

NAIVE BAYES ALGORITHM FOR TWITTER SENTIMENT ANALYSIS AND ITS IMPLEMENTATION IN MAPREDUCE

Zhaoyu Li

Dr. Yi Shang, Thesis Supervisor

ABSTRACT

The scale of social network data generated and processed is increasing exponentially in the Big Data era. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document, and the sentiment analysis on Twitter has also been used as a valid indicator of stock prices in the past. Naive Bayes is an algorithm to perform sentiment analysis. MapReduce programming model provides a simple and powerful model to implement distributed applications without having deeper knowledge of parallel programming. When a new hypothetical MapReduce sentiment analysis system is built to provide certain performance goal, we are lack of the benchmark and the traditional trial-and-error solution is extremely time-consuming and costly.

In this thesis we implemented a prototype system using Naive Bayes to find the correlation between the geographical sentiment on Twitter and the stock price behavior of companies. Also we implemented the Naive Bayes sentiment analysis algorithm in MapReduce model based on Hadoop, and evaluated the algorithm on large amount of Twitter data with different metrics. Based on the evaluation results, we provided a comprehensive MapReduce performance prediction model for Naive Bayes based sentiment analysis algorithm. The prediction model can predict task execution performance within a window, and can also be used by other MapReduce systems as a benchmark in order to improve the performance.