

Chapter 63

Parallel Hybrid BBO Search Method for Twitter Sentiment Analysis of Large Scale Datasets Using MapReduce

Ashish Kumar Tripathi

Delhi Technological University, New Delhi, India

Kapil Sharma

Delhi Technological University, New Delhi, India

Manju Bala

Indraprastha College of Women, New Delhi, India

Abstract

Sentiment analysis is an eminent part of data mining for the investigation of user perception. Twitter is one of the popular social platforms for expressing thoughts in the form of tweets. Nowadays, tweets are widely used for analyzing the sentiments of the users, and utilized for decision making purposes. Though clustering and classification methods are used for the twitter sentiment analysis, meta-heuristic based clustering methods has witnessed better performance due to subjective nature of tweets. However, sequential meta-heuristic based clustering methods are computation intensive for large scale datasets. Therefore, in this paper, a novel MapReduce based K-means biogeography based optimizer(MR-KBBO) is proposed to leverage the strength of biogeography based optimizer with MapReduce model to efficiently cluster the large scale data. The proposed method is validated against four state-of-the-art MapReduce based clustering methods namely; parallel K-means, parallel K-means particle swarm optimization, MapReduce based artificial bee colony optimization, dynamic frequency based parallel k-bat algorithm on four large scale twitter datasets. Further, speedup measure is used to illustrate the computation performance on varying number of nodes. Experimental results demonstrate that the proposed method is efficient in sentiment mining for the large scale twitter datasets.

DOI: 10.4018/978-1-6684-6303-1.ch063

INTRODUCTION

From last one decade, enormous growth of the digital data has been observed (Gantz J, 2018). The social sites such as Instagram, Face-book, Twitter etc., are the major source of the digital data. Such huge availability of data has attracted the user based industries to analyze the sentiments of users for making business strategies. Thus, efficient data mining methods are required for the sentiment analysis of the social media. Twitter, one of the popular social media, provides a prodigious platform for the sentiment analysis. Twitter database has approximately 200 millions of users and nearby 400 million tweets are posted every day. Often, user shares their personal experiences about products or companies. Since, the maximum length of the tweet is 140 characters. Therefore, some short symbols like emoji are available for expressing the sentiments. The study of the tweets can deliver profound viewpoints and emotions about any subject (Asur, S., & Huberman, B. A. 2010). Sentiment analysis methods are mainly classified into three categories: machine learning based methods, hybrid methods and lexicon based methods (Medhat, W., Hassan, A., & Korashy, H. 2014). The methods based on lexicon require prior knowledge of sentiment lexicon to predict the sentiment. However, for the short-hand and emoji based texts, lexicon-based methods fail to perform well (Khan, A. Z., Atique, M., & Thakare, V. M. 2015). Pandey et al. (Pandey, A. C., Rajpoot, D. S., & Saraswat, M. 2017) proposed a hybrid cuckoo search based method for Twitter sentiment analysis and concluded that emoticons are good predictors of the sentiments for short texts. Further, Canuto et al. (Canuto, S., Gonçalves, M. A., & Benevenuto, F. 2016) used the meta level features for prediction of sentiments. Bravo et al. (Bravo-Marquez, F., Mendoza, M., & Poblete, B. 2013) proposed a supervised approach to amalgamate the strengths of emotions and polarities for revamping the twitter opinion prediction. Furthermore, Mohammad et al. (Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. 2015) employed the supervised classifier to analyze emotion stimulus, emotion state, and intent of tweets for the US election. An ontology-based method was introduced for sentiment analysis of tweets by Kontopoulos et al. (Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. 2013) where a sentiment grade was allocated to each different perception in the tweet. Agarwal et al. (Agarwal, B., Mittal, N., Bansal, P., & Garg, S. 2015) presented a novel approach using common sense information taken from ConceptNet based ontology method for the sentiment analysis. Furthermore, SentiCircle method was introduced by Saif, et al. (Saif, H., He, Y., Fernandez, M., & Alani, H. 2016), in which context specific polarity of words was determined. Qiu et al. (Qiu, G., Liu, B., Bu, J., & Chen, C. 2009, July) proposed a semi-supervised algorithm for sentiment mining. Furthermore, Pandarachalil et al. (Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. S. 2015) introduced a distributed unsupervised approach for extraction of the lexicons. Likewise, Fernandez et al. (Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. 2016) proposed an unsupervised algorithm to predict the sentiment polarity of the informal texts using linguistic sentiment propagation model.

Recently, natural language processing (NLP) is used to study the semantics in the features for improving the efficiency of the methods (Kanakaraj, M., & Guddeti, R. M. R. 2015, March). Altnel et al. (Altnel, B., & Ganiz, M. C. 2016) introduced semantic smoothing kernels for extracting class specific semantics using vector space models (VSM). Muhammad et al. (Muhammad, A., Wiratunga, N., & Lothian, R. 2016) proposed a lexicon based method using textual neighborhood for social media opinion mining. Further, Appel et al. (Appel, O., Chiclana, F., Carter, J., & Fujita, H. 2016) introduced hybridized approach based on NLP and fuzzy for the semantic polarity classification. Chen et al. (Chen, T., Xu, R., He, Y., Xia, Y., & Wang, X. (2016)) used the sequence modeling based neural network for the

16 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/parallel-hybrid-bbo-search-method-for-twitter-sentiment-analysis-of-large-scale-datasets-using-mapreduce/308541

Related Content

Partially Supervised Text Categorization

Xiao-Li Li (2009). *Handbook of Research on Text and Web Mining Technologies* (pp. 75-95).

www.irma-international.org/chapter/partially-supervised-text-categorization/21718

XML Mining for Semantic Web

Rafael Berlanga and Victoria Nebot (2013). *Data Mining: Concepts, Methodologies, Tools, and Applications* (pp. 625-649).

www.irma-international.org/chapter/xml-mining-semantic-web/73461

Constrained Cube Lattices for Multidimensional Database Mining

Alain Casali, Sébastien Nedjar, Rosine Cicchetti and Lotfi Lakhal (2010). *International Journal of Data Warehousing and Mining* (pp. 43-72).

www.irma-international.org/article/constrained-cube-lattices-multidimensional-database/44958

SpyNetMiner: An Outlier Analysis to Tag Elites in Clandestine Social Networks

S. Karthika, S. Bose and A. Kannan (2014). *International Journal of Data Warehousing and Mining* (pp. 32-54).

www.irma-international.org/article/spynetminer/106861

Distributed Privacy Preserving Clustering via Homomorphic Secret Sharing and Its Application to (Vertically) Partitioned Spatio-Temporal Data

Can Brochmann Yildizli, Thomas Pedersen, Yucel Saygin, Erkey Savas and Albert Levi (2011). *International Journal of Data Warehousing and Mining* (pp. 46-66).

www.irma-international.org/article/distributed-privacy-preserving-clustering-via/49640