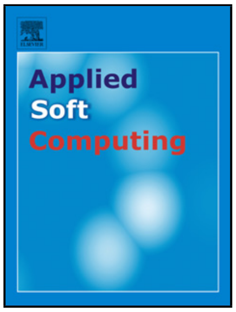
Accepted Manuscript



Title: An Effective Hot Topic Detection Method for Microblog on Spark

Author: Wei Ai Kenli Li Keqin Li

PII: S1568-4946(17)30535-5
DOI: https://doi.org/doi:10.1016/j.asoc.2017.08.053
Reference: ASOC 4443

To appear in: *Applied Soft Computing*

Received date: 26-9-2016
Revised date: 25-8-2017
Accepted date: 28-8-2017

An Effective Hot Topic Detection Method for Microblog on Spark

Wei Ai[a], Kenli Li[a], Keqin Li[a,b]

[a]School of Information Science and Engineering, National Supercomputing Center in Changsha, Hunan University, Changsha, 410082, China
[b]Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

Abstract

With the emergence of the big data age, methods for quickly and accurately obtaining valuable hot topics from the vast amount of digitized textual material have attracted much attention. In this work, we focus on topic detection in microblogs in the big data environment. Different from existing approaches, we solve this problem in a distributed way. Specifically, we propose a non-iterative algorithm called parallel two-phase mic-mac hot topic detection (TMHTD), and implement it in the Apache Spark environment. The proposed TMHTD method includes two phases, i.e., the micro-clustering phase and the macro-clustering phase. To improve the accuracy of hot topic detection, three optimization methods, along with TMHTD, are proposed. To handle large databases, we deliberately design a group of MapReduce jobs to concretely accomplish hot topic detection in a highly scalable way. We compare the TMHTD algorithm with the general single-pass algorithm and the Latent Dirichlet Allocation (LDA) algorithm. Our experiments are carried out on real-life data sets gathered from the Sina Weibo API. Extensive experimental results indicate that the accuracy and performance of the TMHTD algorithm are significant improvements over previous methods. More specifically, the F-measure value of the TMHTD algorithm shows a 6% and 8% improvement over the general single-pass algorithm and the LDA algorithm, respectively. The run time of the TMHTD algorithm is 7 times and twice as superior to the general single-pass algorithm and the LDA algorithm, respectively.

Keywords:
Big data, hot topic detection, microblog, non-iterative clustering algorithm, Spark

1. Introduction

1.1. Motivation

With the advent of the big data era, the amount of data available has increased on a large scale in various fields. Every day, 2.5 quintillion bytes of data are created, and 90 percent of the data in the world today has been produced within the past two years [1]. The vast amount of digitized textual material is now available on the internet, but high-speed connectivity and the explosion in the volume of digitized textual content available online have given rise to information overload [2]. Clearly, although there are large amounts of worthwhile information contained in the huge quantities of data, it is impossible for people to absorb all pertinent information because humans have a limited capacity to assimilate information. Therefore, these characteristics make it extremely challenging to explore the large volumes of data and extract useful information or knowledge quickly and accurately from big data with data-mining methods [3].

Topic detection (TD) has received considerable attention and has been widely accepted as a promising research issue in data-mining to meet this challenge. TD can be seen as a clustering of events that helps analysts separate the wheat from the chaff in massive textual materials. By exploring and organizing the content of textual materials, TD attempts to identify topicsthat will enable us to aggregate disparate pieces of information into manageable clusters automatically. At the same time, hot topic detection provides a way to effectively track hot events and focus on important topics.

Email addresses: aiwei@hnu.edu.cn (Wei Ai), lkl510@263.net (Kenli Li), lik@newpaltz.edu (Keqin Li)