**Abstract**

According to CNNIC’ 38th China Internet Network Development Statistics Report, by the end of June 2016, the number of Chinese netizens has reached 710 million, which is an increase of 21.32 million compared to the end of 2015; Internet penetration rate rose to 51.7%, which is 1.3 percent higher than that of the end of 2015. It can be said that the Internet has become one of the mainstream media in the fields of journalism and communication, which have the enormous influence and the most development potential. It is gradually becoming one of the main carriers to reflect the social public opinion. As can be seen from these data, internet has been integrated into our daily life, but we should be aware that network, an open public opinion platform, works both ways. On one hand, it provides an open environment of public opinion, and it can collect a variety of socially useful comments and suggestions, to play the role of public opinion supervision. But on the other hand, the web information resources inevitably involve a large number of false and illegal information, which misdirects netizens' consciousness. Therefore, public opinion should be processed and analysed digitally, and its macro-dynamic is timely concerned, to grasp the social potential adverse tendency. At the same time, public opinion should be effectively collected and sorted out, in order to further guide and control the consensuses.

Because the huge amount of information and its quick update on the internet, public opinion mining encounters enormous challenges in dealing with massive and complex data, which is constantly updated and supplemented. Due to the high complexity of dealing with these data, traditional computer processing capacity is hard to meet this requirement. Distributed computing technology is needed to achieve large-scale parallel computing.

MapReduce is a parallel programming model and computational framework for processing massive data. It is used for parallel computing on large-scale data level (typically greater than 1 TB). The basic idea of MapReduce is “divide and rule”, that is to say, operation on large-scale data is distributed to sub-nodes to complete commonly, which are under the management of a master node. Then the intermediate results of sub-nodes are integrated to get the final result. Note that, the data sets(or tasks) handled by MapReduce framework must have such characteristics: data sets, which is to be processed, can be broken down into many small data sets, and each small data set can be completely processed in parallel.

In my research, Spark will be used to deal with massive and complex data based MapReduce architecture. Spark is an open source cluster computing framework originally developed in the AMPLab at University of California, Berkeley but was later donated to the Apache Software Foundation where it remains today. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's multi-stage in-memory primitives provides performance up to 100 times faster for certain applications.

As the rapid advancement of Internet and multimedia technologies, social media websites become an important platform for people to access the interested information and share opinions. Concretely, media data emerged from multiple sources, such as news websites, video/photo sharing websites and social network websites; these media data represent different aspects of some real-world topics that people concern and give users a variety of experiences. However, the growing volume of media data makes it difficult for people to find what they are interested in. If topics are discovered from the large amounts of data automatically, users can know what is happening and quickly access the information they concern.

Topic detection is such an effort to discover topics from a collection of documents and group the documents belonging to the same subject. Topic detection for cross-media data integrates data with different modalities from multiple sources and detects hot topics implied in it. Although NIST proposed TDT in the 1990s, so far, most related works only focus on single-source media data, such as news data or social network status text data. Apparently, it is difficult to understand the world and conceive the topics if we do not exploit the media data from multiple sources. However, the traditional TDT approaches are barely suited for cross-media topic detection for the following reasons. The cross-media data have different modalities, and the data representation of each modality varies greatly; besides, the characteristic of the same modality in different sources is not the same. For example, textual information in news articles is abundant, while it (tag information) is sparse and noisy in web videos. To accomplish the goal, we need to overcome three distinct challenges. (1) The overwhelming data in all platforms are inevitably mixed with plenty of useless data, such as daily charts and uninformative messages. This makes the step of emerging data extraction necessary. (2) Since the data are in multiple modalities and from some heterogeneous yet complementary media platforms, revealing the relationships among those data is much more critical and challenging. (3) Based on the extracted emerging data and data relationships, a robust and effective topic detection and elaboration method is expected.

The following is a brief summary of the main contents of my work:

* Build a distributed cluster and build Spark platform for my research.
* Build distributed network crawler system and ensure the normal operation of the system on a daily basis.
* Study master idea, the theoretical basis, the steps, the advantages, the disadvantages and the use of scenarios of the algorithm that has higher frequencies in the actual production. Analyzing problems that we will encounter during processing large data and proposing solutions.
* Study and propose relevant TDT algorithm parallelization scheme, the proposed scheme to be able to solve the problem with TDT algorithm when dealing with a large-scale text data. Also, specific coding implementation.
* Study and propose a scheme on emerging TDT and elaboration using multimedia streams cross different online platforms. Also, specific coding implementation.
* Assess the performance of the proposed scheme, the advantages and disadvantages of this issue.

**Keywords**: Topic Detection and Tracking, Parallel Computing, Cross-Media