Public Abstract
First Name: Zhaoyu
Middle Name:
Last Name: Li
Adviser's First Name: Yi
Adviser's Last Name: Shang
Co-Adviser's First Name:
Co-Adviser's Last Name:
Graduation Term: FS 2014
Department: Computer Science
Degree: MS
Title: Naive Bayes Algorithm For Twitter Sentiment Analysis And Its Implementation In MapReduce

Data has been growing exponentially in recent years. With the development of information highway, data can be generated and collected very fast, and the data is so large that it has exceeded the limit of our conventional processing methods and applications. The social network is one of many data explosion areas. Among all social network medias, Twitter has become one of the most important platforms to share and communicate with friends.

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.