World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:10, 2017

Extracting Opinions from Big Data of Indonesian Customer Reviews Using Hadoop MapReduce

Authors : Veronica S. Moertini, Vinsensius Kevin, Gede Karya

Abstract : Customer reviews have been collected by many kinds of e-commerce websites selling products, services, hotel rooms, tickets and so on. Each website collects its own customer reviews. The reviews can be crawled, collected from those websites and stored as big data. Text analysis techniques can be used to analyze that data to produce summarized information, such as customer opinions. Then, these opinions can be published by independent service provider websites and used to help customers in choosing the most suitable products or services. As the opinions are analyzed from big data of reviews originated from many websites, it is expected that the results are more trusted and accurate. Indonesian customers write reviews in Indonesian language, which comes with its own structures and uniqueness. We found that most of the reviews are expressed with “daily language”, which is informal, do not follow the correct grammar, have many abbreviations and slangs or non-formal words. Hadoop is an emerging platform aimed for storing and analyzing big data in distributed systems. A Hadoop cluster consists of master and slave nodes/computers operated in a network. Hadoop comes with distributed file system (HDFS) and MapReduce framework for supporting parallel computation. However, MapReduce has weakness (i.e. inefficient) for iterative computations, specifically, the cost of reading/writing data (I/O cost) is high. Given this fact, we conclude that MapReduce function is best adapted for “one-pass” computation. In this research, we develop an efficient technique for extracting or mining opinions from big data of Indonesian reviews, which is based on MapReduce with one-pass computation. In designing the algorithm, we avoid iterative computation and instead adopt a “look up table” technique. The stages of the proposed technique are: (1) Crawling the data reviews from websites; (2) cleaning and finding root words from the raw reviews; (3) computing the frequency of the meaningful opinion words; (4) analyzing customers sentiments towards defined objects. The experiments for evaluating the performance of the technique were conducted on a Hadoop cluster with 14 slave nodes. The results show that the proposed technique (stage 2 to 4) discovers useful opinions, is capable of processing big data efficiently and scalable.

Keywords : big data analysis, Hadoop MapReduce, analyzing text data, mining Indonesian reviews
Conference Title : ICCIT 2017 : International Conference on Computing and Information Technology
Conference Location : Bali, Indonesia
Conference Dates : October 23-24, 2017