Journal of Computer Science 10 (7): 1269-1280, 2014
ISSN: 1549-3636
© 2014 Science Publications
doi:10.3844/jcssp.2014.1269.1280 Published Online 10 (7) 2014 (http://www.thescipub.com/jcs.toc)

CLUSTERING TWEETS USING CELLULAR GENETIC ALGORITHM

[1]Amr Adel, [2]Essam ElFakharany and [3]Amr Badr

[1,2]Faculty of Management and Technology,
Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt
[3]Faculty of Computers and Information, Cairo University, Cairo, Egypt

Received 2014-01-06; Revised 2014-02-10; Accepted 2014-02-23

ABSTRACT

As the popularity of Twitter continues to increase rapidly, it is extremely necessary to analyze the huge amount of data that Twitter users generate. A popular method of tweet analysis is clustering. Because most tweets are textual, this study focuses on clustering tweets based on their textual content similarity. This study presents tweet clustering using cellular genetic algorithm cGA. The results obtained by cGA are compared with those obtained by generational genetic algorithm in terms of average fitness, average time required for execution and number of generations. Experimental results are tested with two sets: One of 1000 tweets and the second formed of 5000 tweets. The results show a nearly equal performance for both algorithms in terms of the average fitness of the solution. On the other hand, cGA shows a much faster performance than generational. These results demonstrate that cellular genetic algorithm outperforms generational genetic algorithm in tweet clustering.

Keywords: Clustering, Cellular Genetic Algorithm, Twitter, Tweet Similarity

1. INTRODUCTION

The last years witnessed an enormous growth of internet-based social network siteslike Facebook, Google+, Twitter, YouTube, etc. This has transformed the way by which people communicate and interact with others (Wang, 2010). These social media networks produce a massive amount of data that needs to be properly analyzed. Twitter is one of the most important social media platforms. It can be utilized to share thoughts and coordinate activities, like instant messaging (Honey and Herring, 2009). Postings in Twitter cover an extremely broad variety of topics in diverse fields, from daily life, current events, breaking news, political interpretations to product reviews and other interests. These postings can exist in different formats e.g., short sentences, URL links and direct messages to other users. Each tweet is composed of at most 140 characters. The limited length of Tweets regularly means that the tweets do not certainly include well-developed thoughts, instead they are short and concise; however complete enough so that users can understand the ideas delivered by the tweets (Tumasjan, 2010; Sankaranarayanan *et al*., 2009; Java *et al*., 2007).

This study purposes to address the problem of tweet clustering and the use of cellular genetic algorithm to solve this problem and comparing it with a conventional algorithm such as the generational genetic algorithm.

The rest of this study is structured in the following manner. Section 2 includes a brief overview about Twitter and clustering of documents. Section 3 shows the previous related work concerning Twitter analysis and clustering using genetic algorithms. Section 4 is devoted to the discussion of the problem. Section 5 discusses the data set gathering methodology, description and preparation. Section 6 provides a detailed description of the used algorithm. Section 7 shows the experimental results and its interpretation including a comparison between generational and cellular genetic algorithms according to

Corresponding Author: Amr Adel, Faculty of Management and Technology, Arab Academy for Science and Technology and Maritime Transport, Cairo, Egypt

average fitness, execution time in milliseconds (ms) and number of generations. The final section is dedicated to summarize the conclusion and future work.

2. BACKGROUND

2.1. Twitter

Micro-blogging can be defined as "A form of blogging that allows users to send brief text updates or micro media such as photographs or audio clips" (Sakaki *et al*., 2010). Twitter was created as a micro-blogging website in March of the year 2006 and formallyinitiated in July of the same year by Jack Dorsey, Evan Williams, Biz Stone and Noah Glas (Mosley and Roosevelt, 2012). Twitter is considered to be one of the widelyprevalent micro-blogging platforms in which users are able to generate status messages called "tweets", which are status updates and musings that cannot exceed 140 characters (Liang and Dai, 2013). These messages are broadcasted to a global audience (Conover *et al*., 2011).

Twitter popularity is continuing to increase rapidly. This is demonstrated in **Fig. 1** (quoted from 2012 Social Network Analysis Report, http://www.ignitesocialmedia.com/social-media-stats/2012-social-network-analysis-report/#Twitter, Retrieved on December 28, 2013) that displays statistics of the search traffic on Twitter for the year 2012.

Tweets can be posted from various sources including the Twitter website, Twitter mobile applications in addition to several third party applications/websites. Twitter users also have the control over the privacy features. They can choose to make their tweets public (visible to any one) or private (visible to only some users who get permission from the user). If a user's profile is left public, his/her updates appear in a "public timeline" of recent updates (Java *et al*., 2007). Twitter enables ituser to reply to messages of another user(s) by clicking the reply button on their tweet (Goyal, 2011). Every user is recognized by a user name advanced by "@"symbol (Mosley and Roosevelt, 2012).

Social interaction between Twitter users takes place principally in three ways:

- The "follow" relationship where Twitter users can subscribe to other users' tweets. The follower gets all the status updates of the user that he/she follows. Followers are displayed in chronological order; the most recently selected follower is displayed first. Unlike other social networking sites, the relationship of following and being followed does not require interchange. Twitter supports one-way connection rather than two-way connection. In other words, a user can follow another user and the followed user is not required to follow in return
- Another form of connection that can be defined between two users is "Mention". Mention is the event of referring to other user(s) in a tweet by addressing them directly
- "Retweet" or RT in which individuals can re-transmit content created by another Twitter user, hencemaking itmore visibile. This resembles forwarding an e-mail to other users, in this case the followers. Retweet has an important role in the propagation of information on Twitter (Mosley and Roosevelt, 2012)

"Hashtag" is a unique concept on Twitter (Note: Hash tag is furthermore supported by other social media websites such as: Facebook, Google+, Instagram, YouTube, LinkedIn) that enables users to identify significant keywords in their tweets by adding the prefix '#' before a keyword (without space) in a tweet. Hashtags are used on Twitter to set trending topics, indicate intended audience of a tweet, begin chat rooms and categorize tweets by topic or type. The hash tags allow users to emphasize what they think as important keyword (s) in their tweet. A hashtag beforea topic enables users to get tweets relevant to a specific topic during search to retrieve a list of recent tweets about this topic.

In addition, Twitter offers a search portal (https://Twitter.com/Twittersearch) so that users can constantly monitor or search for tweets either by means of keywords, hashtags or user name, but this service is restricted to only 40 search keywords. Also, Twitter has Application Programming Interface (API) functions to acquire user-specific information. Such information can be used to construct a network of friends Sankaranarayanan *et al*. (2009)

Moreover, Twitter provides clickable "trending topic" terms, that initiate searches for widespread keywords.

Finally, Twitter delivers a location service.Users who send tweets usingportable devices, have the ability to switch on their location. Theusers' latitude and longitude arecaught with the tweet.Location information provided by mobile Twitter applications save the geographical location of the user at the time he/she posted the tweet. In general, the user has the alternative to switch location serviceeither on or off (Mosley and Roosevelt, 2012).

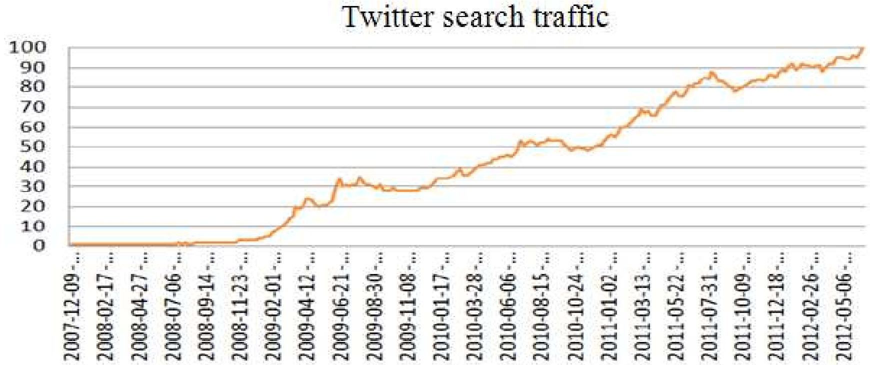


Fig.1. Search traffic on Twitter in 2012

2.2. Clustering

Clustering or cluster analysis is the partitioning of data into collections of similar objects called clusters. Consequently, the samples in a single group are assembledwhile samples of other groups are gathered as a different group. It is a widely used technique for data interpretation and statistical data analysis. Clustering input is a group of data. Clustering process involves measuring similarity and or dissimilarity between data. Clustering output is a group of clusters. Data items in each cluster are similar to each other and dissimilar from items in other clusters. Document clustering can be defined as "Automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters" (Premalatha and Natarajan, 2009, Koteeswaran *et al*., 2012). Similarly, Tweets can be grouped into clusters such that tweets in one cluster tend to be similar to each other, but dissimilar to those in other clusters i.e., minimum inter-cluster and maximum intra-cluster similarity (Mosley and Roosevelt, 2012).

3. RELATED WORK

Twitter analysis is abroad field of research in which researchers have been greatly interested. One of the earliest works in this field is that conducted by Java *et al*. (2007), which focused on studying usage and communities. Conversation and collaboration between users via Twitter was studied by Honey and Herring (2009). One way of analyzing Twitter is cluster analysis of tweets. A lot of studies were performed. Mosley and Roosevelt (2012), applied clustering to 116 keyword indicators extracted from an archive of Twitter insurance posts based on their similarity. A method called Core-Topic-based Clustering (CTC) method was proposed to extract meaningful topics and cluster tweets according to the topics (Kim *et al*., 2012). Another method to improve the accuracy of clustering short text items through the use of Wikipedia as an extrasource of knowledge was proposed (Banerjee *et al*., 2007). Another study exploredhow Twitter can be used to construct a news processing system, from tweets by automatically grouping news tweets into clusters, such that each cluster consists of tweets relating to a particular topic (Sankaranarayanan *et al*., 2009). Perez-Tellez *et al*. (2010), presented and compared a number of different methods based on clustering to determine whether a certain tweet refers to a specific company or not.Application ofk-means clustering technique for masses consisting of a huge number of documents came up with the conclusion that when the documents' content is very short (as in the case of tweets), it is more appropriate to cluster the words instead of the documents. Therefore, a method that clusters the words using the word co-occurrence as a similarity measure was proposed by Khot (2010). Karandikar (2010) used a system for statistical analysis and graphics for clustering tweets based on their topic vectors. He proposed and described a method to determine the most appropriate topic model fortweet clustering. Rangrej *et al*. (2011) compared various document clustering techniques including k-means, SVD-based method and a graph-based approach and compared their performance on short text data collected from Twitter. Tweet Motif that clusters Twitter messages by frequent significant terms was presented by O'Connor *et al*. (2010). Other work has taken into consideration the use of genetic algorithm for cluster analysis of documents. Casillas *et al*. (2003) presenteda genetic algorithm that clusters documents intounidentifiedquantity of clusters. Premalatha and Natarajan (2009) proposed a method for document clustering based on genetic algorithm with Simultaneous mutation operator and ranked mutation rate. Usharani and Iyakutti (2013) proposed anapproachbuilton genetic algorithm for discovering resemblancebetween web documentsdepending on cosine similarity.

To the best of the authors' knowledge, Cellular Genetic Algorithm cGA has not been previously used for clustering tweets. The contribution of this study is the application of cellular genetic algorithm cGA in tweet clustering.

4. THE PROBLEM

In this study, experiments are done with clustering tweets into eight topics defined in advance. The formulation of the problem of clustering tweets based on their similarity is motivated by an essential remark: The tweet content similarity can be used as one of the similarity measures between users where this measure helps to realize whether the users have similar interests. This is an indication of good similarity between users (Goyal, 2011). Since the majority of the user-generated messages on micro-blogging websites are textual information (Liang and Dai, 2013); therefore, the main focus of this study is clustering of tweets based on their textual content similarity. Since English is the most commonly used language in Twitter (Honey and Herring, 2009), the focus is on tweets written in English. Twitter provides a large quantity of short text in the form of tweets where each tweet represents a single document (Rangrej *et al*., 2011).

Goyal (2011) stated that tweet similarity between two users is defined as "the cosine similarity between the documents formed by combining the tweets of a user into one".

Textual contents in Twitter primarily denote tweet text, URLs and hashtags within tweets (Zhang *et al*., 2012) and tweets are considered to be "short texts". Clustering of tweets is a complex problem to solve. The very short length of tweets being only about 140 characters is a problem. Karandikar (2010) stated "Such a short piece of text provides very few contextual clues for applying machine learning techniques". This type of data results in weak performance of most clustering methods due to the informal writing style of tweets that can be full of jargons, misspellings, colloquial and out of vocabulary words with poor grammatical structure (Perez-Tellez *et al*., 2010).

5. DATA AND METHODS

The framework in **Fig. 2** briefly describes the steps of data collection, data preparation, the application of algorithms to the prepared data and comparison of the obtained results. The following sections explain these stepsin details.

5.1. Collection Methodology

For such kind of research, there is no typical dataset available for experimentation. The common practice is to collect datasets from different real world systems (Lu, 2011). Data for this study were gathered using the "Scraping based approach" where Twitter was directly accessed through a web client. The web client is a social network aggregator that pulls content from multiple social networking sites into a single location such that users can access their social network accounts via single interface, without having to sign in to each site alone as shown in **Fig. 3** (quoted from Characterizing user behavior in online social networks, 2009).

Hootsuite.com was selected by the authors to aggregate data. Hootsuite.com is a web site that tracks and archives Twitter posts. To track Twitter messages relevant to a specific topic or user, users canaccess this website and create an archive. This archive will track and archive such Twitter messages. Hoot suite enables users to archive data on social media according towell-defined search criteria. Archives of others can be retrievedonly ifthe archive owner grants an obvious approval. The Hootsuite dashboard is shown in **Fig. 4**

According to Alexa traffic ranks, Hootsuite occupies global rank number 132 as shown in **Fig. 5** (quoted from http://www.alexa.com/siteinfo/hootsuite.com, Accessed on December 29, 2013).

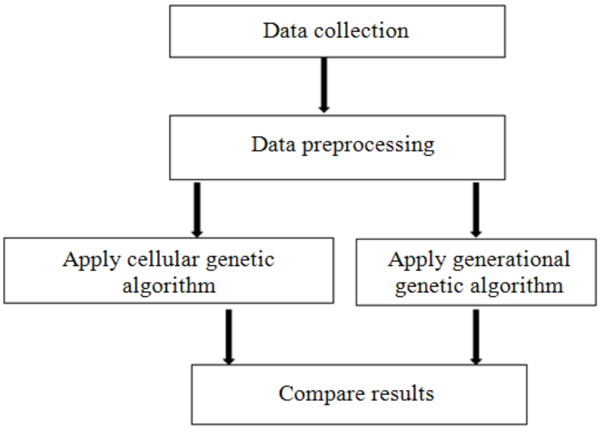

Data collection
Data preprocessing
Apply cellular genetic algorithm
Apply generational genetic algorithm
Compare results

Fig. 2. Framework for data methods, algorithm application and results

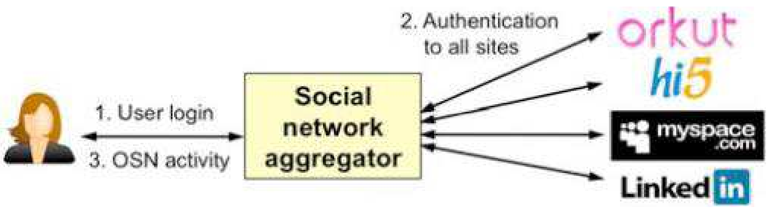
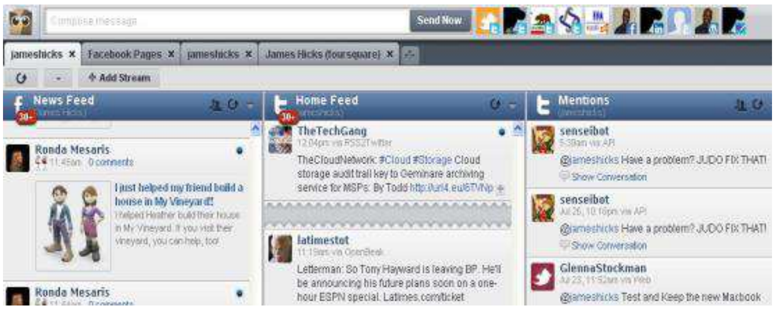


Fig. 3. User interaction with multiple social networks through a social network aggregator



Ronda Mesaris
11:45am 0 comments
I just helped my friend build a house in My Vineyard!
I helped Heather build their house in My Vineyard. If you visit their vineyard, you can help, too!

Ronda Mesaris

TheTechGang
12:04pm via RSS2Twitter
TheCloudNetwork: #Cloud #Storage Cloud storage audit trail key to Geminare archiving service for MSPs: By Todd http://url4.eu/6TVNp

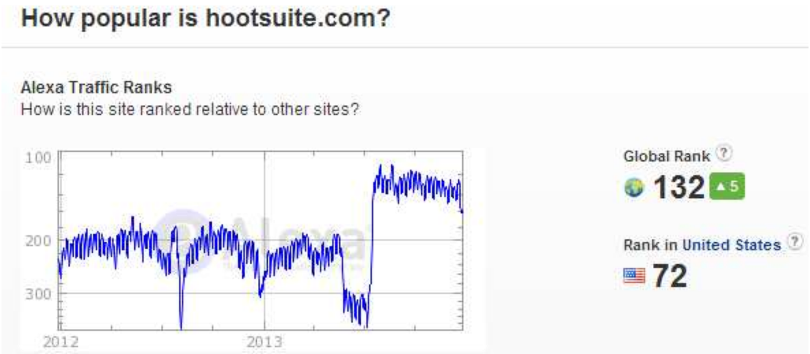
latimestot
11:19am via OpenBeak
Letterman: So Tony Hayward is leaving BP. He'll be announcing his future plans soon on a one-hour ESPN special. Latimes.com/ticket

senseibot
5:39am via API
@jameshicks Have a problem? JUDO FIX THAT!
Show Conversation

senseibot
Jul 26, 10:16pm via API
@jameshicks Have a problem? JUDO FIX THAT!
Show Conversation

GlennaStockman
Jul 23, 11:52am via Web
@jameshicks Test and Keep the new Macbook

Fig. 4. Hootsuite dashboard


How popular is hootsuite.com?

Alexa Traffic Ranks
How is this site ranked relative to other sites?

Global Rank 132 ▲5

Rank in United States 72

Fig. 5. Hootsuite alexa traffic rank on 29th of December 2013

5.2. Data Description and Preparation

For the purpose of this study, tweets were collected based on a set of keywords that describe specific topics in the actual world. Tweets were collected over a 3-day time duration from the 26th of June to the 28th of June 2013. The set of keywords comprise eight variable categories that are intended to be diverse in order to cover different and wide areas of interest: Cinema, Egypt, Film, Hollywood, Iran, Juventus, Messi and Sport.

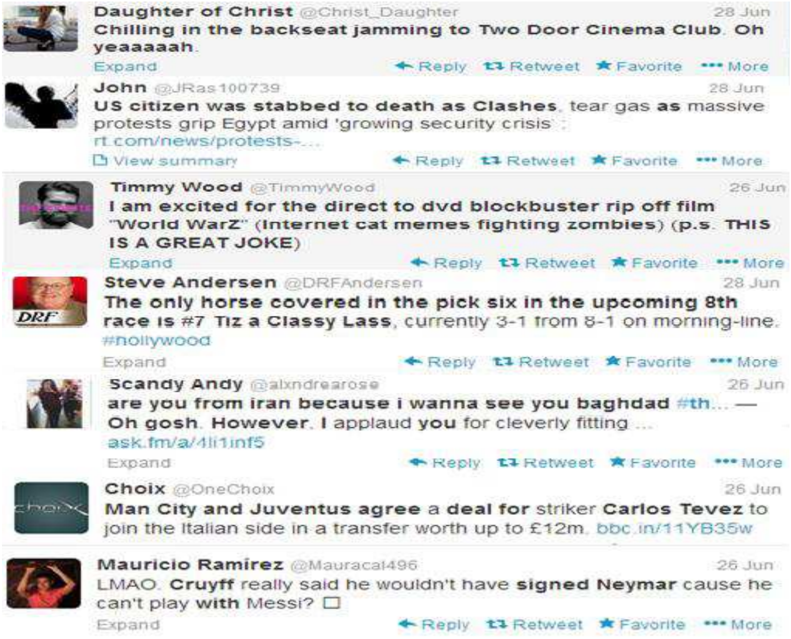
The data gathered using hootsuiteare shown in **Table 1** and sample tweets from the dataset is shown in **Fig. 6**.

The preprocessing of data involved several steps. The first step was the elimination of tweets that:

- Are not in English
- Have too few words (fewer than three)
- Have just a URL
- Duplicate tweets
- All Re-tweets

In addition, all the punctuation and symbols were removed. Such information contains quotation marks, parentheses, punctuation marksplus stray symbols. However, those signs which are really significant for Twitter were kept (@, #).

Two samples of 1000 and 5000 tweets respectively were exploited as test sets over which the experiments were implemented.



Daughter of Christ @Christ_Daughter 28 Jun
Chilling in the backseat jamming to Two Door Cinema Club. Oh yeaaaaah.

John @JRas100739 28 Jun
US citizen was stabbed to death as Clashes, tear gas **as** massive protests grip Egypt amid 'growing security crisis' :
rt.com/news/protests-...

Timmy Wood @TimmyWood 26 Jun
I am excited for the direct to dvd blockbuster rip off film "World WarZ" (Internet cat memes fighting zombies) (p.s. THIS IS A GREAT JOKE)

Steve Andersen @DRFAndersen 28 Jun
The only horse covered in the pick six in the upcoming 8th race is #7 Tiz a Classy Lass, currently 3-1 from 8-1 on morning-line.
#hollywood

Scandy Andy @alxndrearose 26 Jun
are you from iran because i wanna see you baghdad #th... — **Oh gosh. However. I** applaud **you** for cleverly fitting ...
ask.fm/a/4li1inf5

Choix @OneChoix 26 Jun
Man City and Juventus agree a **deal for** striker **Carlos Tevez** to join the Italian side in a transfer worth up to £12m. bbc.in/11YB35w

Mauricio Ramirez @Mauracal496 26 Jun
LMAO. **Cruyff** really said he wouldn't have **signed Neymar** cause he can't play **with** Messi? □

Fig. 6. Sample tweets from the dataset

Table1. Data gathered using hootsuite

The username of the tweet sender
The tweet content
The date and time of tweet posting (according to GMT)
Twitter Identification number of the tweet
Geographic coordinates of the user determining his/her location

6. ALGORITHM

Traditional clustering algorithms were not selected by the authors because such algorithms explorejust a small subset of thepotential clusterings. Thus, the found solution is not guaranteed to be optimal (may get stuck at local minima). Moreover, traditional clustering approaches that require a priori knowledge of the number of clusters, such as K-means, are not suitable to handle large volume of data produced by Twitter and other social media sites (Becker, 2011). Premalatha and Natarajan (2009) used genetic algorithm with ranked mutation operator and demonstrated that it outperforms the traditional algorithms like the K-means. The use of Evolutionary Algorithms (EAs) to handle complicated problems is massive inrecent years. They imitate the biological processes in nature. These algorithms are population-based, which means that they act on a group of prospective solutions (population of individuals) through the application of some operators iteratively to theseindividuals for the sake of finding the finestsolutions. The majority of these algorithms deploys only one population and applies operators to them as a whole (Alba *et al*., 2007). These steps are repeated iteratively until a stopping condition (for example; the maximum number of evaluation limits) is met. The balance (tradeoff) between exploration (diversification) of new solutions and exploitation (intensification) in the search space is an important criterion for performance evaluation of a genetic algorithm and adjusting this tradeoff can improve the overall performance of the algorithm. This tradeoff is represented by "Selection Pressure" which is defined as "A measure of the diffusion speed of the good solutions through the population" (Alba and Dorronsoro, 2008). Reeves (1993) formulated the selection pressure in the following equation:

$$\text{Ø} = \frac{\text{Prob.}(\text{selecting fittest string})}{\text{Prob.}(\text{selecting average string})}$$

Higher exploitation leads to a higher selection pressure as the algorithm tends to converge rapidly to a good enough solution, so it can get stuck to local optimum. Higher exploration leads to a lower selection pressure as the algorithm tends to explore the search space in depth for an optimal solution.

6.1. Cellular Genetic Algorithm

Genetic Algorithms (GAs) imitate the process of natural selection. A population of individuals that represent empirical solutions to a specific problem is preserved. New individuals are then created via reproducing the populationmembers. The new individuals substitute the old ones. Cellular Genetic Algorithms cGAs represent a subclass of Genetic Algorithms where the arrangement of the population is decentralized and the concept of small neighborhood is strongly applied, so the individuals can merely recombine with individuals belonging to its neighbor as shown in **Fig. 7** (Alba *et al*., 2007). Alba and Dorronsoro (2004) stated that “Such a kind of structured algorithms is specially well suited for complex problems”. The existence of small overlapped neighborhoods in Cellular Genetic Algorithms helps preserve a high diversity level for much longerin comparison with other centralized algorithms (Morales-Reyes *et al*., 2009). A behavioral comparison of two different cGAs versus two traditional genetic algorithms, on a large benchmark composed of problems with many different features, revealed that cGA behavior is more robust as it obtains smaller standard deviations than the traditional algorithms. In addition, the cGA shows faster performance (shorter elapsed time) than traditional genetic algorithms (Alba and Dorronso, 2008)

6.2. Chromosome Representation

The population structuretakes the form of a bi-dimensionalgridwith neighborhood defined on it. Each chromosome in the generation represents a candidate tentative solution to the problemand is formed of a sequence of genes. A chromosome is represented as an array of integers of length equal to the number of tweets. Each entry in the array corresponds to a cluster for a tweet. Chromosome reprentation is described in **Fig. 8.**

6.3. Initial Generation

Initial generation is randomly generated from the search space with a fitness value assigned to each individual.

6.4. Fitness Function

The fitness function is used to evaluate the quality of the solution (clustering method). Higher fitness value indicates higher quality of the solution. The used fitness function is a function of cosine similarity. Usharani and Iyakutti (2013) stated that “cosine similarity is a measure of similarity that can be used to compare documents with respect to a given vector of query words. This is quantified as the cosine of angle between vectors”. The function is as follows:

$$\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

6.5. Parent Selection

The aim of the selection operator is to enhance the the population’s quality by granting higher quality individuals a greater possibility to replicate in the following generations. The individual’s quality is evaluated using the fitness function (Khaliessizadeh, 2006). Here, the first parent is selected using the dissimilarity tournament selection operator, while the second parent is chosen by the linear rank selection operator.

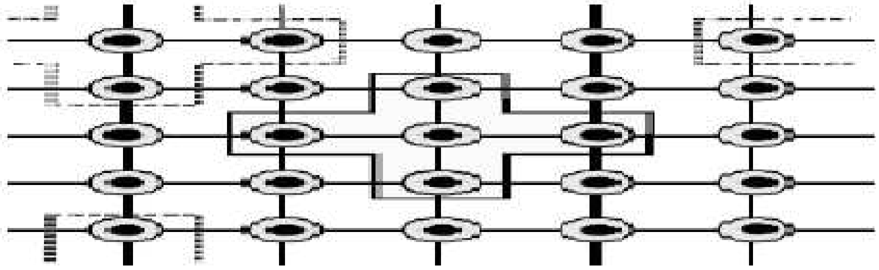
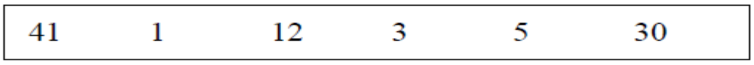


Fig. 7. Cellular genetic algorithm topology



41	1	12	3	5	30

Fig. 8. Representation of chromosome

Table 2. Pseudo-code of Cellular Genetic Algorithm

```
proc evolve (cga)
  GenerateInitialPopulation(cga.pop);
  Evaluation(cga.pop);
  while ! StopCondition() do
    for individual ← 1 to cga.popSize do
      neighbors ← CalculateNeighborhood(cga,position(individual));
      parents ← Selection(neighbors);
      offspring ← Recombination(cga.Pc,parents);
      offspring ← Mutation(cga.Pm,offspring);
      Evaluation(offspring);
      Replacement(position(individual),auxiliary pop,offspring);
    end for
    cga.pop ← auxiliary pop;
  end while
end proc Evolve
```

Table 3. Parameterization of the algorithm

Population size	400 individuals (20*20)
Stopping condition	15,000,000 fitness evaluations
Neighborhood	Linear5
Parent selection	Dissimilarity+ Linear rank
Recombination operator	DPX, $P_c = 1.0$
Mutation operator	Integer mutation, $P_m = 1.0$
Replacement policy	Replace if non worse

Dissimilarity tournament selection operator is an operator that is independent on the relative fitness of the nearby individuals. However, takes into consideration the difference between the respective solutions where two neighbors are chosen in random and the individual which is more dissimilar to the existing individual is chosen. On the other hand, in linear ranking selection, all neighborhood individuals are arranged in order in a list depending on their fitness values, from best to worst, with greater possibility of choosing a parent with a higher rank in this list (Alba and Dorronsoro, 2008).

6.6. Crossover

Recombination (Crossover) operator with a pre-specified crossover probability P_c is applied to the individuals. Here, the applied operator is the two points crossover Distance Preserving Crossover (DPX) operator with P_c=1.0.Theaim of this operator is to produce off springs that have equal distance to every parent. This distance is the sameas the distance in between parents (Misevičius and Kilda, 2005).

6.7. Mutation

Mutation operator with a pre-specified mutation probability P_m is applied to the individuals. Here, the applied operator is the Integer Mutation operator with P_m=1.0. Integer mutation involves the replacement of the integer value of a gene by a new value generated in random (Hugosson *et al*., 2007).

After the application of recombination and mutation operators, fitness value of novel offsprings is calculated. The novel generation replaces the previous one if it is not worse.

6.8. Stopping Criterion

The loop of reproductive cycle is repeated iteratively until the stopping condition is fulfilled. Here, termination occurs when the maximum number of fitness function evaluations (15,000,000 evaluations) is reached.

The pseudo code and parameterization of the algorithm are described in**Table 2** and **Table 3.**

7. RESULTS

As previously mentioned at the end of section 5; the experimental studies were performed over two sets of 1000 and 5000 tweets. The experiments included running each of cellular and generational genetic algorithms for 40 independent runs over the 1000 tweets dataset and 50 independent runs over the 5000 tweets dataset. Both algorithms have been executedusing Java on a single PC 1.90 Ghz under Windows 7 operating system and having 8 GB of memory. The fitness value, execution time and number of generations for each run is recorded and then the average fitness value and execution time (in milli seconds ms) are calculated.

Finally, the values of cellular genetic algorithm are compared to those of generational genetic algorithm to select the most appropriate algorithm that achieves the best fitness i.e., higher quality of clustering at the least time.

Figure 9 and **10** compare the average fitness and average execution time for Cellular and Generational genetic algorithms over the 1000 tweets dataset. **Figure 11** and **12** compare the average fitness and average execution time for Cellular and Generational genetic algorithms over the 5000 tweets dataset. **Figure 13** compares the number of generations produced by each algorithm.

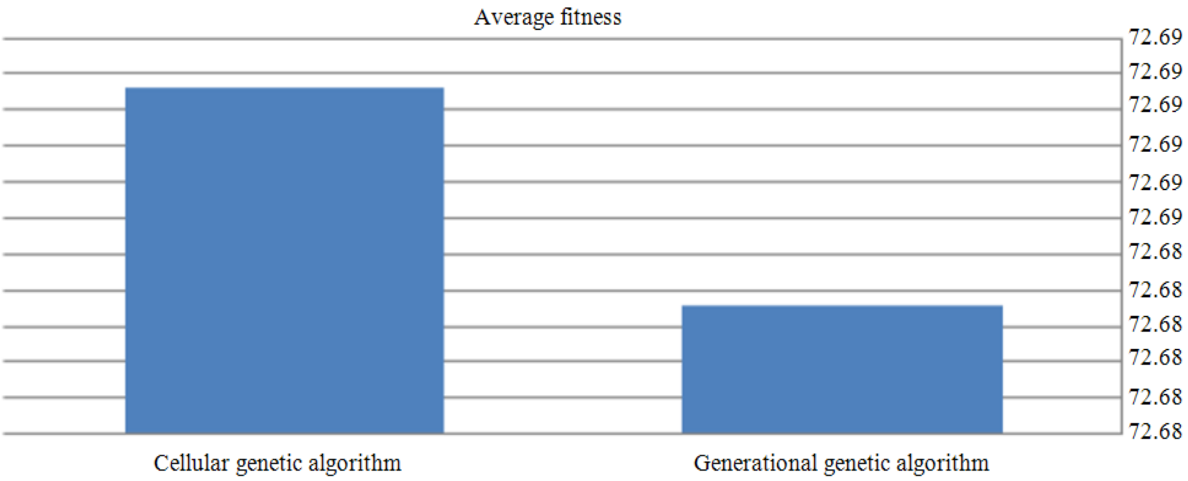


Fig. 9. Average fitness value of generational and cellular genetic algorithms (1000 tweets)

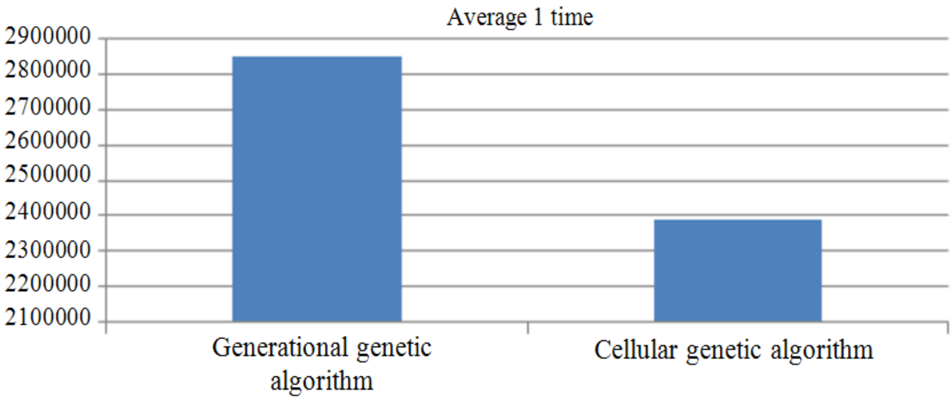


Fig. 10. Average execution time of generational and cellular genetic algorithms (1000 tweets)

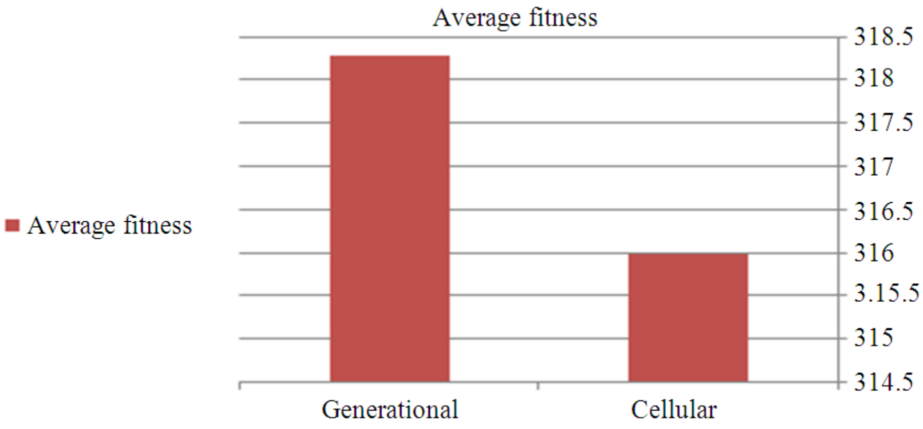


Fig. 11. Average fitness value of generational and cellular genetic algorithms (5000 tweets)

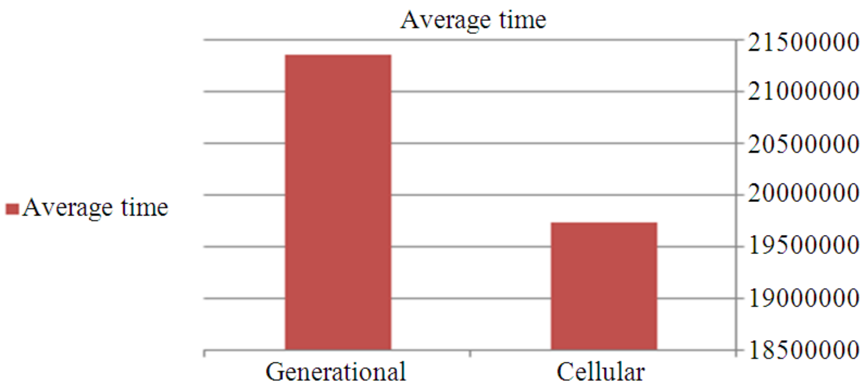


Fig. 12. Average execution time of generational and cellular genetic algorithms (5000 tweets)

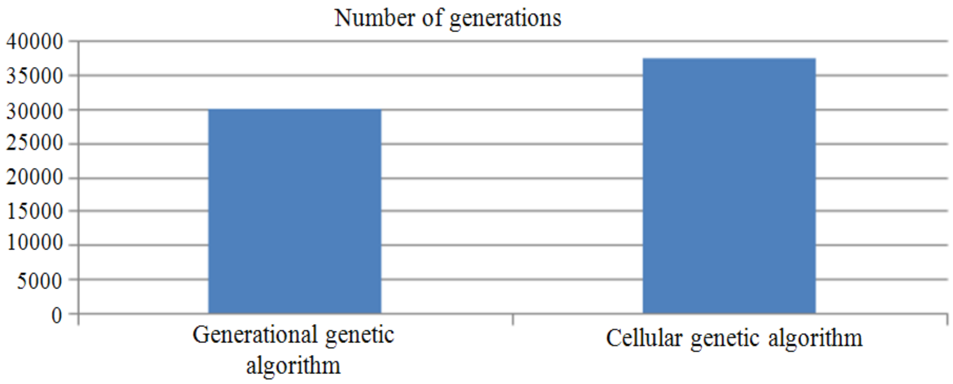


Fig. 13. Number of generations produced by each algorithm

8. DISCUSSION

From the above results, the reader can observe that average fitness of the solutions generated by both algorithms is nearly the same for both sets.The use of small overlapped neighborhood niches in cGA maintains population diversity as it enhances exploration of the search space due to the relatively smooth spread of the finest solutions across the entire population, at the same time exploitation occurs within each neighborhood by genetic operations. In other words, cGA provides a good tradeoff between exploration and exploitation. Therefore; it avoids being stuck into local optima (Alba and Dorronsoro, 2004).

Concerning the average time required for execution, the cellular requires a remarkably shorter time to implement. This also can be attributed to the population structure . The population in cGA is structured into neighborhoods, while it's unstructured in case of generational genetic algorithm. This means that the individual in generational algorithm has to search through the whole population, while cGA individual can interact only with its nearby neighbors

Cellular genetic algorithm gives a larger number of generations than the generational. This means that generational genetic algorithm is more efficient than cGA (as it requires a fewer number ofgenerations to find the solution). The reason is that cGA enhances more exploration, thus induces a lower selection pressure

9. CONCLUSION

This study investigated clustering of tweets based on their textual similarity by the use of cellular genetic algorithm in comparison with the generational genetic algorithm. The experimental tests were performed twice: First; by running each algorithm for 40 independent runs over a set of 1000 tweets. Second; by running each algorithm for 50 independent runs over a set of 5000 tweets.

Comparison between the results of the two algorithms revealed that the quality of the solutions produced by both algorithms (according to the fitness value) is nearly equal, but cGA performs at much shorter time. Therefore, cGA was selected. For future work, the authors plan to test over a larger dataset (composed of 30,000 tweets). Considering the high complexity of the problem, the authors consider the use of parallel computing to minimize the time required for execution

9. REFERENCES

Alba, E. and B. Dorronsoro, 2004. Solving the vehicle routing problem by using cellular genetic algorithms. Proceedings of the Evolutionary Computation in Combinatorial Optimization. Springer Berlin Heidelberg, Apr. 5-7, Springer Berlin Heidelberg, Coimbra, Portugal, pp: 11-20. DOI: 10.1007/978-3-540-24652-7_2

Alba, E. and B. Dorronsoro, 2008. Cellular Genetic Algorithms Electronic Resource. 1st Edn., Springer, New York, ISBN-10: 0387776109, pp: 40-47, 56, 146.

Alba, E., B. Dorronsoro, F. Luna, A.J. Nebro and P. Bouvry, 2007. A cellular multi-objective genetic algorithm for optimal broadcasting strategy in metropolitan MANETs. Comput. Commun., 30: 685-697. DOI: 10.1109/IPDPS.2005.4

Banerjee, S., K. Ramanathan and A. Gupta, 2007. Clustering short texts using wikipedia. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 23-27, ACM New York, NY, USA., pp: 787-788. DOI: 10.1145/1277741.1277909

Becker, H., 2011. Identification and characterization of events in social media. Ph.D. Thesis, Columbia University.

Casillas, A., M.T. Gonzalez De Lena and R. Martinez, 2003. Document clustering into an unknown number of clusters using a genetic algorithm. Proceedings of the 6th International Conference on Text Speech and Dialogue, Sept. 8-12, Springer Berlin Heidelberg, Czech Republic, pp: 43-49. DOI: 10.1007/978-3-540-39398-6_7

Conover, M.D., B. Goncalves, J. Ratkiewicz, A. Flammini and F. Menczer, 2011. Predicting the political alignment of Twitter users. Proceedings of the IEEE 3rd International Conference on Privacy, Security, Risk and Trust, Oct. 9-11, IEEE Xplore Press, Boston, MA., pp: 192-199. DOI: 10.1109/PASSAT/SocialCom.2011.34

Goyal, P., 2011. Semester project report semester project report data mining and analysis on Twitter data mining and analysis on Twitter.

Honey, C. and S.C. Herring, 2009. Beyond microblogging: Conversation and collaboration via Twitter. Proceedings of the 42nd Hawaii International Conference on System Sciences, Jan. 5-8, IEEE Xplore Press, Big Island, HI., pp: 1-10. DOI: 10.1109/HICSS.2009.89

Hugosson, J., Erik Hemberg, A. Brabazon and M. O'Neill, 2007. An investigation of the mutation operator using different representations in Grammatical Evolution. Proceedings of the 2nd International Symposium Advances in Artificial Intelligence and Applications, Oct. 15-17, Wisla, Poland, pp: 409-419.

Java, A., X. Song, T. Finin and B. Tseng, 2007. Why we Twitter: Understanding microblogging usage and communities. Proceedings of the 9th Workshop on Web Mining and Social Network Analysis, Aug. 12-15, ACM New York, NY, USA., pp: 56-65. DOI: 10.1145/1348549.1348556

Karandikar, A., 2010. Clustering short status messages: A topic model based approach. MSc Thesis, University of Maryland.

Khaliessizadeh, S.M., 2006. Genetic mining: Using genetic algorithm for topic based on concept distribution. Proceedings of the World Academy of Science, Engineering and Technology, (SET' 06).

Khot, T., 2010. Clustering Twitter feeds using word co-occurrence. University of Wisconsin.

Kim, S.,S. Jeon, J. Kim and P. Young-Ho, 2012. Finding core topics: Topic extraction with clustering on tweet. Proceedigns of the 2nd International Conference on Cloud and Green Computing, Nov. 1-3, IEEE Xplore Press, Xiangtan, pp: 777-782. DOI: 10.1109/CGC.2012.120

Koteeswaran, S., P. Visu and J. Janet, 2012. A review on clustering and outlier analysis techniques in datamining. Am. J. Applied Sci., 9: 254-258. DOI:10.3844/ajassp.2012.254.258

Liang, P.W. and B.R. Dai, 2013. Opinion mining on social media data. Proceedings of the IEEE 14th International Conference on Mobile Data Management, Jun. 3-6, IEEE Xplore Press, Milan, pp: 91-96. DOI: 10.1109/MDM.2013.73

Lu, C., 2011. Exploiting social tagging network for web mining and search. PhD Thesis, Drexel University, Philadelphia, USA.

Misevičius, A. and B. Kilda, 2005. Comparison of crossover operators for the quadraticassignment problem. Inform. Technol. Control, 34: 109-119.

Morales-Reyes, A., A. Al-Naqi, A.T. Erdogan and T. Arslan, 2009. Towards 3D architectures: A comparative study on cellular GAs dimensionality. Proceedings of Adaptive Hardware and Systems, Jul. 29-Aug. 1, IEEE Xplore Press, San Francisco, CA., pp: 223-229. DOI: 10.1109/AHS.2009.29

Mosley, J. and C. Roosevelt, 2012. Social media analytics: Data mining applied to insurance Twitter posts. Casualty Actuarial Society E-Forum.

O'Connor, B., M. Krieger and D. Ahn, 2010. Tweet motif: Exploratory search and topic summarization for Twitter. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, May 23-26.

Perez-Tellez, F., D. Pinto, J. Cardiff and P. Rosso, 2010. On the difficulty of clustering company tweets. Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, Oct. 26-30, ACM New York, NY, USA., pp: 95-102. DOi: 10.1145/1871985.1872001

Premalatha, K. and A.M. Natarajan, 2009. Genetic algorithm for document clustering with simultaneous and ranked mutation. Modern Applied Sci., 3: 75-82. 1

Rangrej, A., S. Kulkarni and A.V. Tendulkar, 2011. Comparative study of clustering techniques for short text documents. Proceedings of the 20th International Conference Companion on World Wide Web, Mar. 28-Apr. 01, ACM New York, NY, USA., pp: 111-112. DOI: 10.1145/1963192.1963249

Reeves, C.R., 1993. Genetic Algorithms. In: Modern Heuristic Techniques for Combinatorial Problems, In: Blackwell Scientific Publications, Reeves, C.R., (Ed.), Oxford, ISBN-10: 0470220791, pp: 151-196.

Sakaki, T., M. Okazaki and Y. Matsuo, 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th International Conference on World Wide Web, Apr. 26-30, ACM New York, NY, USA., pp: 851-860. DOI: 10.1145/1772690.1772777

Sankaranarayanan, J., H. Samet, B.E. Teitler, M.D. Lieberman and J. Sperling, 2009. Twitterstand: News in tweets. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Nov. 04-06, ACM New York, NY, USA., pp: 42-51. DOI: 10.1145/1653771.1653781

Tumasjan, A., 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, (WSM' 10), pp: 178-185.

Usharani, J. and K. Iyakutti, 2013. A genetic algorithm based on cosine similarity for relevant document retrieval. Int. J. Eng. Res. Technol.

Wang, Z., 2010. Social media data analysis: A dynamic perspective. NATO Research and Technology Organisation, Ottawa, Ontario K1A 0K2, Canada.

Zhang, Y., Y. Wu and Q. Yang, 2012. Community discovery in Twitter based on user interests. J. Comput. Inform. Syst., 8: 991-1000.