Chapter 5
Twitter Data Analysis

Chitrakala S
Anna University, India

ABSTRACT

Analyzing Social network data using Big Data Tools and techniques promises to provide information that could be of use in recommendation systems, personalized service and many other applications. A few of the analytics that do this include sentiment analysis, trending topic analysis, topic modeling, information diffusion modeling, provenance determination and social influence study. Twitter Data Analysis involves analyzing data specifically obtained from Twitter, both tweets and the topology. There are three major classifications on the type of analysis being performed such as Content based, Network based and Hybrid analysis. Trending Topic Analysis in the context of Content based static data analysis and Influence Maximization in the context of Hybrid analysis on data streams using the power of Big Data Analytics are discussed. A novel solution to Trending Topic analysis to generate topic evolved, conflict-free sequential sub summaries and influence maximization to handle streaming data are explained with experimental results.

INTRODUCTION TO TWITTER DATA ANALYSIS

One of the outcomes of the popularity of online social networks is the development of a new field, social network analysis (SNA). This field studies not just the structure of social network but also the behavior of the people who belong to it. One social network that has become popular for analysis is Twitter. Tweets based on a specific topic of interest, once extracted can be analyzed and the results obtained can be used in many applications. Twitter Data Analysis has gained popularity due to few notable reasons. First, obtaining information from Twitter makes it possible for

DOI: 10.4018/978-1-5225-2805-0.ch005

vendors to provide personalized solutions to their customers. Second, unlike other social networks, most accounts of Twitter are public, making it possible to obtain the necessary data. Also, the limitation on the number of characters ensures that the amount of time required to process a single tweet is typically rather small.

Analysis performed on Twitter data can be broadly classified into three categories: Content Based, Network Based and Hybrid Analysis. Techniques which rely solely on the tweets/text produced are named as Content based analysis, whereas techniques that rely on the network structure are called Network based analysis. A combination of both text and structure based analysis is termed as Hybrid analysis. The following sections expose the readers to techniques/methodologies in Twitter Data Analysis and its significance.

Overview

In this chapter, it is intended to show how analytical techniques namely Trending Topic Analysis and Influence Maximization can be utilized to study and mine significant information from a social network such as Twitter. Also, to illustrate their applications in real life business value use cases. It is believed that these illustrations would trigger ideas for researchers in various fields.

Firstly, a study on Trending Topic Analysis technique which is a content based static data analysis is emphasized accounting to the urging need of a complete analyzed summary of the topic under interest, presented in a topic evolved manner.

Secondly a study on Influence maximization technique which is a hybrid data analysis is discussed. It is important as it provides a way to find a small set of users, thus reducing the cost of promoting a product or campaign while simultaneously maximizing the spread of word about them. Distinguishing and critical aspect of the proposed Influence Maximization methodology is that it follows a Big Data approach enhancing its significance many folds.

Motivation

It is evident over the recent years that Twitter has grown from a vague invention to become a mainstream medium for dissemination of messages and the public discussion of news and events. The rapid proliferation of Twitter posts presents a big obstacle for efficient information acquisition. It is impossible for a user to get an overview of important topics on Twitter by reading all tweets every day. In addition, because of information redundancy and the informal writing style, it is time consuming to find useful information about a topic from a huge number of tweets. The tremendous