RESEARCH ARTICLE OPEN ACCESS

A Survey on Event Detection in News Streams

Mrs. Lavanya. S, Kavipriya. R
Assistant Professor
Department Of Computer Science and Engineering
Anna University Regional Centre, Coimbatore
TamilNadu-India

ABSTRACT
The web has become the fastest growing and the most up to date source of information. Web mining is extracting knowledge from data available on the web by applying data mining techniques. News articles are being generated by various news agencies every day on the web. Articles available online is more easily accessible and most often cited by the users. Event detection involves monitoring the continuous stream of news stories to identify events which went unnoticed previously in the news stream and explore new events. In this paper, various techniques for detecting events in the broadcast Newswire stories are analyzed.
Keywords:- Web mining, Online new event detection, News streams, Topic detection and tracking (TDT).

I. INTRODUCTION

Based on the kinds of data to be mined, web Mining can be divided into web content mining, web usage mining and web structure mining. Web Content Mining refers to the extraction of relevant information from the contents of Web documents. Application of text mining to web content has been widely researched. Issues addressed in text mining are topic detection, clustering of web documents and classification of Web Pages.

Event detection problem is a part of topic detection and tracking (TDT) [6]. TDT is a DARPA funded project with UMass, CMU and Dragon systems. The domain of TDT comprises all broadcast news – written and spoken in multiple languages. It includes first story detection, segmentation and event tracking. The topic is a seminal activity or event which considers all associated events. The event is an occurrence reported at a particular time and place with consequences. It is defined by a list of stories that discussing the single event. New event refers to those stories that discusses an event which has not been reported already in previous stories. Real-time detection of the events and discovery of their evolutions should be explored to more effectively present news stories.

The query driven approach helps the user in finding the requested information. It can be applied only if the nature of events is known precisely. The new event detection does not imply any prior knowledge about the event. Therefore predefined queries does not support this task. Methods which do not solely depend on the query driven approach will be effective for event detection. Detecting events provide a conceptual structure of the news stories and facilitates better navigation for users in news spaces. Event detection can be applied to financial or stock markets, media analyses and intelligence gathering.

II. EVENT DETECTION: A LITERATURE SURVEY

A. Approaches for detecting and tracking news events

Yiming Yang, Jaime Q. Carbonell, Ralf D. Brown, Thomas Pierce, Brian. Archibald, and Xin Liu [1] proposed topic detection and tracking (TDT) to devise an intelligent system that automatically detects novel events from large volumes of news stories. This method accepts news stories from various TV channels and radio broadcasts as input. The subtasks of TDT includes segmentation of speech recognized input into news stories, detection of events from segmented news streams, tracking user interested events. Event detection task is unsupervised and is divided into two forms [8]:

1. Retrospective detection.
2. Online detection.

Retrospective detection discovers previously undetected events in news stories. It employs clustering methods like group average clustering algorithm (GAC) and incremental clustering algorithm (INCR). GAC makes use of divide and conquer strategy performing agglomerative clustering. INCR algorithm processes the documents sequentially and produces clusters incrementally. Online event detection flags onset of new events from live news feed using TF-IDF values and INCR algorithm. Event tracking task is a supervised task and uses methods like k-nearest neighbour classification, decision tree induction for tracking events of interest.

B. Event detection in social streams

Hassan Sayyadi, Matthew Hurst and Alexey Maykov [2] presented an algorithm for new event detection, which detects events by creating keyword graph and using community detection methods. Events are characterized by a set of

keywords. Keywords are extracted from the news articles which comprises named entities. The key factor is the dependency between the extracted keywords. More than one event can be denoted by the same set of terms causing ambiguity. Thus a graph is constructed using the extracted keywords called key graph. Each node in the graph represents a keyword whereas the edge represents co-occurrence of the keywords in multiple documents.

The proposed algorithm performs three tasks, namely building the key graph, community detection and document clustering. For each keyword term frequency (TF), document frequency (DF) and inverse document frequency (IDF) values are computed to determine its relevancy and association with other keywords. A node is removed if the keyword has low document frequency. An edge is removed if the keywords co-occurrence is below some threshold value. A community is characterized by a set of keywords which convey a meaningful information by establishing links between the nodes. Keywords within a community are densely linked. Inter community links can also be established by using a measure called betweenness centrality score. This task represents a hypothesis of an event. The algorithm iteratively removes the edges. Community of keywords helps to find the documents representing a particular event. Document clustering involves creating a key document which consists of keyword communities and comparing it with those documents found in original corpus. The similarity and variance between the two documents is calculated. Similar documents are clustered and documents depicting high variance with respect to the key document are filtered. The resulting document discusses an event.

C. Online new event detection

James Allan, Ron Papka and Victor Lavrenko [3] performed event detection using a clustering algorithm and threshold model. The major components of the model are the properties of an event. For event tracking, filtering methods are deployed. The event detection follows an online setting strictly, i.e., processing one news story at a time. The proposed work encompasses properties of event identity which determines whether two events are the same. A system incorporating the event identity properties, performs new event detection by comparing the newly arrived story in the stream with the existing ones. The algorithm used for new event detection is a modified version of single pass clustering. Query representation of the news story is built using techniques like feature extraction and selection. The threshold model assigns a threshold value for each query. It also compares new query against the previous queries and checks if any existing query gets triggered. Based on the result, the algorithm concludes whether the event is a new one or old.

A simple approach for tracking is using a query based approach whose capabilities is limited when events change frequently. Thus, adaptive tracking is used. It rebuilds the query after tracking a new story for a particular event. The evaluation measures for detection task are false alarm rate and misses. If a system fails in detecting new event, it is called a miss while false alarm is if the system detects a new event when actually it does not occur. System effectiveness or the efficiency can be evaluated using the above measures.

D. Detecting bursty events in the news stream

Wei CHEN, Chun CHEN, Li-Jun ZHANG, Can WANG, Jia-Jun BU [4] monitors the news stream for a predefined duration to identify bursty events. It is represented using features (i.e., keywords). Bursty event comprises bursty features whose frequency increases as the corresponding event occurs. The steps involved are identified bursty features in the current window for different periods, grouping the bursty features detected and formulating the bursty events, each being associated with a power value corresponding to its bursty level, discovering the evolution of events. Bursty features are identified using an online multi resolution burst detection (OMRBD) algorithm.

The algorithm uses sliding window with maximum of size 4 to obtain bursty features at different time periods. After identifying bursty features and its bursty period, the correlation between features are obtained using the cosine similarity measure. If the features are highly correlated, then it possibly discusses the same event. A clustering method called affinity propagation is proposed for identifying events from bursty features. It takes as input the correlation between bursty features and forms clusters for the events. The power value of each event is assigned. Events with higher values represents the significant events at a particular time. The related events for the detected bursty event are tracked along the timeline using cosine similarity based information retrieval technique. The power values of the related events determine its level of association with the bursty event.

E. Event detection using names and topics

Giridhar Kumaran and James Allan [5] perform new event detection (NED). It involves monitoring the news stream to identify stories that report on a new event. In this work, NED is treated as a binary classification problem. Each news story has three representations on the basis of named entities. Since the occurrence of new event does not follow a pattern and is almost instantaneous, named entity is used. The proposed method considers those measures, calculating the difference between the new story and existing ones. It uses baseline confidence score, named entity overlap and topic-term scores as features.

Named entities like person, location, organization, etc. are identified using BBN IdentiFinder. When two stories depict the same event, then the named entities and topic terms will be similar. Cosine similarity metric is used to calculate the

similarity between new story and those reported in the past. Confidence score is the maximum similarity value and it ranges from 0 (new event) to 1 (old event). The other features are cosine similarity between the named entities and topic terms. Support vector machine (SVM) classifier is trained on the above features. The system performance is evaluated using miss and false alarms.

F. Contextual analysis for event detection

W. Lam, H. M. L. Meng, K. L. Wong, J. C. H. Yen [6] presented a method called contextual analysis for event detection in a continuous stream of Newswire stories. The proposed method doesn't only depend on keywords for describing an event, but takes into account the concept terms, named entities like person, location, organization and story terms. The concept terms are extracted based on the relationship between the events in concept database and those sentences found in the new event. A score is computed for each concept term and those with high scores are used for representing the story.

Concept terms are useful when two stories discussing the same event uses different vocabulary. The information obtained from these terms along with its weights are used for event detection. The named entities are extracted by transformation-based part-of-speech tagger and its corresponding weights are calculated using TF-IDF weighting scheme. Similarly story terms are extracted and weighted based on its location in the story. The relationship between stories or events are determined using scores of concept relevance, named entity relevance, story feature relevance and a time adjustment scheme. Event detection model is composed of three components:

1. Similarity calculation component.
2. Grouping the relevant elements by means of agglomerative clustering.
3. Event identification.

A gross translation module translates the Chinese event into English. The representation of the event by the combined features provides better characterization of the event than the traditional keyword approach.

III. CONCLUSIONS

Event detection involves processing large volumes of news articles continuously over time. This paper analyses various techniques used for detection task. The efficiency of each technique can be evaluated using the standard evaluation measures like false alarm rates and misses.

REFERENCES

[1] Yiming Yang, Jaime Q. Carbonell, Ralf D. Brown, Thomas Pierce, Brian. Archibald, and Xin Liu "Learning approaches for detecting and tracking news events" *IEEE Intell. Syst. 14 (1999) 32-43.*

[2] Wei CHEN, Chun CHEN, Li-Jun ZHANG, Can WANG, Jia-Jun BU "Online detection of bursty events and their evolution in news streams" J. Zhejiang Univ.-Sci C 11 (2010) 340-355.

[3] Giridhar Kumaran and James Allan "Using names and topics for new event detection"

[4] Hassan Sayyadi, Matthew Hurst and Alexey Maykov "Event detection and tracking in social streams", *3rd Int'l AAAI Conference on Weblogs and Social Media, ICWSM '09, AAAI, 2009*

[5] James Allan, Ron Papka and Victor Lavrenko "Online new event detection and tracking". *In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery Special Interest Group on Information Retrieval, 1998, p 37-45.*

[6] W. Lam, * H. M. L. Meng, K. L. Wong, J. C. H. Yen "Event detection using contextual analysis". *Int. J. Intell. Syst., 16 (4): 525-546. [doi: 10.1002/int.* 1022]

[7] J. Allan et al., "Topic Detection and Tracking Pilot Study: Final Report", *Proc. DARPA Broadcast News Transcription and Understanding Workshop, Morgan Kaufmann, San Francisco, 1998, pp. 194-218.*

[8] Yang Y, Pierce T, Carbonell J. "A study on retrospective and on-line event detection". *In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, p 28-36.*