On Summarization and Timeline Generation for Evolutionary Tweet Streams

Short-text messages such as tweets are being created and shared at an unprecedented rate. Tweets, in their raw form, while being informative, can also be overwhelming. For both end-users and data analysts, it is a nightmare to plow through millions of tweets which contain enormous amount of noise and redundancy. In this paper, we propose a novel continuous summarization framework called Sumblr to alleviate the problem. In contrast to the traditional document summarization methods which focus on static and small-scale data set, Sumblr is designed to deal with dynamic, fast arriving, and large-scale tweet streams. Our proposed framework consists of three major components. First, we propose an online tweet stream clustering algorithm to cluster tweets and maintain distilled statistics in a data structure called tweet cluster vector (TCV). Second, we develop a TCV-Rank summarization technique for generating online summaries and historical summaries of arbitrary time durations. Third, we design an effective topic evolution detection method, which monitors summary-based/volume-based variations to produce timelines automatically from tweet streams. Our experiments on large-scale real tweets demonstrate the efficiency and effectiveness of our framework.

EXISTING SYSTEM:

- Tweets, in their raw form, while being informative, can also be overwhelming. For instance, search for a hot topic in Twitter may yield millions of tweets, spanning weeks. Even if filtering is allowed, plowing through so many tweets for important contents would be a nightmare, not to mention the enormous amount of noise and redundancy that one might encounter.
- To make things worse, new tweets satisfying the filtering criteria may arrive continuously, at an unpredictable rate. Implementing continuous tweet stream summarization is however not an easy task, since a large number of tweets are meaningless, irrelevant and noisy in nature, due to the social nature of tweeting. Further, tweets are strongly correlated with their posted time and new tweets tend to arrive at a very fast rate.

DISADVANTAGES OF EXISTING SYSTEM:

Unfortunately, existing summarization methods cannot satisfy the above three requirements because:

(1) They mainly focus on static and small-sized data sets, and hence are not efficient and scalable for large data sets and data streams.

(2) To provide summaries of arbitrary durations, they will have to perform iterative/recursive summarization for every possible time duration, which is unacceptable.

(3) Their summary results are insensitive to time. Thus it is difficult for them to detect topic evolution.

PROPOSED SYSTEM:

- In this paper, we introduce a novel summarization framework called Sumblr (continuouS sUMmarization By stream cLusteRing).
- The framework consists of three main components, namely the Tweet Stream Clustering module, the High-level Summarization module and the Timeline Generation module.

- In the tweet stream clustering module, we design an efficient tweet stream clustering algorithm, an online algorithm allowing for effective clustering of tweets with only one pass over the data.
- The high-level summarization module supports generation of two kinds of summaries: online and historical summaries.
- The core of the timeline generation module is a topic evolution detection algorithm, which consumes online/historical summaries to produce real-time/range timelines. The algorithm monitors quantified variation during the course of stream processing.

ADVANTAGES OF PROPOSED SYSTEM:

- We design a novel data structure called TCV for stream processing, and propose the TCV-Rank algorithm for online and historical summarization.
- We propose a topic evolution detection algorithm which produces timelines by monitoring three kinds of variations.
- Extensive experiments on real Twitter data sets demonstrate the efficiency and effectiveness of our framework.

SYSTEM ARCHITECTURE:

SYSTEM SPECIFICATION

Hardware Requirements:

- System : Pentium IV 3.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 14’ Colour Monitor.
- Mouse : Optical Mouse.
- Ram : 1 GB.

Software Requirements:

- Operating system : Windows Family.
- Coding Language : J2EE (JSP,Servlet,Java Bean)
- Data Base : MY Sql Server.
- IDE : Eclipse Juno
- Web Server : Tomcat 6.0