IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 4, April 2019

Timeline Generation After Summarization of Evolutionary Tweet Streams

Madhuri Jiwe

M.E (CSE), CSMSS Chh. Shahu College of Engineering, Aurangabad (Maharashtra), India

Abstract: In a world of rapidly evolving communication systems managing intelligible and required data is a tough ask. Short-t messages in text form such as tweets are being created and shared at an unprecedented rate. Tweets, in their raw form while being informative, can also be enormous. It's a nightmare for both end-users and data analysts to plow through millions of tweets which contain huge amount of noise and redundancy. In this paper, an innovative continuous summarization framework called Sumblr to reduce the problem. Sumblr is way different than the traditional summarization methods which focuses on static and small scale data set, rather it is designed to deal with large scale, dynamic and fast arriving tweet streams.

Keywords: Continuous Summarization, Timeline, Tweet Stream, Summary

I. INTRODUCTION

There is wide increment in the short text messages due to the popularity of micro blogging services such as Pinterest, Tumblr, Twitter, Flattr and Weibo. Consider Twitter, which receives over 500 million tweets per day[1] which includes news, quotes, views, opinions and many more. If we randomly search for any hot topic we will get millions of tweets. It would be next to impossible to sort through the incoming and existing tweets for required data for and in a given time duration. One way out of this information overload is to summarize the stream of tweets. But traditional methods of summarizations do not do justice when it comes to tweet summarization. Tweet streams are relentless and fast paced, so we have to consider time and space related features of the incoming tweets. The summarization of tweets primarily should consider two main factors 1. Tweets generated at random times 2. Time classified and real time tweets. This process also takes into account gliding through the required tweet data according to the subject and also backward and forward integration of relevant data. Hence the need for continuous summarization, as we have proposed in this paper.
The challenges in this summarization are many for example the barrage of incoming tweets overwhelming data to sort out only relevant tweets and the fast paced changes. We propose to solve these issues by 1. Evolution of topic i.e. mechanism to detect sub topic changes and when they happen 2. Efficiency – creation of a very efficient summarization algorithm 3. Flexibility – tweets generated at random times should be summarized.

In this paper, we have introduced summarization framework called Sumblr (continuouS sUMarization By stream cLusteRing). [1] This innovative structure is divided into three major modules as follows 1. Tweet stream clustering – an online clustering algorithm is designed to cluster the tweets in only one pass over the data. 2. High-level summarization – in this module two kinds of summaries are generated: online summaries and historical summaries. 3. Timeline generation – the core of this module is topic evolution detection algorithm, which considers online/historical summaries to generate real-time/range timelines.

II. LITERATURE SURVEY

The process of discovering patterns in large data sets ("big data") we have data mining which is the analysis step of the KDD or "Knowledge Discovery in Databases" process[1], this is an interdisciplinary subfield of computer science[2][3][4]. The main purpose of data mining process is to extract required data from a data set and convert it to an understandable structure for further use [2]. Except the raw analysis step, it involves database and data processing, data management aspects, model and complexity considerations visualization, inference considerations, interestingness metrics, post - processing of discovered structures, and online updating [2]. The term is misleading, because the goal of data mining is the extraction of information and patterns from large amount of data, not the extraction of data itself [5].
For mining data an unsupervised data mining algorithm called BIRCH (balanced iterative reducing and clustering using hierarchies) is used on large datasets to perform hierarchical clustering[10]. The ability to cluster incoming dynamic and incremental multidimensional data in an order to produce the best quality clusters for a given set of resources (time constraint and memory) is the best advantage of BIRCH. In most cases, BIRCH requires only a single scan of the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 4, April 2019

database for generating clusters. Its inventors claim BIRCH to be the "first clustering algorithm proposed in the database area to handle 'noise' (data points that are not part of the underlying pattern) effectively" [10].
Clustering is an important area for many fields including data mining [3], statistical data analysis [8][9][10], compression [10], vector quantization, and other business applications[2]. Grouping together of the similar data is the basic clustering problem. The most general approach is to view clustering as a density estimation problem[8][5][9]. There is a hidden unobserved variable named "cluster membership". We assume that the data is arrived from a mixture model with hidden cluster identifiers. In general, mixture model M having K clusters Ci, i = 1 , 2 , … , K , assigns a probability to a data point x:
$Pr(x|M) = \sum Wi. Pr(x|Ci,M)$, where Wi are the mixture weights. The estimation of the parameters is the problem of the individual Ci, considering the number of clusters K is known. Finding parameters of the individual Ci is the main problem of clustering optimization which maximize the chance of the database given the mixture model. For general assumptions on the distributions for each of the K clusters, the EM algorithm [1] [7] is a popular technique for estimating the parameters. The assumptions made by K-Means algorithm are: 1. Spherical Gaussian distribution can effectively model each cluster, 2. Each cluster has a data item. 3) The mixture weights (Wi) are considered equal. Note that K-Means [4][6] is only defined over numeric (continuous-valued) data since the ability to compute the mean is required.

III. EXISTING SYSTEM

The proposed system consist of a new framework called Sumblr (continuous sUMarization By stream cLusteRing). It mainly consists of three main module: 1. Tweet stream clustering module 2. High level summarization 3. Timeline generation module. This can be well illustrated from the following fig:

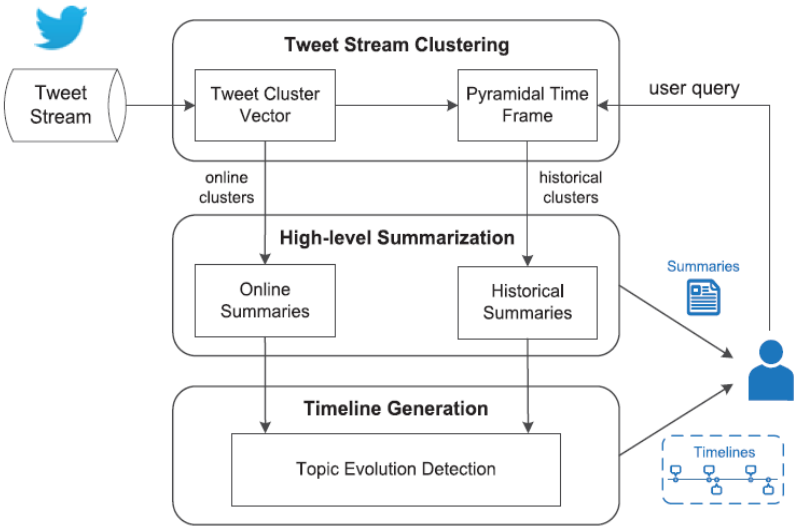


Fig1. The framework of Sumblr[1]

III.1. Tweet Stream Clustering

There are two types of data online and offline. This module maintains the online statistical data. It efficiently clusters data based on topic and maintain compact cluster information.

III.2. High-Level Summarization

There are two types of summaries in this module: online and historical summaries. What is currently discussed among the public is described in online summary and the historical summary helps people to understand the main happenings during a specific period.

III.3. Timeline Generation

The core of the timeline generation module is a topic evolution detection algorithm which produces real-time and range timelines in a similar way. The algorithm discovers sub-topic changes by monitoring quantified variations during the course of stream processing. A large variation at a particular moment implies a sub-topic change, which is a new node on the timeline. The main process is described in Algorithm 3. First the tweets are binned by time (e.g., by day) as the stream proceeds. This sequenced binning is used as input of the algorithm. Then, we loop through the bins and append new timeline nodes whenever large variations are detected (lines 4-5).

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 8, Issue 4, April 2019

IV. PROPOSED SYSTEM

The existing system consists of base paper and the enhanced features to overcome the drawback of existing system. The focus is on the problem of multi topic document clustering by leveraging natural composition of documents in text segments. A segment based approach to clustering multi topic documents such as each document may be assigned to one or more clusters. A novel clustering framework for clustering multi topic documents works as follows [11]: i) Each document in the collection is modelled with a set of segments-sets, which is identified according to the multi topics of the document. ii) Using a document clustering algorithm the segment sets from all documents are clustered. iii) The segment set clustering leads to possibly "soft " classification of the original documents. Segment clustering is done by using spherical k-means algorithm. SK-Means is majorly based on the partional clustering paradigm[12] and it is widely used for clustering documents due to its low memory and computational requirements and its ability to find high quality solutions. SK-Means introduces a similarity tolerance threshold t ∈ [0..1], along with the number k of desired clusters.

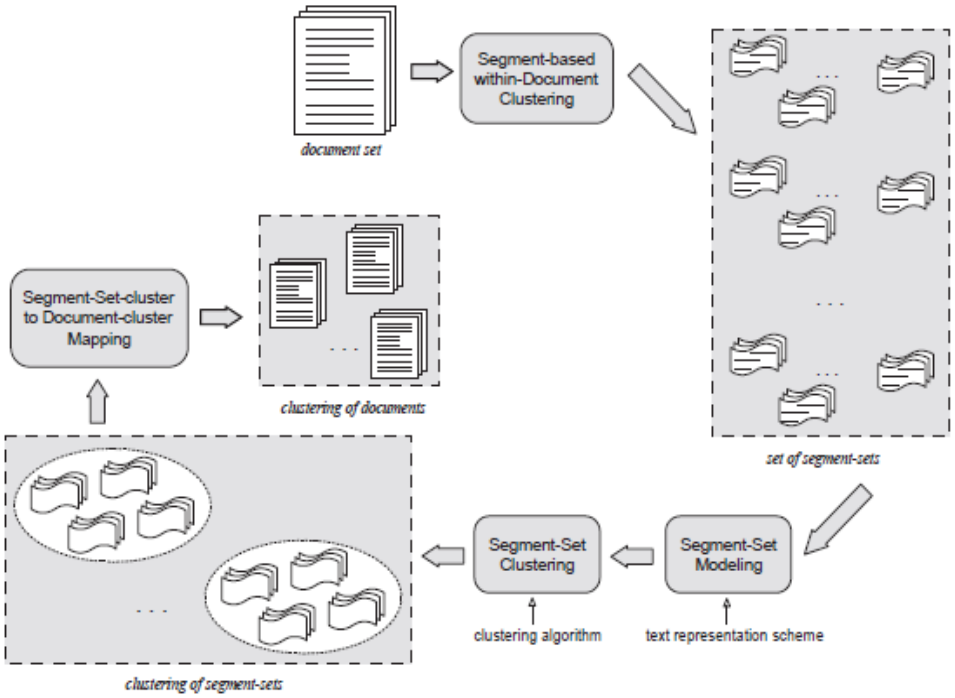


Fig2: Segment based document clustering[12]

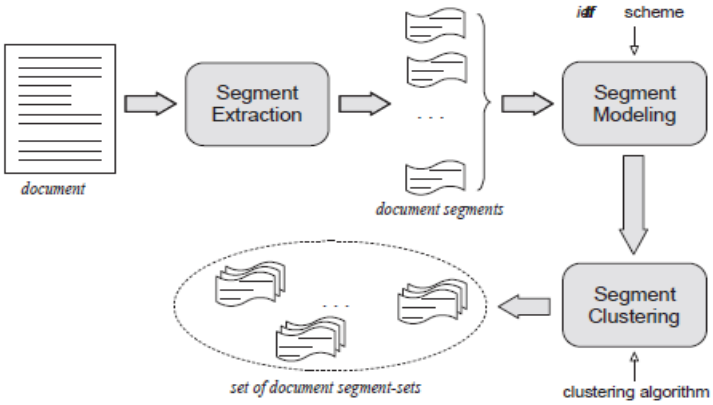


Fig3. Segment based within document clustering[12]

Sk-Means in that, for each iteration of the algorithm, the instances xi are allocated according to the following condition: $C(xi) = \{Cj \in C \mid \cos(xi, cj) \geq \max Simi \times t\}, \forall xi \in X$, where Cj is the jth cluster, cj is its centroid, and $\max Simi = \max 1 \leq j \leq k\{\cos(xi, cj)\}$. The second overlapping clustering algorithm is the spherical variant of the "fuzzy" version of k-Means, which is called Fuzzy Spherical k-Means (FSk-Means). The overlapping cluster feature is enabled by using a matrix of degrees of membership of objects w.r.t. clusters, and a real value $f > 1$. The latter is usually called "fuzzyfier", or fuzzyness coefficient, and hence it controls the "softness" of the clustering solution.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 4, April 2019

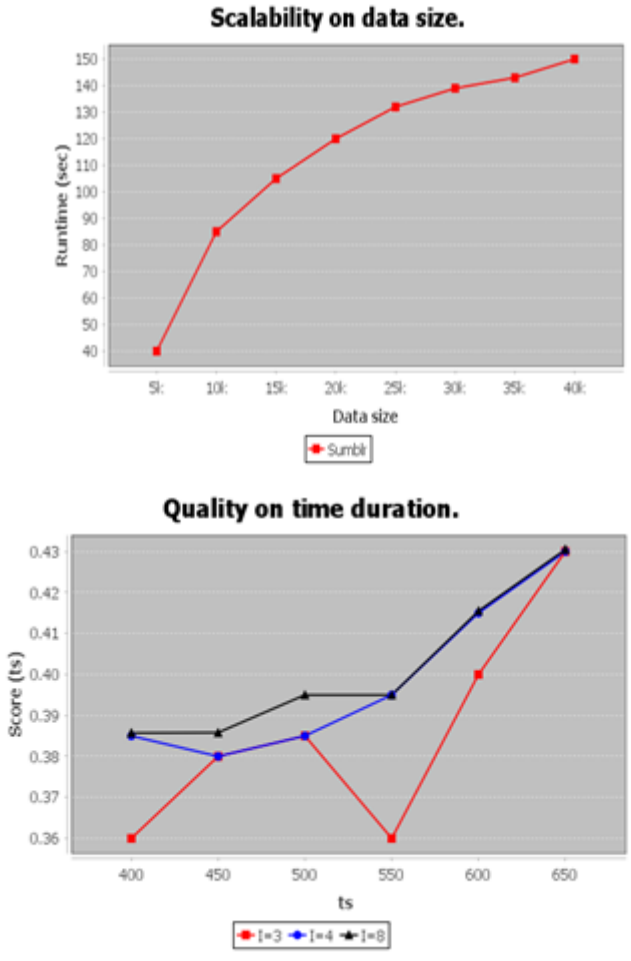
Input: A reference classification $\Gamma = \{\Gamma_1, \ldots, \Gamma_h\}$ and a clustering $\hat{C} = \{\hat{C}1, \ldots, \hat{C}k\}$, with $k > h$, for a given set of text objects.

Output: A partition $C = \{C_1, \ldots, C_1\}$ of $\hat{C}$.

Algorithm:

1: initialize C as a set of h indexed, empty sets C_i;
2: get a clone Γ' of Γ;
3: initialize a matrix $\mathcal{F}$, s. t.
$\mathcal{F}(i,j) = 2P_{ij}R_{ij}/(P_{ij} + R_{ij}), \forall \Gamma_i \in \Gamma', \forall \hat{C}j \in \hat{C}$;
4:find, $\langle i^*, j^* \rangle = \text{argmax}_{i,j}\{\mathcal{F}(i,j)\}$;
5: $\text{nSearches} \coloneqq 0$; $\text{used}_I \coloneqq \Phi$; $\text{used}_J \coloneqq \Phi$;
6: while(nSearches $<$ k)do
7: $C_{i^*} \coloneqq C_{i^*} \cup \hat{C}_{j^*}$;
8: $\Gamma'_{i^*} \coloneqq \Gamma'_{i^*} \cup \hat{C}_{j^*}$;
9: $\mathcal{F}(i,j) \coloneqq 2P_{i^*j}R_{i^*j}/(P_{i^*j} + R_{i^*j})\mathcal{F}(i^*,j), \forall \forall \hat{C}j \in \hat{C}$;
10:$\text{used}_I \coloneqq \text{used}_I \cup \{I^*\}$; $\text{used}_J \coloneqq \text{used}_J \cup \{j^*\}$
11: if (nSearches $\geq$ h) then
12: $\text{used}_I \coloneqq \Phi$;
13: find, $\langle i^*, j^* \rangle = \text{argmax}_{i,j}\{\mathcal{F}(i,j)\}, i/\in \text{used}_I, j/\in \text{used}_J$;
14: $\text{nSearches} \coloneqq \text{nSearches} + 1$;
15: return C;

V. RESULTS

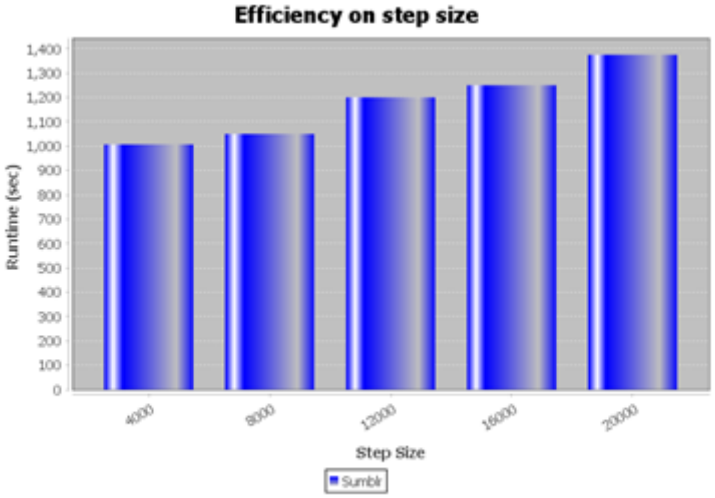


IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 4, April 2019



VI. CONCLUSION

We have studied there are various methods to summarize the tweets by forming the clusters using the clustering algorithms. It's difficult to summarize the big size tweets which are fast and are continuous by traditional summarization techniques also they can't focus on static and small scale data set. So we introduced a prototype called Sumblr which is supporting continuous tweet stream summarization. Sumblr generates summaries and timelines in the context of streams, which is suitable for distributed systems and also evaluate on more complete and large scale dataset, which deals with fast arriving tweet streams on large scale. The multi topic summarization is done by implementing spherical K-Means algorithm. Multi topic clustering is approached by SK-Means through segment based clusters. The experimental results shows that clustering multi-topic documents through a segment-set-based decomposition of the documents tends to significantly improve the identification of the various topics of each document and to favor the assignment of documents into multiple clusters according to their topics of large scale data. Summarization of the trending topics on tweets is also done here. For future work aim will be to design such an extend version to obtain the results in less time for the large scale data.

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my guide who gave me the golden opportunity to do this wonderful project on the topic Timeline generation after summarization of evolutionary tweet streams which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them.

REFERENCES

[1]. On summarization and timeline generation for evolutionary tweet stream.
[2]. A framework for clustering evolving data stream.
[3]. T. Zhang, R. Ramakrishnan, and M. Linvy," BIRCH: An Efficient Data Clustering Method For Very Large Databases."
[4]. R.C.T. Lee, clustering analysis and its applications, advances in information systems. Toum, vol.8, Plenum Press, New York,1981.
[5]. Richard Duda, and Peter E. Hart, Pattern Classification and scene analysis, wiley, 1973.
[6]. A. Jain, R. Dubes. Algorithms for clustering data, Pentice Hall, New Jersey, 1998.
[7]. C.C. Aggarwal. A framework for diagnosing changes in evolving data streams . ACM-MOD conference, 2003
[8]. R.Yan, X.Wan, J.Otterbacher, L.Kong, X. Li and Y.Zhang, "Evolutionary timeline summarization." A balanced optimization framework via iterative substitution." In Proc. 34th Int. ACM SIGIR conf. Res. Develop. Inf. Retrival, 2011.
[9]. G. Erkan and D.R. Radev, "LexRank: Graph-Based lexical centrality as salience intext summarization," J. Artif. Int.Res.Vol.22, no.1
[10]. L.O. Callaghan et al. streaming- Data Algorithms for High Quality Clustering. ICDE conf. 2002.
[11]. A Segment-based Approach To Clustering Multi-Topic Documents, Andrea Tagarelli, George Karypis.
[12]. A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice-Hall, 1988.

BIOGRAPHY

Madhuri Jiwe is currently studying in Masters of Computer Science and Engineering at CSMSS Chh. College of Engineering, Aurangabad (Maharashtra). She has completed her BE(CSE) from SPWEC, Aurangabad in 2014. Her research is clustering in data mining.