Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 619 – 626

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Clustering Models for Data Stream Mining

R. Mythily[a,*], Aisha Banu[b], Shriram Raghunathan[c]

[a] *Assistant Professor, Department of Information Technology, B.S. Abdur Rahman University, Chennai India*
[b] *Professor, Department of Computer Science and Engineering, B.S. Abdur Rahman University, Chennai India*
[c] *Professor, Department of Computer Science and Engineering, B.S. Abdur Rahman University, Chennai India*

Abstract

The scope of this research is to aggregate news contents that exists in data streams. A data stream may have several research issues. A user may only be interested in a subset of these research issues; there could be many different research issues from multiple streams, which discuss similar topic from different perspectives. A user may be interested in a topic but do not know how to collect all feeds related to this topic. The objective is to cluster all stories in the data streams into hierarchical structure for a better serve to the readers. The work utilizes segment wise distributional clustering that show the effectiveness of data streams. To better serve the news readers, advance data organization is highly desired. Once catching a glimpse of the topic, user can browse the returned hierarchy and find other stories/feeds talking about the same topic in the internet. The dynamically changing of stories needs to use the segment wise distributional clustering algorithm to have the capability to process information incrementally.

Keywords: Data streams; information retrieval; data mining

1. Introduction

Data mining analyses a large number of observational data sets, finds unsuspected relationships and summarizes the data in novel ways that are both understandable and useful for the user. The wide-spread use of distributed information systems leads to the construction of large data collections in various fields.

* Corresponding author. Tel.: +91-965-963-3777
E-mail address: mythily@bsauniv.ac.in

doi:10.1016/j.procs.2015.02.107

Data in real world keeps changing continuously with the updates of information where the upcoming data is combined along with existing data available. The size of data keeps growing continuously with frequent updates. To deal with this data stream mining is used. Due to large volumes of data streams, it is important to construct data mining algorithms to work efficiently with huge amounts of data. Data stream is a continuous flow of information or data. Data streaming has the ability to make sure that enough data is being continuously received without any noticeable time lag.

In the case of news websites that generate frequent updates to the news readers. The goal of this work is to summarise news reports using segment wise distributional clustering to produce data clusters which is user specific. The rest of the paper is organized as follows. In the next section, the relevant work in the domain is reviewed. In section 3, the proposed clustering model is explained in detail. The experiments and results are described in section 4 while section 5 concludes the paper.

2. Related work

[1] proposed a method for managing RSS feeds from different news websites. A Web service was used to provide filtered news items extracted from RSS feeds. The result was categorized based on text categorization algorithms for efficiently managing and filtering unwanted data. An analytical model [2] was proposed to examine how RSS feeds have impact on the number of visitors, the total traffic load, and the profit of websites in a competitive setting. The explosive growth of data on web demand lead to an approach [3] for classification of RSS feed news items by considering only the key concepts of the domain for classification instead of all the terms, which curbs the problem of dimensionality. In [4] proposed that the system takes RSS feeds of news article and applies an online clustering algorithm so that articles belonging to the same news topic can be associated with the same cluster. Using the feature vector associated with the cluster, the images from news articles that form the cluster are extracted. A framework [5] for content-based web newspaper articles and to broadcast the news stories aggregation and its retrieval. In [6], emergency alert systems are proposed that demands a push notification for the infrequency of events and the urgency for notifying the parties about them. An application of e-commerce on personality searching based on RSS was proposed in [7]. In [8] uses RSS to support ubiquitous learning based on media richness theory. The proposed model visualizes the RSS as a data stream. The aim is to propose a generalized method for content aggregation and clustering.

3. Proposed Approach

The news updated every day on news sites is displayed on the web. The news stories are displayed to the user categorically. The different categories of news reports include Business, Sports, Politics, Education, and Technology. The information related to the several categories are displayed in it. The information is updated on the news sites frequently. The news reports are the events that take place on a particular day. The news reports are displayed in XML format. The overall model is given in Figure 1. Data pre-processing is done on the incomplete, noisy and inconsistent information obtained on news reports. The news reports updated on the news site providing the required information to the user are pre-processed. Incomplete data involves data lacking attributes, noisy data involves data containing errors and inconsistent data include data having discrepancies in the code.

To overcome the errors pre-processing performs cleaning, integration, transformation, reduction of news reports. This indicate filling up the missing values, aggregating reports based on relevancy and consolidating data by replacing the original information using news aggregators. Once the data is pre-processed the cleaned data is stored in data repository. Data repository contains cleaned data. The news stories are updated frequently on the web pages. When more and more of the information is gained regarding the specific event frequently. The readers are provided with the updated news stories chronologically on the news sites. The updated feeds are also stored along with the pre-processed data in the repository.