CLUSTERING TWEETS USING CELLULAR GENETIC ALGORITHM

Amr Adel, Essam ElFakharany and Amr Badr

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

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 sites like 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...
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 formally initiated in July of the same year by Jack Dorsey, Evan Williams, Biz Stone and Noah Glass (Mosley and Roosevelt, 2012). Twitter is considered to be one of the widely prevalent 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 its user to reply to messages of another user(s) by clicking the reply button on their tweet. 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, hence making it more visible. 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 before a 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 using portable devices, have the ability to switch on their location. The users’ latitude and longitude are caught 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 service either on or off (Mosley and Roosevelt, 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 assembled while 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 explored how 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 of k-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 for tweet 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) presented a genetic algorithm that clusters documents.

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 steps in 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 website that tracks and archives Twitter posts. To track Twitter messages relevant to a specific topic or user, users can access this website and create an archive. This archive will track and archive such Twitter messages. Hootsuite enables users to archive data on social media according to well-defined search criteria. Archives of others can be retrieved only if the 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).

![Fig. 2. Framework for data methods, algorithm application and results](image-url)
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 hootsuite are 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 marks plus 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.
6. ALGORITHM

Traditional clustering algorithms were not selected by the authors because such algorithms explore just a small subset of the potential 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 in recent 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 these individuals for the sake of finding the finest solutions. 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:

\[
\varnothing = \frac{\text{Prob.(selecting fittest string)}}{\text{Prob.(selecting average string)}}
\]

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 population members. 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 neighborhood 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 longer in 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 Dorronsor, 2008).

6.2. Chromosome Representation

The population structure takes the form of a bi-dimensional grid with neighborhood defined on it. Each chromosome in the generation represents a candidate tentative solution to the problem and 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 representation 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 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 (Khalieesizadeh, 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.
Table 2. Pseudo-code of Cellular Genetic Algorithm

1. proc evolve (cga)
2. GenerateInitialPopulation(cga.pop);
3. Evaluation(cga.pop);
4. while ! StopCondition() do
5. for individual ← 1 to cga.popSize do
6. neighbors ← CalculateNeighborhood(cga, position(individual));
7. parents ← Selection(neighbors);
8. offspring ← Recombination(cga.Pc, parents);
9. offspring ← Mutation(cga.Pm, offspring);
10. Evaluation(offspring);
11. Replacement(position(individual), auxiliary pop, offspring);
12. end for
13. cga.pop ← auxiliary pop;
14. end while
15. end proc Evolve

Table 3. Parameterization of the algorithm

<table>
<thead>
<tr>
<th>Population size</th>
<th>400 individuals (20*20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stopping condition</td>
<td>15,000,000 fitness evaluations</td>
</tr>
<tr>
<td>Neighborhood</td>
<td>Linear5</td>
</tr>
<tr>
<td>Parent selection</td>
<td>Dissimilarity+ Linear rank</td>
</tr>
<tr>
<td>Recombination operator</td>
<td>DPX, P_c = 1.0</td>
</tr>
<tr>
<td>Mutation operator</td>
<td>Integer mutation, P_m = 1.0</td>
</tr>
<tr>
<td>Replacement policy</td>
<td>Replace if non worse</td>
</tr>
</tbody>
</table>

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. The aim of this operator is to produce off springs that have equal distance to every parent. This distance is the same as 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 executed using 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)
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 of generations 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


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


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