) presented a new heuristic algorithm to solve the maximum clique problem. The authors compared the performance of new heuristic algorithm with the algorithms of Carraghan-Pardalos (CP), cliquer algorithm and MCQD+CS algorithm by using extensive trials on three different kind graphs. With the graphs of DIMACS benchmark and certain dense synthetic graphs, the results shows that the new heuristic algorithm performs the same as the CP algorithm, but slower than other algorithm (cliquer and MCQD+CS). For large sparse graphs, the new heuristic algorithm runs several times faster than the other three algorithm [18].

Soleimani-Pouri et al., (2013) presented a new hybrid algorithm that contains Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) algorithms to find a maximum clique in a popular social networks. For valuation of the proposed algorithm, many trials applied on some social

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 26-33 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp18.pdf

network datasets. The result showed that the new hybrid algorithm makes an improvement on the basic of the ACO algorithm in simply and quickly manner and the Simulation results on popular datasets indicate improvement of outcomes for proposed hybrid algorithm in comparison with the other algorithm [18].

Behind classification document measure techniques, there are different similarity measures which have been used for text document clustering [2] analyzed three of these measures such as squared Euclidean distance, Cosine similarity, and relative entropy. The results show convergence of measures in performance.
Dongen et al., (2012) investigated two classes of transformations of Cosine similarity and Pearson and Spearman correlations into metric distances. The authors derived metric distance using metric preserving functions. The Cosine similarity, which is a standard measure used in information retrieval, puts anti-correlated objects maximally far apart. While Pearson and Spearman collates correlated and anti-correlated objects [16].

Deshpande et al., (2014) presented a comparative study for document similarity measure techniques including Jacard similarity measure, Metric similarity measure, Euclidean Distance measure, and Cosine similarity measure. This study concluded that Cosine similarity is the most effective and the simplest among other techniques, due to its scalability and implementation simplicity [19].

3. METHODOLOGY

The framework of the proposed approach starts by calling Twitter API, in order to collect the most trending topics. These topics are used then to retrieve users who are most likely interested with these topics with the latest five tweets. After that those tweets are normalized to get the RD from them. On the other hand, the trending topics that are extracted firstly via Twitter API are used to get raw tweets, which are normalized to get the TD. RD and TD are added to the Cosine measure for calculating their similarity. The calculated similarity is used then in detecting communities. And finally, the Depth First Search algorithm is used to find cliques of users. Figure. 1 describes the general framework of this approach.

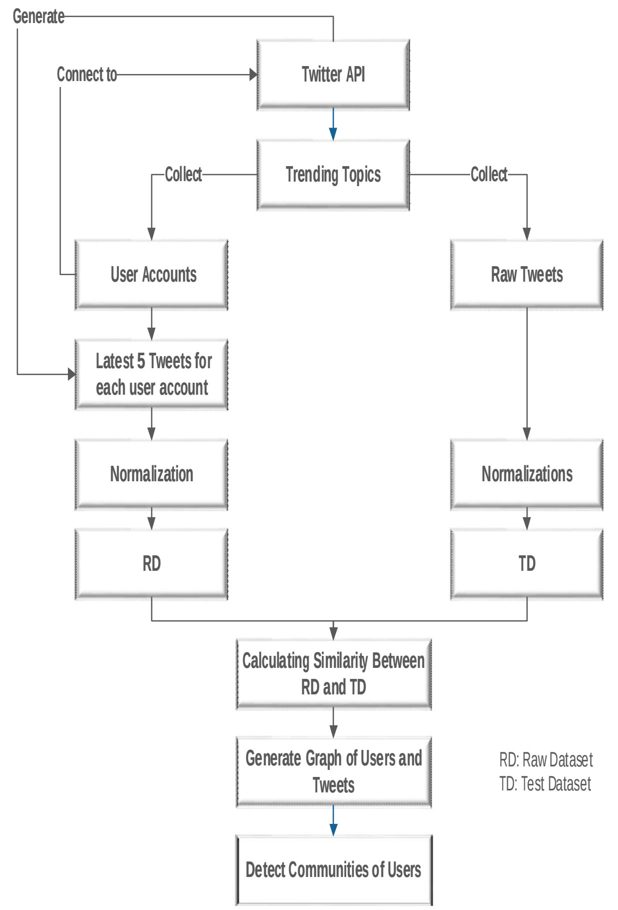

Fig. 1 GENERAL FRAMEWORK

3.1. PROPOSED MECHANISM

This section describes the proposed mechanism for this research which starts with requesting tweets from Twitter via REST API. The retrieved tweets are filtered based on the trend topic, in order to use them for building RD and TD. Then, after the communities are detected, the cliques are found. The following are the required steps for this mechanism:

3.1.1. REST API: firstly, the user should collect the initial dataset, which includes users' tweets that are related to the chosen trending topic. Those tweets are requested via Twitter REST API, which is used with so many web sites for rendering some specific tweets in their pages. The process for request REST API is shown in Figure. 2 The user should create an account on Twitter website, in order for him to access REST API using this account .

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 26-33 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp18.pdf

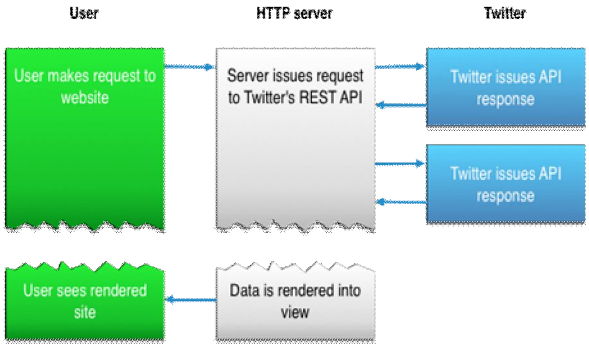

Fig. 2 REST API

The steps below discuss the all required steps to make configuration for our API:

- Create new Twitter accounts.
- Create an API key for our application https://dev.twitter.com/apps (we should fill all requirements to get successful application, such as: Name, Description, Website, and Callback URL).
- After creating our application, we can access what we need to authenticate Twitter using OAuth, namely (Consumer key, Consumer secret, Access token, Access token secret)
- Now we can connect with Twitter API.

3.1.2. Trending Topics: the data that are retrieved form REST API contains two lists; the first one is the list of tweets that are related to the trending topic, while the other one includes the user accounts that are relate to the tweets list. The trending topics which is chosen to be studied in this research are about hot topics in the world. The following are the Hashtags that are selected for this purpose:

- Public opinion issues and polls.
- Topics related to health organizations.
- Contemporary issues.
- Tourism topics.

3.1.3. Raw Dataset (RD): the initial dataset is filtered to build the Raw Dataset (RD). The list of users' tweets is minimized to keep only the latest 5 tweets for each user. While the list of user account is kept as it is.

3.1.4. Test Dataset (TD): the testing dataset is the collection of retrieved tweets of the chosen hash tags. This dataset will be used to build the community, by comparing users' interests with this dataset.

3.1.5. Tweets' Normalization: within this step, the test of each tweet in TD is modified by removing the unrequired characters that will not be used with later steps, such as numbers, characters from other languages and symbols such as emotion.

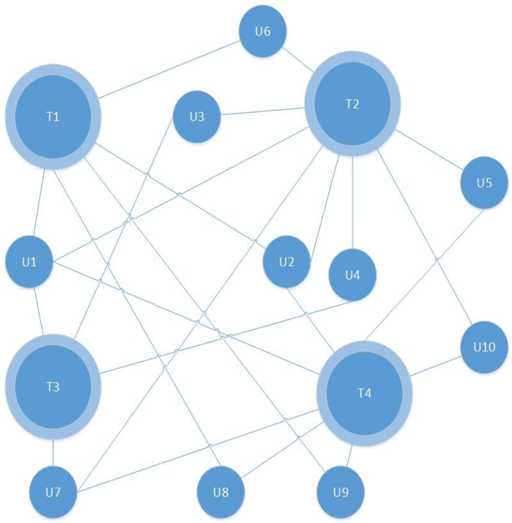

Fig.3 COMMUNITY EXAMPLE

3.1.6. Community Detection in this step will be founded using the Cosine similarity measure, which uses both RD and TD to build the detected community. TD is determined to be the main component of the proposed community, since each tweet in TD is as a node in the detected community, while RD are the edges of this community. Such nodes will be connected directly with its corresponding nodes from RD, depending on their relations. These relations are clarified in figure. 3 Also in this step, the detected communities is divided into smaller communities, based on the strongest connections between nodes. A sample set of community is shown in Figure 4. In which, each user node has one link with the most corresponding tweet node

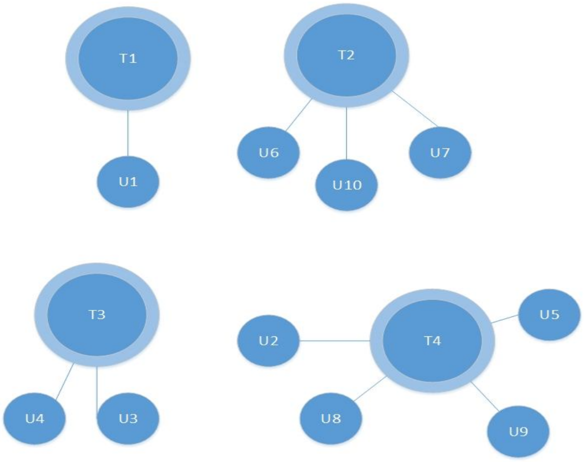

Fig.4 SMALLER COMMUNITY BASED ON STRONGEST CONNECTION

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 26-33 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp18.pdf

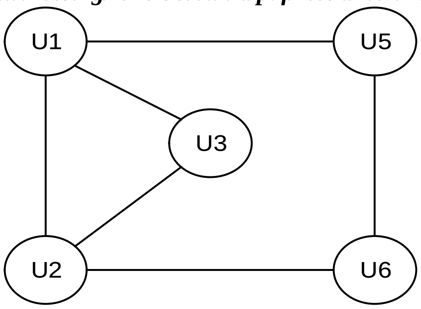

Fig. 5 EXTRACT CLIQUE OF USERS

3.1.7.Extracting Cliques: in this step the users who share the same tweet are linked together to extract all cliques which contain group of people with common interests. A sample set of cliques is shown in figure 5.

Each tweet in both RD and TD is represented as a vector. Vectors are set of terms, and each term has a weight that reflects its importance on that RD or TD. There are several methods that can be used to calculate this weight. The method is frequency-inverse document frequency (TF-IDF) can be used to calculate the weight, where the term frequency (TF) represents the term frequency of a tweet in TD. While the inverse document frequency (IDF) represents the importance of a term regarding to the entire dataset, which is the number of tweets in the dataset divided by the number of tweets containing a term. TF-IDF method is presented as follows:

$$TF = \frac{\text{number of occurrence of term in Tweet}}{\text{number of terms in Tweet}} \qquad (1)$$

$$IDF = \log\left(1 + \frac{N}{nj}\right) \qquad (2)$$

Where:

- N is the total number of tweets
- nj is the number of tweets containing the term

After finding TF and IDF, the formula is applied as follows:

$$TFIDF = TF * IDF \qquad (3)$$

3.2. Cosine Similarity

TF-IDF is used to compare the tweets in RD vectors with the tweets in TD vectors using the well-known Cosine similarity measure, which has been used in so many experiments over the years, due to its reliability [19].Cosine similarity measures the cosine of the angle between two vectors, the value of cosine similarity is bounded by the interval [0, 1] this measure has been used in information retrieval and text mining [16].

Two vectors with attributes the RD and TD are used in cosine similarity cos (θ), which is represented using a dot product and formatted as follows:

Value	Evaluation
Smaller than 0.5	No Connection
Greater than or equal 0.5	Connect

Table 1. RANGES OF COSINE SIMILARITY MEASURE

$$Cosine\ Similarity(RD,TD) = \frac{RD[0]*TD[0]+\cdots+RD[n]*TD[n]}{\|RD\|*\|TD\|} \qquad (4)$$

Where tweets length is computed as follows:

$$\|RD\| = \sqrt[2]{RD[0]^2 + RD[1]^2 + \cdots + RD[n]^2} \qquad (5)$$

$$\|TD\| = \sqrt[2]{TD[0]^2 + TD[1]^2 + \cdots + TD[n]^2} \qquad (6)$$

For text matching, the vectors RD and TD are usually the term frequency vectors of the tweets. Cosine similarity resulted value can be set in the range between 0 to 1, because the TF cannot be negative, the angle between the two vectors must be less or equals to 90°.

After the cosine similarity between the RD and TD is calculated, the community is being detected. This community connects each node (tweets) with the related edge (users) based on Cosine result. If the result of Cosine similarity is 1, the user account is linked to the tweet in the community. On the other hand, if the result of Cosine similarity is 0, then the user does not have a link with the tweet. Table 1 shows the ranges of Cosine similarity measure.

4. EXPEREMENTS AND EVALUATION

The sample dataset that are planned to be used in this research are real dataset, which can be retrieved from twitter via REST (REpresentational State Transfer) APIs. REST provides programmers with sets of functionality to be used to read or update specific data inside Twitter; such as, reading or writing Twitter data, adding new tweet, reading author profile and follower data, and many other functions.

Three types of datasets are presented in this research:

- Raw Dataset (RD): The initial dataset which is retrieved directly with regard to the latest five tweets of each user.
- Testing Dataset (TD): Collection of tweets that are present specific topic, which is determined in this research to be hash tags.
- Final Dataset (FD): the final extracted cliques from the detected communities.

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 26-33 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp18.pdf

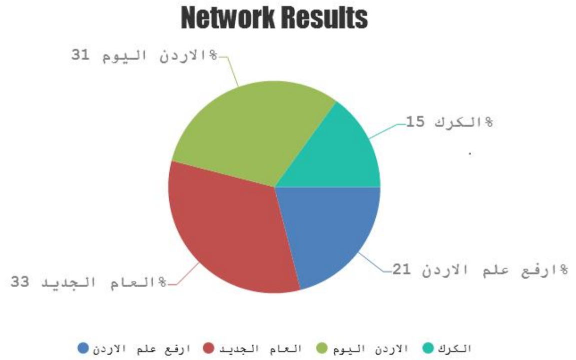

FIG.6 DISTRIBUTION OF USERS BETWEEN HASHTAGS

According to Twitter API, we collect our datasets for TD and RD according to Arabic Hashtags, which is recently posted new Hashtags on recent time. For our experiment we collected 34 user accounts with five tweets for each one, so the total tweets in raw dataset is 170 tweet according to different Hashtag topics.

The next phase we collected our TD, using same technique in the first level, such as:

- ارفع علم الاردن
- العام الجديد
- الاردن اليوم
- الكرك

We collected 20 tweets in each of the above Hashtags; so TD have 80 different tweets.

Communities of users show the last level in our research, and the relationships between different users' accounts on Twitter. FIG. 6 shows the distribution of users between Hashtags.

Figures below show the distribution of users' tweets in the same Hashtag for all testing Hashtags list, where x-axis represents tweet id from TD and y-axis represents Cosine similarity value.

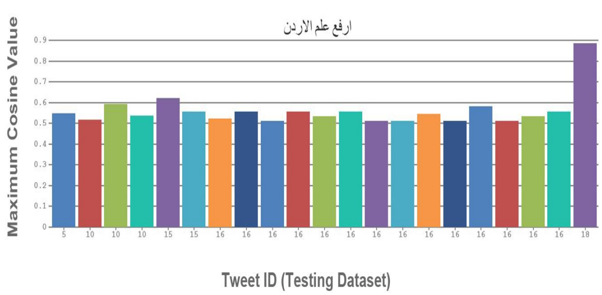

FIG.7 THE DISTRIBUTION OF USERS' TWEETS IN THE SAME HASHTAG

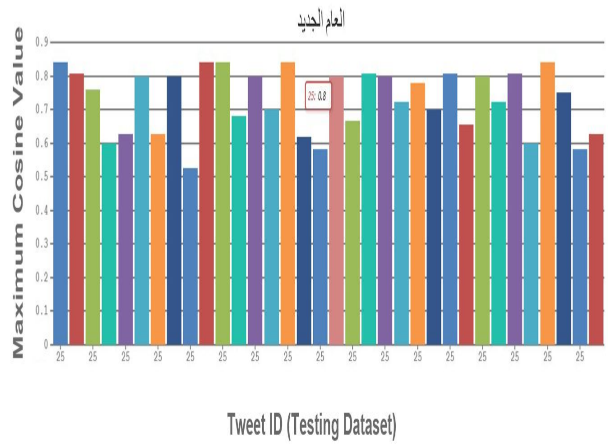

FIG.8 THE DISTRIBUTION OF USERS' TWEETS IN THE SAME HASHTAG

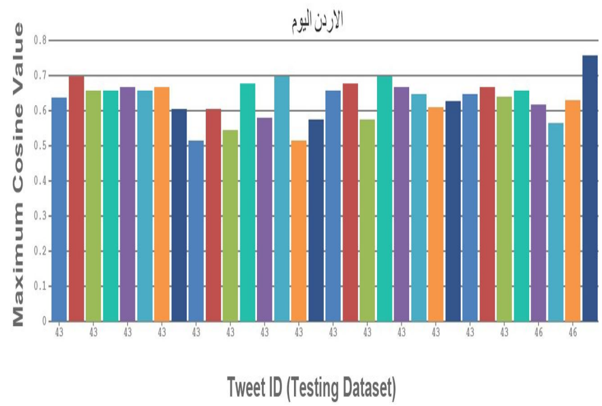

FIG.9 THE DISTRIBUTION OF USERS' TWEETS IN THE SAME HASHTAG

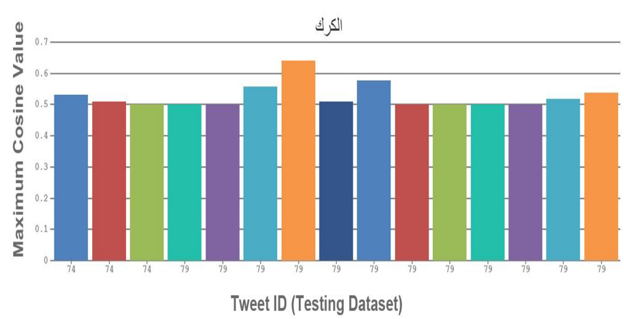

FIG.10 THE DISTRIBUTION OF USERS' TWEETS IN THE SAME HASHTAG

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 26-33 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp18.pdf

Comparing RD with TD using the program, results in values of Cosine similarity. Repeating the comparison manually produced convergence results.

To verify our work, we traced back the tweets which have higher value of Cosine similarity till we reached to the users where the accounts have a Cosine value equal to or more than (0.5) are selected. Then we gathered the users in one community. Subsequently, we checked the interest of each user using his/her profile which was in accordance with the results found by the method used in this research.

5. CONCLUSION

This work intends to generate graph of users by detecting communities of people on Twitter with common interests, based on tweets Hashtags for specific topics in Arabic language. The system is limited to Arabic tweets with the number of characters for each tweet. The common interests will be detected using the Cosine similarity measure, which promises to add more flexibility for measuring the similarity between nodes TD and RD. Moreover, extracting all available cliques in the community according to the strong connection between TD and RD will lead to more accurate result. The cliques of users will be detected using Depth First Search algorithm. Generally, the evaluation results are reasonable and logical. Generated data which is calculated automatically shows similarity with data in user accounts. Hence, this experiment is proved successful.

6. FUTURE WORK

In the future work we intend to continue our work by applying any maximum clique algorithm to previous data. So we can detect maximum number of users who have the strongest common interests.

Moreover we intend to classify user interests according to categories based on location so that governments and human right organization can make use of it .for example; they can make several studies on society samples. It can also alarm intelligence with any imminent risk.

7. RECOMMENDATIONS

To get benefits from the results of this work, we have to expand the search by checking the latest updates on the relevant researches. This work should be developed to belong to specific space, in order to make this work more accurate. In addition to become more comprehensive it should include English words.

References

[1] Du, N., Wu, B., Xu, L., Wang, B. and Pei, X. **A parallel algorithm for enumerating all maximal cliques in complex network**. In *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)* (pp. 320-324). IEEE, December.2006

[2] Huang, A. **Similarity measures for text document clustering**. In *Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008),Christchurch,NewZealand*,April.2008, (pp. 49-56),

[3] Fortunato, S**. Community detection in graphs**. Physics reports, 486(3), pp.75-174, 2010

[4] Kwak, H., Lee, C., Park, H. and Moon, S. **What is Twitter, a social network or a news media?**. In Proceedings of the 19th international conference on World wide web, , April.2010, (pp. 591-600). ACM

[5] Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Yan, E., Li, J. and Dong, T. **Community-based topic modeling for social tagging**. In Proceedings of the 19th ACM international conference on Information and knowledge management, October.2010, (pp. 1565-1568). ACM

[6] Rajawat, S., Hemrajani, N. and Menghani, E. **Solving Maximal Clique Problem through Genetic Algorithm**. In INTERNATIONAL CONFERENCE ON METHODS AND MODELS IN SCIENCE AND TECHNOLOGY (ICM2ST-10) Vol. 1324, No. 1, November.2010, pp. 235-238. AIP Publishing.

[7] Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M.A. and Benevenuto, F. **Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In Proceedings of the Workshop on Languages in Social Media**. June.2011, (pp. 58-65). Association for Computational Linguistics.

[8] Jin, X., Wang, C., Luo, J., Yu, X. and Han, J**. LikeMiner: a system for mining the power of'like'in social media networks**. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, August.2011, (pp. 753-756). ACM.

[9] Poblete, B., Garcia, R., Mendoza, M. and Jaimes, A. **Do all birds tweet the same?: characterizing twitter around the world**. In Proceedings of the 20th ACM international conference on Information and knowledge management, October.2011, (pp. 1025-1030). ACM.

[10] Wang, X., Wei, F., Liu, X., Zhou, M. and Zhang, M. **Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach**. In Proceedings of the 20th ACM international conference on Information and knowledge management, October.2011, (pp. 1031-1040). ACM.

ISSN 2319-7595

International Journal of Information Systems and Computer Sciences, Vol.7. No.1, Pages : 26-33 (2017)
Special Issue of ICSIC 2017 - Held during 23-24 September 2017 in Amman Arab University, Amman-Jordan
http://www.warse.org/IJISCS/static/pdf/Issue/icsic2017sp18.pdf

[11] Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z. and Zha, H. **Like like alike: joint friendship and interest propagation in social networks**. In Proceedings of the 20th international conference on World wide web, March.2011, (pp. 537-546).ACM.

[12] Lim, K.H. **Detecting Like-minded Communities with Common Interests on Twitter**. 18th School of Computer Science & Software Engineering Research Conference, Fairbridge, Pinjarra, Western Australia, November.2012

[13] Lim, K.H. and Datta, A**. Following the follower: detecting communities with common interests on twitter**. In Proceedings of the 23rd ACM conference on Hypertext and social media, June.2012a ,(pp. 317-318). ACM.

[14] Lim, K.H. and Datta, A. **Finding twitter communities with common interests using following links of celebrities**. In *Proceedings of the 3rd international workshop on Modeling social media*, June. 2012b ,(pp. 25-32). ACM.

[15] Soleimani-Pouri, M., Rezvanian, A. and Meybodi, M.R. **Finding a maximum clique using ant colony optimization and particle swarm optimization in social networks.** In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)* (pp. 58-61). IEEE Computer Society.

[16] Dongen, S. and Enright, A.J. **Metric distances derived from cosine similarity and pearson and spearman correlations**. *arXiv preprint arXiv:1208.3145*, 2012

[17] *"Top Sites". Alexa Internet. Retrieved May* 13, *2013. Available from:* http://www.alexa.com/topsites. Accessed November 7, 2016.

[18] Pattabiraman, B., Patwary, M.M.A., Gebremedhin, A.H., Liao, W.K. and Choudhary, A. **Fast algorithms for the maximum clique problem on massive sparse graphs**. In *International Workshop on Algorithms and Models for the Web-Graph* (pp. 156-169). Springer International Publishing, December.2013

[19] Deshpande, R., Vaze, K., Rathod, S. and Jarhad, T. **Comparative Study of Document Similarity Algorithms and Clustering Algorithms for Sentiment Analysis**. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* ISSN 2278-6856, Volume 3, Issue 5, 2014

[20] Santoro, R., Ferragina, P., Piccinno, F. **On analyzing hashtags in twitter.** 2015

[21] Twitter. Twitter API. Internet, Sep 2016. Available from: https://dev.twitter.com.Accessed December 28, 2016

[22] *"AboutTwitter,Inc.".Twitter.Availablefrom:*https://about.twitter.com/company. Accessed December 28, 2016.