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.

The data stored in the data repository is clustered based on the specific event. (The clustering approach is described later in this section). The upcoming news reports are added to appropriate existing clusters. The various event centric clusters are then stored in database in the form of tables. Queries are used to draw information from the tables. When user chooses a feed subscribed earlier all the event specific information is retrieved from the database containing the event centric clusters. The incoming news reports from the websites are made available to the news readers to gain information. These news reports are the events recorded in real life. The news reports describe various events. The news readers when interested in updating themselves regarding a particular event would subscribe for the RSS feed. The user on returning to the home page when chooses the desired feeds to update knowledge. This is carried out by extracting the incoming news reports displayed on the news site. The extracted news reports are used to design news feeds that are stored in the data base. The news feed once stored in the database are clustered using segment-wise distributional clustering. The event specific information is listed to the user chronologically with latest news stories given the top most priority and the existing news stories with lowest priority enabling the user to update information about the specific event. If the user is not satisfied with the information provided, the user is given a chance to browse for further information on other news sites to keep track of the events.

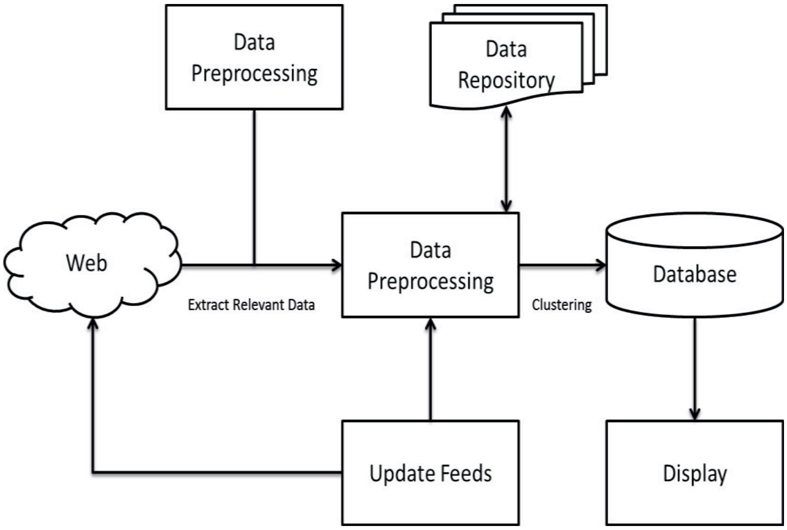

Figure 1. Proposed architecture.

Segment-wise distributional clustering process is used to evaluate the clusters based on the probabilistic distribution of data stream segments. Clustering is carried out on news reports related to a particular event to produce event-centric clusters. Event centric clusters are clusters that contain information related to the specific event. The news readers or the user when requests for desired information a search of the desired information is stored in the database. Once all information related to the requested event is obtained separately on clustering the data or information is given back to the user. The retrieved information is listed on the news sites. The Clustering system obtains its source (news reports) via the RSS feeds. The RSS feeds from the selected sources are downloaded periodically and parsed. The relevant news reports are then downloaded and pre-processed from the surrounding text and HTML code, in such a way that only the news report content remains as simple text.

- INPUT
    - Getting news reports
- OUPUT
    - Display user-specific feeds

The news readers are allowed to read news stories by logging into the news website. The new users are provided with registration form to be filled in for logging into the news site. Once the user logs in the user is provided with news stories covered from several places in real life every day. The latest news stories are displayed with the top most priority when compared to old stories.

The news stories are categorised based on the appropriate field. The top news stories which provide further information related to the event once the user selects them. This information is updated frequently using incoming news reports. The incoming news reports increase the size of information made available to the user. The news reports might contain irrelevant, incomplete and noisy data. These data are eliminated or removed using data cleaning. The irrelevant data is replaced with new data and all the relevant news stories are aggregated and are displayed as news reports to the news readers. When a user re-logins into the site next time, the user is provided with the feeds option. The feeds option contains all the desired information marked by the user as important. Once the news reports have been converted to their corresponding representation, the similarities between the different news reports are calculated. Similarly news reports are then clustered based on the specific event.

- INPUT
    - Clustering the feeds that are marked as important and saved
- OUTPUT
    - Display the clustered feeds

Clustering the news reports into event-centric clusters, the quality of the actual clusters is evaluated. Results are displayed to the user.

- INPUT:
    - Obtain user-specific information
- OUTPUT:
    - Display the relevant information

Once the information that is obtained related to the specific events using event centric clusters obtained by segment-wise distributional clustering. Segment-wise distributional clustering is nothing but the probabilistic density profiles on stream segments. Segment-wise distributional clustering is an enhancement of density distributional clustering algorithm. Density distributional algorithm evaluates the clusters based on their density. The new information is added to the existing clusters with upcoming news reports. Once all the clusters are evaluated in the database the cluster that are efficient are given back to the user as information. These clusters they are represented as the information on the news sites. The news sites will display the information to the news readers to help them keep updated regarding the specific event. This helps the users to update their knowledge frequently. Segment wise distributional algorithm to summarise news reports to produce RSS feeds which is user specific was used to cluster the feeds. Segment wise distribution clustering is an indirect representation of density distribution of the data.

Distributional clustering can be considered a generalization of the clustering problem, which clusters an indirect feature representation of the density distribution of data sets. In the case of data streams, the data sets correspond to the different segments in the data stream. This is a natural case for data streams, in which the contiguous segments represent particular temporal portions of the data stream. As opposed to standard clustering, distributional clustering provides insight into the density behaviour of data sets, rather than individual points. This is quite critical for many high volume streaming applications, in which the knowledge is embedded into groups of data points, rather than the individual data points. But, the key challenge is that to perform second level clustering of histograms with the first level clustering of individual data points done simultaneously. To overcome this in segment wise distribution clustering two sets of micro-clusters are maintained. One is based on individual data points and second level is based on higher level histogram. First level is primary micro-clusters and second level is secondary micro-clusters. Initially, the existing data sets on news sites are clustered using k-means approach. It aims at partitioning the existing data sets into k-clusters. Each data belongs to cluster with nearest mean using the approach two levels of clusters have been created. The set of data are divided into segments based on the category they belong to. Each segment contains w data sets. Each events specific segment is saved in cluster histogram having event centric clusters and re-clusters the data sets into kd clusters using k-means approach.

The upcoming new data sets are added to primary clusters created, the new data sets are added to the closest cluster that they are related to. For primary cluster created on specific event, centroid is evaluated using primary micro-cluster. Next, the upcoming data sets are taken as set of w points and determine the cluster histogram used on last window. Distance to the closest micro-cluster is evaluated using secondary cluster. Then the segment is added to existing closest micro-cluster. Finally the clustered data sets are displayed to the user.

3.1. Segment-wise distributional clustering Algorithm

Distributional clustering provides insight into the density behaviour of data sets, rather than individual points. This is quite critical for many high volume streaming applications, in which the knowledge is embedded into groups of data points, rather than the individual data points. But, the key challenge is that to perform second level clustering of histograms with the first level clustering of individual data points done simultaneously.To overcome this in segment wise distribution clustering two sets of micro-clusters are maintained. One is based on individual data points and second level is based on higher level histogram. First level is primary micro-clusters and second level is secondary micro-clusters. Initially, the existing data sets on news sites are clustered using k-means approach. It aims at partitioning the existing data sets into k-clusters. Each data belongs to cluster with nearest mean using the approach two levels of clusters have been created. The set of data are divided into segments based on the category they belong to. Each segment contains w data sets. Each events specific segment is saved in cluster histogram having event centric clusters and re-clusters the data sets into kd clusters using k-means approach. The upcoming new data sets are added to primary clusters created, the new data sets are added to the closest cluster that they are related to. For primary cluster created on specific event, centroid is evaluated using primary micro-cluster. Next, the upcoming data sets are taken as set of w points and determine the cluster histogram used on last window. Distance to the closest micro-cluster is evaluated using secondary cluster. Then the segment is added to existing closest micro-cluster. Finally the clustered data sets are displayed to the user. The pseudocode is given below

```
• Histogram window-W
• No.of.micro-clusters- K
• No of distributional clusters - Kd
• Output – Cluster Histogram H (c1....ck,W) where c1 to ck are set of micro clusters and W is the window size
• Step 1: Cluster initial Number data points using k-means approach
• Step 2 : Create Primary and secondary level micro clusters
• Step 3: Divide Stream of initial Number data points into segments
• Step 4: Each segment contains w points
• Step 5: Convert each segment into cluster histogram
• Step 6: Consider each cluster histogram as k-dimensional point
• Step 7:  Re-clustered k-dimensional points into kd different clusters using k-means approach
• Step 8: Every data point arriving are assigned to primary clusters.
• Step 9: Repeat step 4
• Step 10: Assign incoming data point to the closest micro cluster
• Step 11: Determine centroid using primary micro-cluster
• Step12: Receive set of w points
• Step13: Determine the cluster histogram used on last window (w points)
• Step14: Repeat step 6
• Step15: Assign to closest micro-cluster
• Step16: Compute distance using secondary micro-clusters
• Step17: Consider it as segment
• Step18: Assign current segment to closest micro-cluster First point
• Second point And so on
```

4. Implementation and results

A Survey was conducted to determine the efficiency of our system designed. The survey was conducted for the users of our system to determine the efficiency of our system when compared to other systems that they must have used earlier. The factors that were chosen to determine the variation in the quality of our system when compared to other systems are the number of relevant information given to the user when user chooses a given feed. Opinions were obtained from the user to determine if the knowledge gained could reach their satisfactory levels or were much below their expectation. The system was designed using segment-wise distribution clustering. As segment-wise distribution clustering algorithm is much more efficient when compared to density based clustering. Density based clustering is designed to prioritize clusters based on density. Whereas segment-wise distribution clustering is the probabilistic distribution of data on stream segments.

For the user to fill in these details the users were allowed to use the system to subscribe for the feeds they were interested in and retrieve later the information they needed with priority. In addition users were also provided with additional functionality in our system where they were displayed with information regarding the same event that they have subscribed for on the web by choosing desired option in our site. Finally users were allowed to compare the results of both the systems to rate the system efficiency. The categorisation of news stories is carried out and the numbers of relevant clusters are thus obtained this is shown in Table 1. Figure 2 shows the relevant information retrieved on specific categories of news reports.

Table 1. Categorization of news stories.

Technology-Adobe Creative Cloud	8	1	5	4	---
Business-Sebi-Sahara case	3	---	4	--	1
Health-spring allergies	6	5	6	3	---
Sports-Australia cricket	8	7	5	---	---
Sports-India in Olympic fold	9	9	7	3	---
Technology-Edu-Slide i-1017 tablet	6	7	5	---	---
Business-DTDC e-retail logistics firm	4	5	6	---	---
Heath-Bogus diet secret	6	7	---	---	---
Business-shares	10	8	5	---	1
Business-IL&FS Engg-contract	6	5	5	4	---
Sports-MSD best captain of India	9	4	4	4	---
Health-World Sleep Day	2	5	8	3	---
Health-hand-foot-mouth disease in vietnam	2	3	7	---	2
Sports-advancement of Nadal	4	8	5	2	---
Technology-Star Bolt	4	7	4	4	---
Technology-SSTL	5	5	5	1	---
Technology-Nokia Lumia 928-metal casing	7	8	6	---	---
Sports-Champions League	10	9	4	2	---

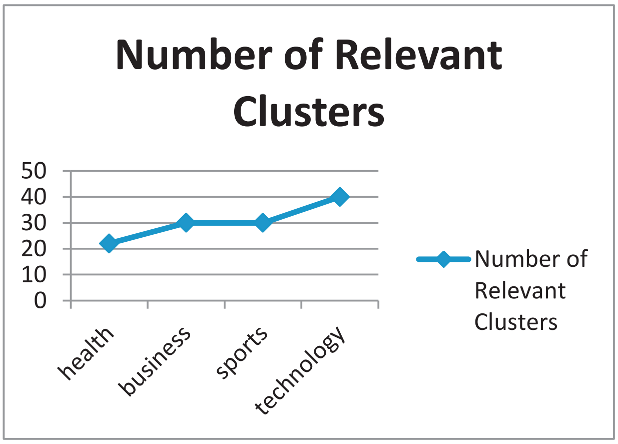

Figure 2 Relevant Clusters

The event specific information obtained by the user is related to the news reports the user subscribed for.

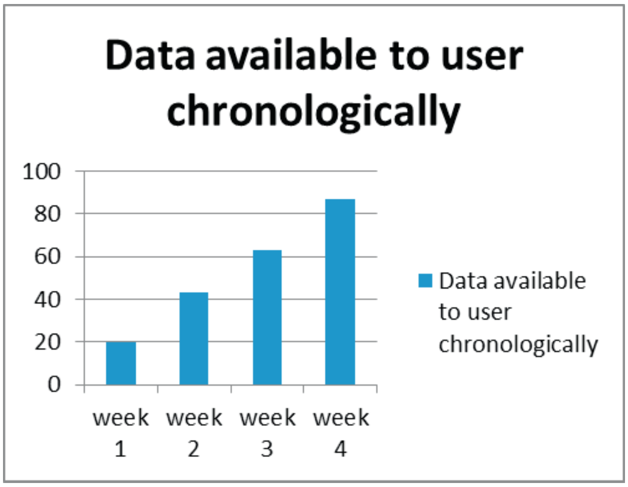

Figure 3. Data available to users.

Data is updated on the news site chronologically and the data is made available to the user. The Figure 3 depicts the data updated on the news site frequently. The information is retrieved in chronological order, with top stories given more priority and the old stories given the least priority. The performance of data updated on our system and the data updated in other systems is evaluated.

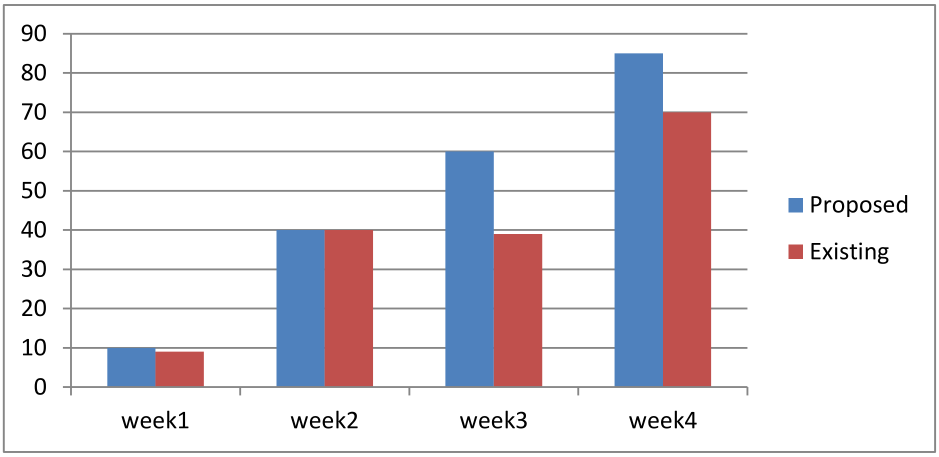

Figure 4. Efficiency of system.

The graph (Figure 4) compares our system with other online news websites. Based on the relevant information updated and retrieved in our system when compared to other system. In the first two weeks, the information that could be drawn from websites in both the systems was the same. A rapid development is obtained in the last two weeks in gathering news reports from our system when compared to other systems. The graph infers that the performance of our system is more efficient in comparison to other systems developed in retrieving information to the user.

5. Conclusion

The Concept Event Specific information retrieval system has been developed to help the users to keep track of daily reports. The clustering system reads incoming news reports as RSS streams, and clusters them according to the event they are describing. The clustering is performed on incoming news reports. The number of clusters to produce is not known beforehand and new events are detected automatically. It reduces the burden on user for exploring the contents of web sites to acquire the required information with minimal seek time. In future research the algorithm will be extended.

References

1. Saha S., Sajjanhar A., Gao S., Dew R., Zhao Y. Delivering categorized news items using RSS feeds and web services in CIT 2010 : 10th *IEEE International Conference on Computer and Information Technology Proceedings*, IEEE Computer Society, Los Alamitos, Calif., pp. 698-702, 2010.
2. Ma D. Offering RSS Feeds: Does It Help to Gain Competitive Advantage? *HICSS '09. 42nd Hawaii International Conference*, 2009.
3. Agarwal S., Singhal A., Punam B., Classification of RSS Feed News Items using Ontology *In proceedings ISDA 2012 - 12th International Conference on Intelligent System Design and Applications*, November 27-29, 2012, Kochi, pp.491- 496.
4. Sankaranarayanan J, Samet H. Images in News *20th International Conference on Pattern recognition*. pp. 3240-3243, 2010.
5. Messina A., Montagnuolo, M. Content-based RSS and broadcast news streams aggregation and retrieval *International Conference on Digital Information Management*, 2008
6. Filippo G., Ravinder S., May M.J., Gunter C. A., Shin W. Emergency Alerts as RSS Feeds with Interdomain Authorization *Proceedings of the Second International Conference on Internet Monitoring and Protection*, p.13, July 01-05, 2007
7. Jian Z., Hanshi W., Application of e-commerce personality searching based on RSS *The 2nd IEEE Int. Conf. on Information Management and Engineering*, 2010, pp. 197-199
8. Yu-Feng L., Yang-Siang S., Using RSS to support ubiquitous learning based on media richness theory *Proceedings of the 2009 IEEE international conference on Virtual Environments*, Human-Computer Interfaces and Measurement Systems, p.287-291, May 11-13, 2009, Hong Kong, China