IRP proposal

Pengfei Gao, s1144374

March 7, 2012

Abstract

In the recent days, the social networking media has been widely used by more and more people. There is a growing need to develop a powerful and efficient system tools to mining the useful and important information from the huge amount of data generated by users from those social networking media. The aim of this project is to develop such a real-time First Story Detection system to detect current hot events from the real-time stream of Twitter. The project will use Apache Distributed Stream Computing platform(S4) to process continuous unbounded streams of data and locality-sensitive hashing method to extract the features from the tweets.

1 Purpose

-a statement of the problem to be addressed. This should include arguments as to why solving the problem is important; e.g., because it will enable certain applications, or lead to interesting scientific discoveries.

Twitter is a popular real-time information network that users can present and find the latest ideas, stories, news and opinions on what they interested. For example, when the news the pop star Michael Jackson's death came out, 22.64% of the tweets posted contains the phrase "Michael Jackson", which enables the detection of the current hot events promptly.

2 Background

-a short description of how previous work addresses (or fails to address) this problem, leading to a rationale for the hypotheses that you intend to test, and a convincing argument about how that hypotheses might solve the problem.

3 Methods

-A description of the methods and techniques you intend to use to test your hypotheses (e.g., data analysis procedures, experimental design etc), indicating that alternatives have been considered and ruled out on sound scientific grounds.

"First Story Detection" deals with spotting breaking news. For example, as soon as an earthquake happens, we want to know about it. We do not care about follow-up stories.

Now, Twitter can be an excellent source of first stories and we have shown how they can be spotted using scalable algorithms based upon locality sensitive hashing. However, our system only works on a single processor and is not real-time.

S4 (the "real time hadoop" by Yahoo) is a distributed processing system which spreads computation over a network of machines. Tasks communicate with each other via remote procedure calls (there is no use of disks) using key-value pairs. Because of this, processing is scalable and fast. S4 is written in java.

This project will look at how S4 can be used for First Story Detection in Twitter. Tweets will be provided. The focus here will be on investigating how to design such a system, looking at questions such as throughput rate, robustness and scaling and how.

Locality-sensitive hashing

4 Evaluation

-Details of the metrics by which you will evaluate the outcomes of your research; e.g., by comparing the output of your system with some gold standard, or with the ways in which humans perform a task, etc.

Implemented a system based on S4 to spot breaking news in Twitter.

5 Outpus

- A description of what the outputs of the projects will be: e.g., these might include an extension or change to some existing theory or to some piece of software, some new data (e.g., annotated linguistic data), and so on.

6 Workplan

-A timetable or research plan, detailing what will be done to complete the proposed project, and when these tasks will be completed by.

References

[1] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.