Abstract

Amr A. Abd El-Rehim El-Sayed

CLUSTERING TWEETS USING CELLULAR GENETIC ALGORITHM

Social media has become an essential part of the daily online experience. They enable information sharing and communication between online users. The unprecedented huge amount of user-generated content produced by social media, needs to be analyzed in a proper manner. Twitter has emerged as an extremely popular micro-blogging social media platform in the recent years, with the number of registered twitter accounts has reached 645,750,000 active accounts by the year 2013, generating an average number of 58 million tweets per day (according to the statistical analysis website statisticbrain.com). As the popularity of Twitter continues to increase rapidly, it is extremely necessary to analyze the huge amount of data that Twitter users generate. Twitter is an essential source of real time information in a wide variety of interests including sports events, advertising, political campaigns, mass emergencies, crisis events, health care, etc. A popular method of tweet analysis is clustering. Because most of the posted Twitter messages are textual in nature, this study focuses on clustering tweets based on their textual content similarity. Moreover, since the English language is the most popular language over Twitter (34% of all tweets are in English according to the Report generated by the British-American news website, technology and social media blog "Mashable"), the study focuses on clustering Twitter messages written in English. The Scraping based technique was employed in this study to gather data from Twitter, using Hootsuite as a social network aggregator. The gathered Twitter messages were collected over a 3-day time duration from the 26th to the 28th of June 2013, based on a set of keywords that describe diverse specific topics in the actual world, in order to cover wide areas of interest. Experimental studies were performed over three datasets of different sizes. Genetic Algorithms belong to the class of evolutionary computational algorithms, which are population based optimization techniques designed for finding globally optimal solutions from a pool of feasible solutions (individuals). Genetic Algorithms are probabilistic search methods whose mechanisms are analogous with the natural process of biological evolution to discover solutions to problems. A subclass of Genetic Algorithm: Cellular Genetic Algorithm was used in the study to cluster tweets. Based on the literature review this study is one of the earliest attempts for tweet clustering through the use of Cellular Genetic Algorithm cGA, which can improve the performance of clustering in comparison with the traditional clustering algorithms, those clustering algorithms that require a priori knowledge of the number of clusters such as K-means. The results obtained by cGA are compared with those obtained by a conventional Genetic Algorithm: Generational Genetic Algorithm genGA. The comparison takes place according to four parameters: the average fitness value, the average time required for execution, the number of generated clusters, in addition to the number of generations. The obtained results indicate a better overall performance of cGA in comparison to genGA.