Detection and Tracking of Real-World Events from Online Social Media User Data Using Hierarchical Agglomerative Clustering Based System

AMJAD JUMAAH FRHAN
Department of Telecommunication and Information Technology
University Politehnica of Bucharest
ROMANIA.

Abstract: - Event detection and tracking is always been an efficient strategy of automation. Detecting significant real-world events from the given database or documents using past knowledge has garnered immense research interest in the recent years. Researches have garnered huge in numbers which focuses on utilizing the data like updates, status messages, shared pictures, etc. in social media to identify the occurrence of events. The most popular events of environmental, political, cultural or everyday importance are detected and tracked for various applications all over the world. However detecting the number of common events from the social media content requires efficient strategies as the size of the content and number of users is large, leading to large data to be processed. In order to avoid the limitations of the existing event detection schemes, this paper presents a new approach named Event WebClickviz. This model visualizes the user data and then analyses the similarity between the data to detect the events. Initially the event detection process is considered as a clustering problem as best results are obtained for clustering algorithms. Named Entity recognition with Topical PageRank is employed for extracting the key terms in the texts while the temporal sequences of real values are estimated to build the event sequences. The features are extracted by applying the concept of sentiment analysis using term frequency–inverse document frequency (TF-IDF). Based on these features the content is clustered using Hierarchical Agglomerative clustering algorithm. Thus the event is detected with high efficiency and they are visualized better using the proposed model. The simulation results justify the performance of the proposed Event WebClickviz.

Key-Words: - Event detection, visualization, Named Entity recognition, Topical PageRank, Hierarchical Agglomerative clustering, term frequency–inverse document frequency, online social networks, WebClickviz, clickstream

1 Introduction

Social media and online Social Networks are the most generally utilized administrations alongside web search engines these days. The Web 2.0 period got a great deal of progressive changes the way World Wide Web content is produced and used. Data produced from Web 2.0 are of immense value since they reflect parts of real-world social orders. Additionally, data are effortlessly available since they can be gathered through web-crawlers or public APIs. These two qualities constitute the primary inspiration for researchers examining online social networks. Social media has turned into the mainstream platform for people discuss recent events, which makes it a significant information source of recent events and conclusions about them [1], [2]. Nonetheless, information overload in social media makes it difficult to gain information identified with a specific event. Event detection and tracking attempt to take care of this issue by gathering and aggregating information identified with fascinating events. Event detection and tracking assumes an essential part in the security area. As a social sensor, it gives data to impact analysis, decision making and public administration [2].

Event detection and tracking isn't a trivial task due to the following reasons. Firstly, the number of tweets or messages in social media is extensive. Secondly, there are different expressions of the same meaning, and many informal expressions in social media. Moreover, the online contents in social media keep updating and new events keep appearing. Information sharing is a typical component over online social media today from which novel data investigation applications are possible. A prominent technique, known as sentiment analysis, analyses user opinions in order to extract the expressed emotion about products [3], benefits, or even political figures [4]. Marketing found an ideal platform since now organizations can investigate a vast volume of open data and distinguish patterns [5], compelling profiles [6], and specialists [7] or to give customized notices and records [8]. From another point of view, social researchers examine knowledge cascades [9], information propagation [9] or community dynamics [10]. In health care, researchers have been able to track and predict diseases like influenza [11].

From all the above applications, the assignment of event recognition emerges because of its multifaceted nature and social effect. Various event related data is shared through different social media sites like Flickr, Twitter, Tumblr, YouTube, and so forth. This huge user base makes these social media stages a portion of the biggest and quickest data sources [12]. Twitter and Facebook were utilized to spread data amid the counter government challenges in Egypt amid the Arab Spring [13], [14]. Data around an earthquake in Japan was tweeted inside 2 seconds of the quake contrasted with 20 minutes for an alarm to be issued. Social media is likewise generally utilized by open authorities for correspondence and effort to the group [15]. Broadly speaking event detection is the problem of automatically identifying significant incidents by analysing social media data. Such events can be a concert, an earthquake or a strike.

As the discussion of recent events will cause the burst of word frequency signal in social media, term weighting based techniques utilizes such burst examples to distinguish event words. These methodologies recognize words with striking scores as event words, contrasted with scores in other time windows. For instance, [29] utilizes peakiness score, which is like TF-IDF metric, and [30] utilizes drifting score, which is standardized term frequency of a n-gram with respect to other n-grams in a similar time window. In any case, these strategies do not have a powerful approach to blend words identified with similar events.

To tackle this issue, clustering based techniques group words utilizing frequency and content similarity. These strategies utilize distinctive similarity measurements. EDCoW [23] compute cross correlations by wavelets of word frequency. ET [31] utilizes both frequency and text content to process similarity. In [32] use the frequency of mention marks as a signal of words, and utilize similarity of this signal to group event words. Contrasted with other event detection and tracking techniques, clustering based strategies have better performance in general.

Hence most methodologies handle event location also to a clustering issue. Clustering can be performed on the printed highlights of users' messages (Topic Clustering) or on their spatio-transient characteristics (Spatio-Temporal clustering). A portion of the recognized bunches relate to genuine events while others are simply gatherings of comparable messages. The distinguishing proof of the event bunches is regularly handled with scoring capacities or machine learning classifiers [16]. Some methodologies use novelty tests [17] while others concentrate on sentiment peaks [18] and keyword bursts [19]. A typical component in numerous techniques is a change identification part important to distinguish that 'something occurred strange'. Change is identified through measurable examination of the messages' substance or the system's structure. The real purpose of research is that the stream of data is gone before by the data significance, interest and importance to it. The characteristics of this data provide us an opportunity to develop a model to automatically detect the abnormal/important events.

The proposed model named as Event WebClickviz is based on the previously developed WebClickviz model [20] for user behaviour analysis based on their clickstream data. In this paper, the Event WebClickviz aims at utilizing the clickstream and message data from the social media and utilizes them to detect the events automatically.

The remainder of the article is organized as: Section 2 discusses some the most related research works. The proposed methodologies are discussed in Section 3. Section 4 focuses on the Event WebClickviz performance analysis results. Finally, Section 5 makes a conclusion about the proposed model while also suggesting future directions of research.

2 Related Works

For detecting new event occurrences, previous researches usually try to find the abnormality in the collected a data. The collected data may be text, image or any multimedia data, which is made feasible by the developments of wireless multimedia sensor networks [21].

The abnormalities in such data are different going from the unpredictable thickness of significant archives, the tedious utilization of specific keywords, to the adjustments in day by day schedules of data et cetera. As the event discovery is considered as a clustering problem, various calculations have been created for that reason.

He et al [22] utilize a Discrete Fourier Transformation (DFT) technique to discover crests in the frequency space of the catchphrase motions after these were gathered in five element sorts as indicated by the power range quality length and the periodicity.

Weng and Lee [23] expand the strategy by utilizing a wavelet change, the creators infer that the transient data of the signs is lost if DFT is utilized despite the fact that it is a critical property. Events are separated from data sources by clustering keywords utilizing modularity-based graph partitioning with respect to the wavelet transformation. Both methodologies investigate the evidence of events as indicated by the unusual instability of the frequency of the keywords; however they have not considered the mix of these keywords in tweets to express the event features.

Li et al [24] acquaint an approach with build a framework for extricating events by examining tweets, which are followed from Twitter by utilizing a specific sort of channel. Authors center to distinguish Crime and Disaster related events by group tweets with Social features, for example, Twitter-Specific features and the point's features, yet just substance based features are considered.

Ciglan et al [25] proposed Wikipop, a personalized event detection system based on Wikipedia page view statistics. The Wikipop system presents to the user a set of Wikipedia articles that are popular based on his/her interest. The popularity of an article is based on the increase in page views of the article. The assumption behind their approach is events covered in public sources trigger an increase in the number of visits of corresponding Wikipedia articles.

Ahn et al [26] also uses Wikipedia page views to detect events. A set of articles whose daily page views for a fifteen day period substantially increases over the previous fifteen day period are identified. These articles are clustered using k-means and topic modelling to group similar articles together. Each detected cluster corresponds to an event.

Data from social media come in great volume and velocity. Subsequently, algorithms ought to be online and versatile in memory and computational resources. High data volume makes batch preparing computationally infeasible. In any case, in social networks, there is a rich measure of structure accessible in determining the key events in the network. For instance, an event relating to Mideast Unrest may frequently compare to content streams traded between individuals who are firmly connected to each other based on geological proximity. While the utilization of linkages in order to determine clusters and examples has been generally considered by the social networking group [33], [34], [35], [36], these strategies are ordinarily intended for static networks.

Some clustering strategies have as of late likewise been intended for dynamic networks; however they don't utilize the content of the hidden network for the mining procedure. On the contrary, some current strategies for pattern discovery in networks use both content and structure [36, 37], however these techniques are not characterized for the issue of event detection in the worldly situation. A technique in [38] is intended to measure the diffusion and spread characteristics of known and prevalent events, but is not designed for new event discovery.

In online event investigation, the incremental clustering algorithm [39] is utilized to cluster documents one by one in sequential request. A document is esteemed a member from a specific cluster if the similarity between its content and that of alternate documents in the cluster is over a predefined limit. If no cluster is sufficiently comparable to the document, another cluster is made, and the document is dealt with as the principal document of the newly created topic.

The execution of the incremental clustering algorithm on the TDT task is astounding, as long as the contents of the topic maintain a high level of similarity amid the topic's life expectancy. However, the textual content of a long term event or a multi-subject event may change after some time to reflect topic changes or distinctive parts of the event.

Allan et al. [40] noticed that transiently estimated documents likely identify with a similar event. By continually raising the similarity limit of the event detection technique for each time increase, it is conceivable to prevent transiently remote documents being grouped into a similar cluster.

In this manner, documents with comparative content that identify various events can be effectively recognized.

Yang et al. [41] connected the VSM to the task of news TDT and utilized a time window with a rotting capacity (TW-DF) to model the worldly relations amongst documents and events. In this strategy, the measure of the time window manages the quantity of past documents to be considered for clustering, and the decaying function weights the impact of a document in the window as indicated by the gap between it and the newest document.

For the online approach, Morinaga and Yamanishi [42] utilize a Gaussian Mixture Model to manage content information and a time-stamp-based marking down learning algorithm that tracks topic structures adaptively by erasing obsolete statistical data. For computational simplicity, the Gaussian Mixture Model utilized in [42] accepts that the covariance lattices of every Gaussian circulation are inclining.

For online event analysis, Surendran and Sra [43] proposed incrementally Built Aspect Models (BAMs) to powerfully find new topics from document streams. BAMs are probabilistic models intended to suit new topics with the spectral algorithm and utilize an "overlay in" approach like that of the first PLSI. This approach holds all the conditional probabilities of the old words, given the old dormant factors, and the phantom advance is utilized to measure the probabilities of new words and new documents.

The new model turns into the beginning stage for finding consequent new topics with the objective that the idle factors in the BAM model can be developed incrementally. Under this approach, the probabilities of old factors are held, and the probabilities of new dormant factors are identified as required while the streaming data is being processed. This is a magnificent method for applying incremental algorithms to online content investigation. In spite of the fact that this is another topic (or inert) detection system, it isn't an online content clustering approach; accordingly, its motivation varies from that of online event analysis. Chakrabarti et al. [33] proposed a structure of evolutionary clustering. They contended that evolutionary clustering should simultaneously streamline the clustering accuracy of snapshots and the clustering consistency along a timeline.
K-Means and agglomerative hierarchical clustering algorithms are used to satisfy their prerequisites. These greedy methodologies give users a smooth perspective of cluster changes when the data drifts from the present clusters. However, in the TDT new event detection and tracking task, it is a necessity that each news document must be gathered with documents identified with a similar real-world event. Evolutionary clustering has a tendency to keep up the consistency of clustering by sacrificing the clustering accuracy; consequently, it isn't appropriate for event detection tasks.

As most methods discussed in literature have certain drawbacks, it affects the overall performance. The limitations of these methods urge the researchers in developing more efficient technique. The proposed Event WebClickviz model is such an attempt.

3 Data Presentation & Event Detection Process

Figure 1 shows the overall procedure of the proposed model. The textual data from the social media is collected and it is pre-processed followed by the transformation into suitable format. In most cases, these dataset is transformed into discrete signals for better analysing. Given a temporal corpus of news or messages that are generated by users on a Social network system (SNS), the messages can be distributed in a range of time.
The messages are counted with a fixed time rate. The texts, either tweets or messages posted in social media, is represented as

$$txt = x, U_n, v, t \tag{1}$$

Where x is one of the set of users (U_n) who posted messages with v terms in SNS at time t.

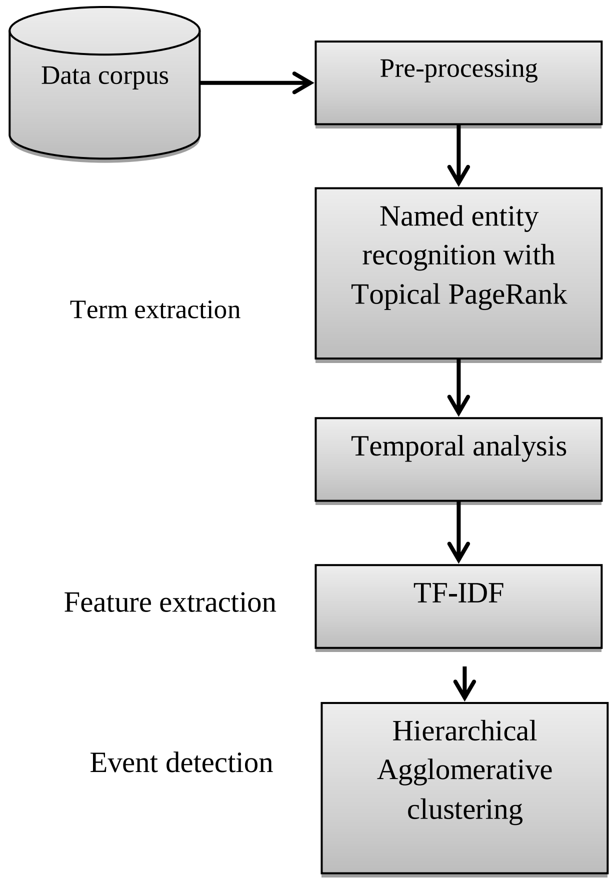

Figure.1. Proposed model Procedure

The social reactions are the reactions performed by the users on posting texts or while seeing the original texts from another user. Based on these texts, the social reactions to the text by the users is represented as

$$R(txt) = txt_i \tag{2}$$

Where txt_i is the smaller portion of the original text.
Named entity recognition with Topical PageRank
The keywords are extracted from the dataset using the Named entity recognition with topical PageRank method. Named-entity recognition is a subtask of data extraction that seeks to discover and categorize named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The model is based on the following assumptions.
There is a set of topics in dataset, each represented by a word distribution. Each user has topic interests modelled by a distribution over the topics. When a

user wants to write a text, they first choose a topic based on their topic distribution. Then chooses a bag of words one by one based on the chosen topic. However, not all words in a text are closely related to the topic of that text; some are background words commonly used in texts on different topics. Therefore, for each word in a text, the user first decides whether it is a background word or a topic word and then chooses the word from its respective word distribution. Topical PageRank was introduced by Liu et al [27] to identify keywords for future key phrase extraction. It runs topic-biased PageRank for each topic separately and boosts those words with high relevance to the corresponding topic. It is defined as

$$R_t(w_i) = \lambda \sum_{j:w_j \to w_i} \frac{e(w_j,w_i)}{O(w_j)} R_t(w_j) + (1 - \lambda)P_t(w_i) \quad (3)$$

Here $R_t(w)$ is the topic specific PageRank score of word w in topic t, $e(w_j,w_i)$ is the weight for edge $(w_j \to w_i)$, $O(w_j)$ is the summation of weight and λ is the damping factor ($\lambda = 0$ to 1). $P_t(w_i)$ is the topic specific preference value of each word. Based on this method, the keywords are extracted from the given dataset with better efficiency. A sliding window T_k is fixed to collect the data of a specific period from the total stream.

3.1 Term Frequency- Inverse Document Frequency (TF-IDF)

In order to determine the significance of the keywords in the dataset, the features such as occurrence based measurements are calculated. For this purpose the TF-IDF [28] is employed, which is a numerical statistic intended for determining the importance of a word in a text or document. TF-IDF is one of the most commonly used term weighting schemes in today's information retrieval systems. Despite its popularity, TF-IDF has often been considered an empirical method, specifically from a probabilistic point of view, with many possible variations.

Term Frequency $tf_{t,d}$ of term t in dataset d is defined as the number of times that t occurs in d. Similarly Inverse Document Frequency is the estimate of the rarity of a term in the whole data collection. If a term occurs in all the data samples of the collection, its IDF is zero. The tf-idf weight of a term is the product of its tf weight and its idf weight.

The samples have temporal characteristic, so the documents are ordered sequences in terms of time. Hence, the score for a given keyword that is computed on the sequence is a variation of its importance indexed by the sample positions. In addition to measuring the score based on tf-idf, occurrence feature of each of the Keywords and diffusion speed of original messages between users is required as well for the analysis.

The keyword score ($S_w(k,i)$) is estimated as the frequency of occurrence of that word in the text based on the context features. The major features are occurrence score and diffusion degree. The occurrence score $S_{w,occurence}(k,i)$ s given by

$$S_{w,occurence}(k,i) = \frac{|B_i^W|}{|B_i|+1} \times \log \frac{|U_{B_j}B_j|}{\left|U_{B_j^W}B_j^W\right|+1} \quad (4)$$

Here $B_i^W$and B_i are the subset of the micro-texts with term word w and all texts that are in the dataset respectively. Based on the occurrence score the frequency of occurrence can be estimated. The Diffusion degree $S_{w,D.Degree}(k,i)$ is estimated as

$$S_{w,D.Degree}(k,i) = \frac{|U_{B_i}R_i(txt_w)|}{|U.R(txt)|+1} \quad (5)$$

Where $R_i(txt_w)$ is the set of user reactions in the collected sample while $R(txt)$ is the set of user reactions in the whole corpus. Thus the features are estimated based on which the keyword score is computed. These features are also helpful in the clustering process for event detection.

3.2. Hierarchical Agglomerative Clustering

Hierarchical clustering builds a binary tree over the data. The leaves are singular data items, while the root is a solitary cluster that contains the majority of the data. Between the root and the leaves are intermediate clusters that contain subsets of the data. The fundamental thought of hierarchical clustering is to make "clusters of clusters" going upwards to build a tree.

There are two principle reasonable ways to deal with shaping such a tree. Hierarchical agglomerative clustering (HAC) begins at the base, with each data in its own particular individual cluster, and consolidates together. Divisive clustering begins

with the greater part of the data in one major group and afterward cleaves it up until the point that each datum is in its own singleton group. In this proposed event detection model, HAC is used.
For this phase, hierarchical agglomerative clustering on a semantic network of texts that is obtained in order to determine the potential clusters. The clusters include adjacent points which are close in terms of time and frequency.
Each cluster is a potential candidate depending on the existence condition. Hierarchical agglomerative clustering is the bottom-up approach of the hierarchical clustering. The main motivation behind the proposed algorithm is discovering a structure in text. The event is characterized as the abnormal texts which are different from the normal texts.
The algorithm successively grows "coherent" segments by appending lexically related paragraphs, or by merging larger segments.
The result is hierarchical structure, called dendrogram, where text segments correspond to its sub-trees.
The dendrogram represents the internal hierarchy of the text discourse, similar to an intention structure. Using sentences as the elementary segments for the algorithm makes sense for a number of reasons.
The paragraph is a universal linguistic structure, representing a coherent textual segment. Allowing a boundary in the middle of the sentence is thus counter to the author's intention.
In addition, the size of a paragraph, unlike a sentence, contains sufficient lexical information for the proximity test. The proximity test selects the closest pair of segments, based on which the events can be determined. The test is based on repetition of words, a well-recognized indicator for lexical cohesion. The test computes the cosine between the representative term vectors of the segments.

$$Proximity\ (w_i, w_i + 1) = \sum_{k=1}^{n} \frac{e(w_{k,i}).e(w_{k,i}+1)}{\|w_i\|\|w_i+1\|} \tag{6}$$

Where $e(w_{k,i})$ is the weight of word $w_{k,i}$ while $\|w_i\|$ is the vector length. Then the boundary is determined for the coherent text for identifying the abnormal occurrence of the texts. Thus the events are identified.

Algorithm: Event WebClickviz

Initialize dataset d
Extract txt
Estimate $R(txt)$
Extract key words
Estimate tf
Estimate idf
Multiply tf & idf
Compute $S_w(k,i)$
Partition the keywords to elementary segments
While more than one segment left do
Apply a proximity test to find similar segments $w_i, w_i + 1$
Merge $w_i, w_i + 1$ into one segment
End while
For each node i
Set boundary between two segments
End for
Each segment = event
End

4 Event Webclickviz Performance Analysis

In the experiments, a social media dataset from the FACEBOOK website is employed as the experimental data. There were 5999 records in the raw file from a period of August 2017 to September 2017. After data cleaning, there were 3222 records left from 243 user sessions.
There were 8 different kinds of activities during the user sessions in these data. As the primary work of WebClickviz is to visualize the data, the process begins from clickstream data visualization, followed by event detection.
The performance of the Event WebClickviz is compared with that of the WebClickviz, Pattern WebClickviz and Social pattern-WebClickviz, to estimate its efficiency.

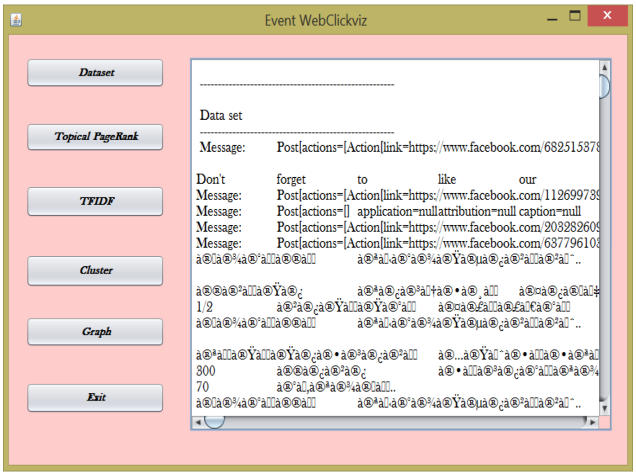

Figure.2. Dataset representation

Figure 2 shows the loading process of the data into the proposed model. The data are represented as texts initiated in the Eq. 1. Figure 3 shows the text extraction results using the Named entity recognition with topical PageRank method. It can be seen that the proposed approach effectively extracts the text from the given dataset.

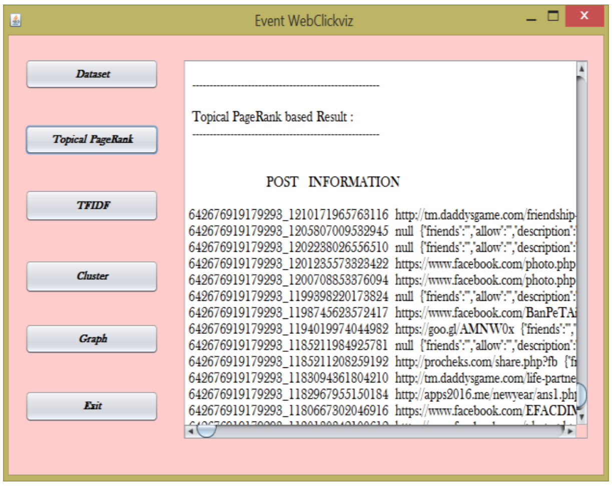

Figure.3. Text extraction

Figure 4 shows the feature extraction process carried out using TF-IDF method. This method is much efficient for the social media data because of its ability extract the features based on the statistical scores.

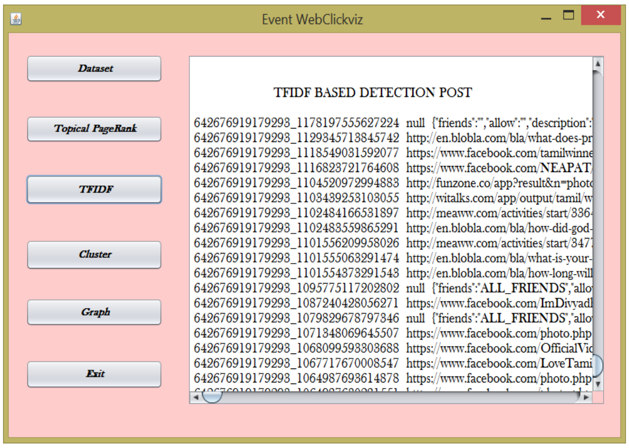

Figure.4. TF-IDF based feature detection

Figure 5 shows the clustering results. It can be seen that the proposed approach has segmented the messages and reactions and utilized them to predict the possible event occurred based on the estimated coherent features.

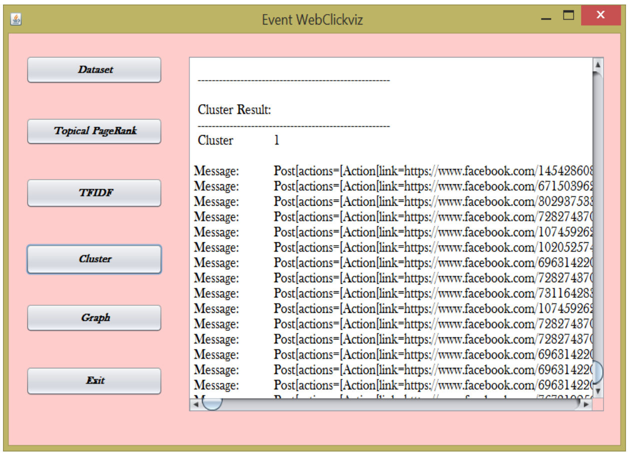

Figure.5. Clustering results

Based on the above clustering results, the events are identified. The refined list of events identified by the proposed model is shown in Figure.6. It can be evident that for the refined form shows fewer actions happened at that time period which means that this event has the highest probability of occurrence.

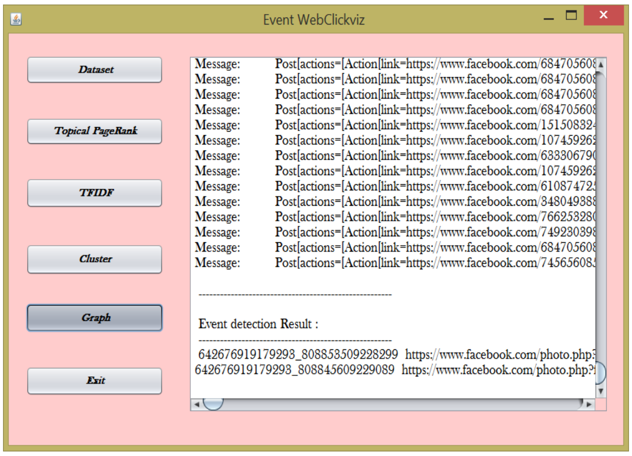

Figure.6. Refined event detection

Performance comparison

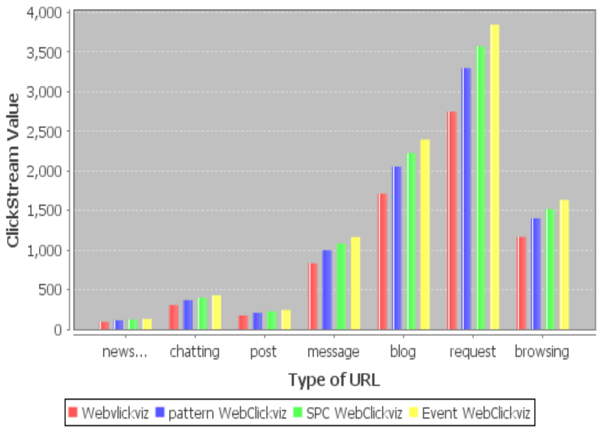

Figure.7. Clickstream value

Figure 7 shows the clickstream value comparison of the proposed Event WebClickviz with the other WebClickviz based models. It is see that the proposed model improves the detection process by identifying the larger actions in the same dataset.

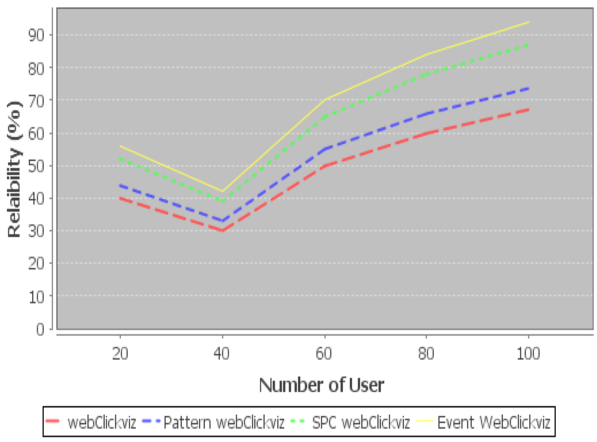

Figure.8. Reliability

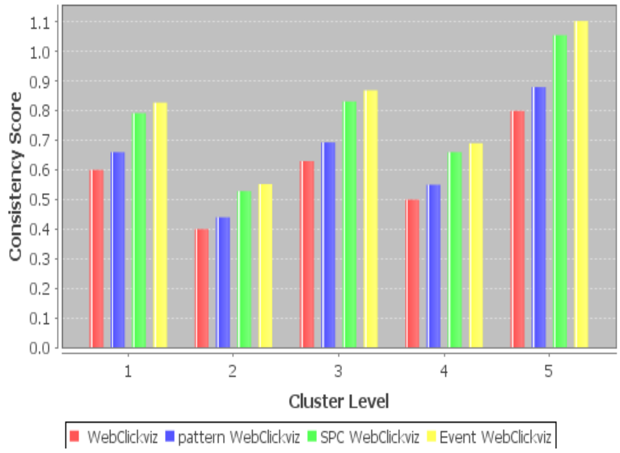

Figure.9. Consistency score

Similar to clickstream value, the reliability and consistency score comparisons are shown in Figure 8 and 9 respectively. It is found that the proposed Event WebClickviz has better values in both the cases, thus justifying its performance efficiency.

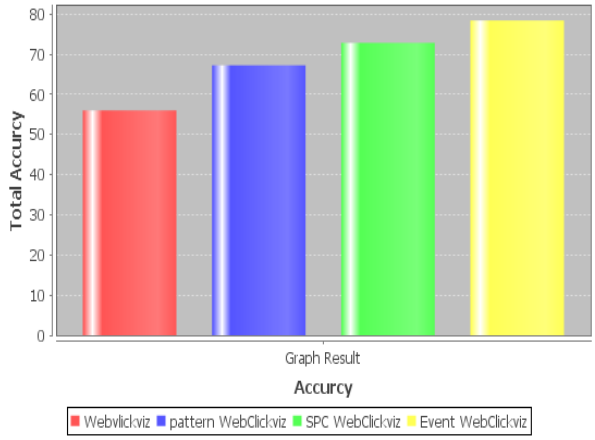

Figure.10. a) Accuracy

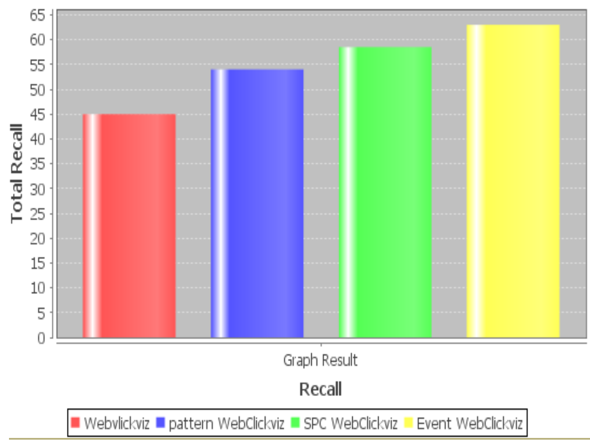

Figure.10. b) Recall

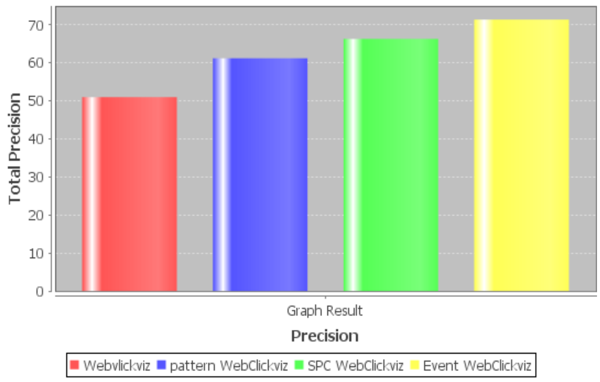

Figure.10. c) Precision

Figure 10 shows the a) accuracy, b) recall and c) Precision comparison of the Event WebClickviz with the other WebClickviz based models. As in the other performance metrics, Event WebClickviz outperforms the other models with higher values of accuracy, recall and precision. This is due to the novel process of the proposed approach in extracting the texts and features from the data corpus.

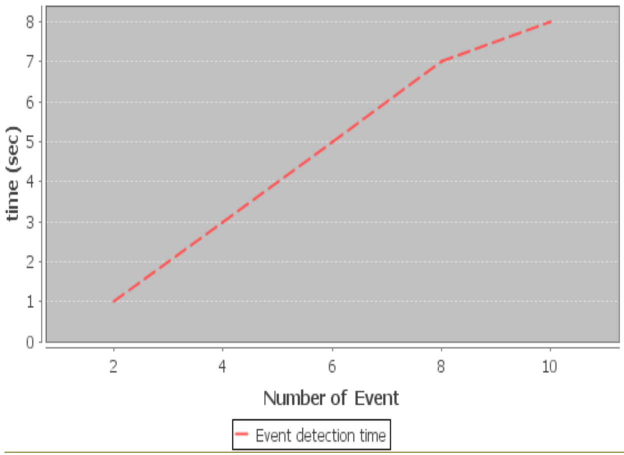

Figure.11. Event detection time

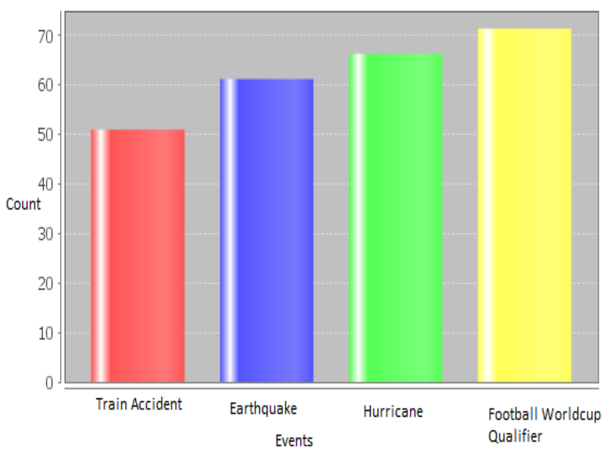

Figure.12. Major Event detection

Figure 11 shows the event detection time of the Event WebClickviz model. It is evident that the proposed model has less time for detecting the events which is in seconds. Thus the proposed approach is not only accurate but it is also very faster in performance.

During the specified duration of analysis, the major event which has higher count value is shown in Figure 12. It can be seen that from the data collected the event of football match has higher count in that particular period from August 2017 to September 2017. The main reason behind this result is that maximum number of users who posted messages during that period was fond of football matches. Thus similar to this, the behaviour of mass crowd cab also be analysed with this approach.

5 Conclusion

In this paper, an approach for detecting events from social media is presented. The proposed model which was developed on the basis of WebClickviz was named as Event WebClickviz efficiently detects the events with higher accuracy. The experiments conducted on Facebook dataset justify the claim. It is worth noting that this model has been developed not only for Facebook but suitable for text based social media like Twitter. It is also possible to utilize descriptions from images used in multimedia social networks for event detection. In the future, multiple source data will be combined to test the diverse performance of the vent detection model. Similarly instead of collecting data from social networks and utilizing it is planned to extend for collaborative collection of data and processing in parallel systems.

References:

[1] Go, A., Huang, L., & Bhayani, R. (2009). Twitter sentiment analysis. Entropy, 17, 252.

[2] Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In International conference on discovery science (pp. 1-15). Springer, Berlin, Heidelberg.

[3] Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11(538-541), 164.

[4] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, 10(1), 178-185.

[5] Mathioudakis, M., & Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In Proceedings of the 2010 ACM SIGMOD International Conference

on Management of data(pp. 1155-1158). ACM.

[6] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. Icwsm, 10(10-17), 30.

[7] Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., & Gummadi, K. (2012). Cognos: crowdsourcing search for topic experts in microblogs. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 575-590). ACM.

[8] Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., & Yu, Y. (2012). Collaborative personalized tweet recommendation. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (pp. 661-670). ACM.

[9] Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., & Kellerer, W. (2010). Outtweeting the twitterers-predicting information cascades in microblogs. WOSN, 10, 3-11.

[10] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591-600). ACM.

[11] Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PloS one, 6(5), e19467.

[12] Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., & Zha, H. (2010). Time is of the essence: improving recency ranking using twitter data. In Proceedings of the 19th international conference on World Wide Web (pp. 331-340). ACM.

[13] Starbird, K., & Palen, L. (2012). (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In Proceedings of the acm 2012 conference on computer supported cooperative work (pp. 7-16). ACM.

[14] Alkazemi, M. F., Bowe, B. J., & Blom, R. (2012). Facilitating the egyptian uprising: A case study of facebook and. Cases on Web, 2, 256.

[15] Mills, A., Chen, R., Lee, J., & Raghav Rao, H. (2009). Web 2.0 emergency applications: How useful can Twitter be for emergency response?. Journal of Information Privacy and Security, 5(3), 3-26.

[16] Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM, 11(2011), 438-441.

[17] Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 181-189). Association for Computational Linguistics.

[18] Valkanas, G., & Gunopulos, D. (2013). How the live web feels about events. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 639-648). ACM.

[19] Abdelhaq, H., Sengstock, C., & Gertz, M. (2013). Eventweet: Online localized event detection from twitter. Proceedings of the VLDB Endowment, 6(12), 1326-1329.

[20] Frhan, A. J. (2017). Website Clickstream Data Visualization Using Improved Markov Chain Modelling In Apache Flume. In *MATEC Web of Conferences* (Vol. 125, p. 04025). EDP Sciences.

[21] Al-Ariki, H. D. E., & Swamy, M. S. (2017). A survey and analysis of multipath routing protocols in wireless multimedia sensor networks. *Wireless Networks*, *23*(6), 1823-1835.

[22] He, Q., Chang, K., & Lim, E. P. (2007). Analyzing feature trajectories for event detection. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 207-214). ACM.

[23] Weng, J., & Lee, B. S. (2011). Event detection in twitter. International Conference on Weblogs and Social Media *11*, 401-408.

[24] Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012). Tedas: A twitter-based event detection and analysis system. In *Data engineering (icde), 2012 IEEE 28th international conference on* (pp. 1273-1276). IEEE.

[25] Ciglan, M., & Nørvåg, K. (2010). WikiPop: personalized event detection system based on Wikipedia page view statistics. In *Proceedings of the 19th ACM*

international conference on Information and knowledge management (pp. 1931-1932). ACM.

[26] Ahn, B. G., Van Durme, B., & Callison-Burch, C. (2011). WikiTopics: What is popular on Wikipedia and why. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages* (pp. 33-40). Association for Computational Linguistics.

[27] Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 366-376). Association for Computational Linguistics.

[28] Chowdhury, G. G. (2010). *Introduction to modern information retrieval*. Facet publishing.

[29] Shamma, D. A., Kennedy, L., & Churchill, E. F. (2011). Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 355-358). ACM.

[30] Benhardus, J., & Kalita, J. (2013). Streaming trend detection in twitter. *International Journal of Web Based Communities*, *9*(1), 122-139.

[31] Parikh, R., & Karlapalem, K. (2013). Et: events from tweets. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 613-620). ACM.

[32] Guille, A., & Favre, C. (2014). Mention-anomaly-based event detection and tracking in twitter. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 375-382). IEEE.

[33] Chakrabarti, D., Kumar, R., & Tomkins, A. (2006). Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 554-560). ACM.

[34] Chi, Y., Song, X., Zhou, D., Hino, K., & Tseng, B. L. (2007). Evolutionary spectral clustering by incorporating temporal smoothness. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 153-162). ACM.

[35] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

[36] Yang, T., Jin, R., Chi, Y., & Zhu, S. (2009). Combining link and content for community detection: a discriminative approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 927-936). ACM.

[37] Zhou, Y., Cheng, H., & Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment*, *2*(1), 718-729.

[38] Lin, C. X., Zhao, B., Mei, Q., & Han, J. (2010). PET: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 929-938). ACM.

[39] Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study final report. Proc. DARPA Broadcast News Transcription and Understanding Workshop.

[40] Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 37-45). ACM.

[41] Yang, Y., Pierce, T., & Carbonell, J. (1998). A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 28-36). ACM.

[42] Morinaga, S., & Yamanishi, K. (2004). Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 811-816). ACM.

[43] Surendran, A. C., & Sra, S. (2006). Incremental aspect models for mining document streams. *Lecture notes in computer science*, *4213*, 633.