Recent Approaches for Trend Detection from Text Documents and Real Time Streams-A Study

[1]Swaraj K P , Dr. Manjula D, Adhithyan V, Hemnath K B, Manish Kumar L
Department of Computer Science and Engineering,
College of Engineering, Anna University
Chennai, TamilNadu, India

***Abstract* -Growth of unstructured data or semi-structured data in different formats is astounding and it is forecasted that the number will be touching 40 Zettabytes (ZB) by 2020.Unstructured data are basically data that does not fit easily into traditional relational type data base systems. This type of Data comprises of emails, word processing documents, messaging content, PDF files, spread sheets, multimedia, video, digital pictures and graphics, mobile phone GPS records, and social media content which combines all the other elements on a gigantic scale. Among this, unstructured data in the form of text is utmost important and it is the major content in text documents and social networking websites. As the volume and diversity of unstructured text data grows, trending topic detection and analysis have become much more important issues in order to timely utilize information from unstructured data in the form of text. Much headway has been made towards automating the process of trend detection in the recent years but still there is lot to be done to address the big data issue. This paper traces the different methods and systems that are being currently utilized for detecting trends from textual data and social streams.**

Keywords—Text Mining, Text Classification, Topic Detection, Trend Detection, ETD, Twitter

I. INTRODUCTION

Topic detection and Trend analysis have been one of the longing areas of research where expertise is short [1], [4], [5], [6], [7], [8], [9], [10]. Monitoring research trends has always been a concern of industries and policy makers of science and technology, since it helps in efficient allocation of resources and forecasting the future technology. A topic becomes trending when people began to show more and more interest in that topic for a particular time period. Considering the case of a research journal, when a topic is trending for a particular time, more and more articles are published related to that topic. Similarly when something becomes trending in social media, more and more messages tends to appear related to that event. Identifying topics and trends are given high priority by policy makers and these topics have also been termed as hot topics, upward trends or emerging trends. As depicted by their literal connotation, they seem to highly correlate with the strength of attention received over time .Many researchers have put their effort on automatic topic detection and trend analysis and have come up with various techniques and methods. But due to the problem of big data, this area requires much more number of techniques in order to find a suitable one for the job within a short amount of time.

Large amount of text are published in internet in the form of web pages, documents, research articles and other type of text. Manually classifying all this text, detecting topics, segregating them into corresponding areas and finding the emerging subtopics from each topic is a tiresome and time consuming task for an average person. Even if one seeks the help of an expert it takes a lot of man-hours to finish the job. Hence automatic topic extraction and trend detection systems are of high importance especially for those who are newly diving into research arena.

Interestingly in the recent years a lot of focus was on online social networks, in which the social network is empowered as an internet application. Some examples of such online networks are Twitter, Facebook, LinkedIn and MySpace. These social networks have rapidly grown in popularity, because they are no longer constrained by the geographical limitations. Among those, Twitter is a very prevalent online social networking and micro-blogging service that allows its users to post and share text-based messages which are commonly called as tweets. There are huge numbers of active users in twitter and daily tweets generated by them are enormous .Hence they collectively can be effectively used to provide solutions to several interesting problems such as public sentiment analysis and hot topics trend detection. Detection of popular keywords is very important particularly to find out hot issues and trends from tweets. Several methods have been proposed recently to provide new approaches for detecting trend and bursty keywords from Unstructured Text, Twitter stream data and other social networking websites. In this paper some of the existing approaches are reviewed to support the future research in the area of topic trend detection.

II. REVIEW OF EXISTING METHODS

A. A Survey of Emerging Trend Detection in Textual Data Mining

April Kontostathis et al. have published their detailed survey paper on emerging trend detection in Textual Data Mining [1]. This literature review clearly indicates that a lot of effort has been put into research in this area and hence much progress has been made towards automating the process of detecting emerging trends. All of the projects which were reviewed rely on a human domain expert to segregate the emerging trends from noise in the

system. Survey mentions about several systems that detect emerging trends in textual data. Some of the systems are semi-automatic, requiring user input to begin processing; others are fully automatic producing output from the input corpus without guidance. For each Emerging Trend Detection- ETD system, detailed descriptions are provided for components including linguistic and statistical features, learning algorithms, training and test set generation, visualization and evaluation. Literature also provides a brief overview of several commercial products with capabilities for detecting trends in textual data followed by an in industrial view point which describes the relevance of trend detection tools .Survey also delivers an overview of how such tools can be used .

An emerging trend is considered as a topic area that is growing in interest and utility over time. For example, Extensible Mark-up Language (XML) emerged as a trend in the mid-1990s.An Emerging Trend Detection, ETD application takes as input, a collection of textual data and identifies topic areas that are either novel or are exceptionally growing in importance within the corpus. Current applications in ETD have been classified into two categories, fully automatic and semi-automatic. The fully automatic systems take in a corpus and output a list of emerging topics. A human reviewer then scrutinises these topics along with the supporting evidence found by the system to determine which the truly emerging trends are. These systems often include a visual component that allows the user to track the topic in an intuitive manner. A detailed description of several semi-automatic and fully automatic ETD systems used for research purposes or educational purposes is elaborated in the survey. Some of them are mentioned below.

TOA (Technology Opportunities Analysis) - TOA is a semi-automatic trend detection system for technology opportunities analysis. Abstracts from technical database such as INSPEC, COMPENDEX, US patents are input to the system. Potential keywords from the abstracts are extracted manually by domain experts. These keywords are then combined into queries using appropriate Boolean operators to generate comprehensive and accurate searches. The queries are then input to the Technology Opportunities Analysis Knowbot (TOAK), a custom software package also referred to as TOAS Technology Opportunities Analysis System. TOAK extracts the relevant document abstracts and provides analysis of the data by using information such as word counts, date information, word co-occurrence information, citation information and publication information to track activity in a subject area. It is the responsibility of users to detect trends from this semi-automatic method.

The TimeMines system -It takes free text data with explicit date tags and develops an overview timeline of statistically significant topics covered by the corpus. TimeMines relies on Information Extraction and Natural Language Processing techniques to gather the data. The system employs hypothesis testing techniques to determine the most relevant topics in a given time frame. Only the most significant and important information as determined by the program is presented to the user. TimeMines begins processing with a default model that assumes the distribution of a feature depends only on a base rate of occurrence that does not vary with time. Each feature in a document is compared to the default model. A statistical test is used to determine if the feature being tested is significantly different than what the model would expect. If so the feature is kept for future processing, otherwise it is ignored. Finally a threshold is used to determine which topics are most important and these are displayed via the timeline interface

THEME RIVER - Theme River is yet another trend detection tool that summarizes the main topics in a corpus. Then it presents a summary of the importance of each topic via a graphical user interface. The topical changes over time are shown as a river of information. The river is made up of multiple streams. Each stream represents a topic and each topic is represented by a color and maintains its place in the river relative to other topics. Like TOA and Time Mines, Theme River does not presume to indicate which topics are emergent. The visualization is intended to provide the user with information about the corpus.

PATENT MINER - The Patent Miner system was developed to discover trends in patent data using a dynamically generated SQL query based upon selection criteria input by the user. The system is connected to an IBM DB2 database containing all granted United States (US) patents. There are two major components to the system. They are Phrase identification using sequential pattern mining and trend detection using shape queries.

HDDI- A new method is proposed called the Hierarchical Distributed Dynamic Indexing (HDDI) system .The HDDI system supports core text processing including information feature extraction, feature subset selection , unsupervised and supervised text mining machine learning as well as evaluation for many applications including ETD. They describe their approach for the detection of emerging trends in text collections based on semantically determined clusters of terms. The HDDI system is used to extract linguistic features from a repository of textual data and to generate clusters based on the semantic similarity of these features. The algorithm takes a snapshot of the statistical state of a collection at multiple points in time. The rate of change in the size of the clusters and in the frequency and association of features is used as input to a neural network that classifies topics as emerging or non-emerging.

Detailed information relating to linguistic and statistical features, training and test set generation, learning algorithms of several semi-automatic and fully automatic ETD systems has been discussed above. But disadvantage is that all of the systems rely on human domain expertise to separate emerging trends from noise in the system. Hence there is still a large scope for improvement even if much

progress has been made towards automating the process of detecting emerging trends.

B. Network Analysis of Keywords

An interesting novel network approach is proposed by Arjun Duvvuru et al. for uncovering trends in an area of research by analysing keywords appearing in research articles [2]. In this network approach, keywords are represented as nodes. A pair of keywords is linked if they appear in the same article as in Figure 1. Each link is assigned a weight which represents the number of co-occurrences of the pair in different articles. A statistical and visual analysis of the network's structural and temporal characteristics is then performed to provide a broad understanding of how keywords are organized and research areas evolved over time. Accordingly keywords organize themselves into three categories: topical keywords, complimentary keywords and diverse keywords. Networks are built corresponding to articles published for a particular time window. Networks built from articles published in two successive time windows are then compared .Through this comparative analysis it is possible to detect some interesting keyword patterns and keyword node strength. Emerging research areas are identified from this analysis. Results can be used for updating academic programs and course curricula.

The idea of this approach is based on the findings from networks analysis of various systems that most complex systems are characterized by three topological features: (1) the statistical abundance of hubs or nodes with a high number of links compared to the average degree of the network <k> and (2) the presence of scale-free node degree distributions P(k), where the probability P of finding a node of degree k follows a power-law P(k)~k-γ (3) Topological properties such as node degree and clustering coefficient can extract patterns from the information. Based on these observations, research is performed on network analysis of keywords that appear in scholarly articles. The objective is to understand the structural organization of keywords and identify trends in keywords by detecting changes in topological properties such as node strength and link weight distribution. The findings can be used in academia to identify emerging research areas.

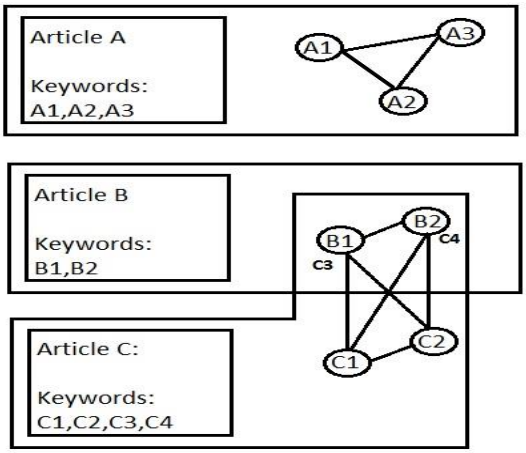

Fig.1. Network of keywords

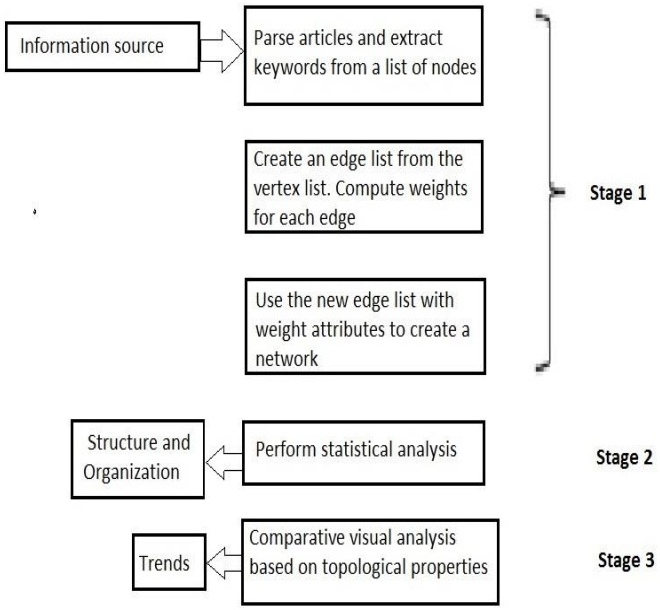

Fig.2. 3-staged method for identifying trends

Degree is the number of links or connections incident on a node and is denoted by k. The greater a node's degree, the more central it is relative to the others. Strength of a node is similar to the degree but the former takes into consideration the weights of the links incident on the node. The strength s of a node is the sum of the weights of all links incident on that node. In a weighted network, strength is considered a better measure of centrality than degree. Clustering coefficient C1of a node is the ratio of the actual number of connections between all neighbours of the node to all possible connections between the neighbours. Thus the clustering coefficient is a measure of cohesiveness of a set of nodes or a network (average clustering coefficient). The weighted clustering coefficient Cw accounts for the weighted connections. Flow diagram as in Fig.2. depicts the 3 staged method used for identifying trends. In stage 2, the weighted network created in stage 1 is subjected to statistical analysis. The primary objective of this analysis is to determine if the nature of weights on links between keywords is random or systematic. The rationale behind this is to unearth the underlying mechanism, if any, that is responsible for weights between keywords. The secondary objective of the statistical analysis is to provide a clear picture on how keywords (nodes) are organized in the journal. This method of analysis determines if the network characteristics such as degree or strength create a hierarchy of keywords. For instance, the organization of keywords by topical areas can be ascertained from this analysis. Additionally, keywords can be classified into different categories based on how they connect to each other. Finally in stage 3, visual maps are created for the networks from different time periods. These maps are color coded to show variations in different network characteristics such as strength and weights. This aids in picturing different characteristics and patterns in the network, which in turn facilitates the comparative analysis through which trends in research areas are detected

C. Detecting Temporal Trends of Phrases in Research Documents

Conventional methods take into consideration only frequency of word and do not consider the nature of terms or the importance indices separately. In this paper, Hidenao abe et al. proposed an improved method for detecting trends of phrases by combining term extraction methods, importance indices of the terms and trend analysis methods [3]. Although importance indices of the technical terms play a key role in finding valuable patterns from various documents, temporal changes of them are not explicitly treated by conventional methods. Since those methods depend on particular index in each method, they are not robust in changes of terms. In order to robustly detect remarkable temporal trends of technical terms in given textual datasets, a method is proposed based on temporal changes in several importance indices by assuming the importance indices of the terms to be a dataset. After detecting the temporal trends, comparison is done between the emergent trends of the technical phrases to some emergent phrases given by a domain expert for checking match.

The main steps in this approach are

1) Technical term extraction in a corpus
2) Importance indices calculation
3) Trend detection

An overview of the proposed method is illustrated below

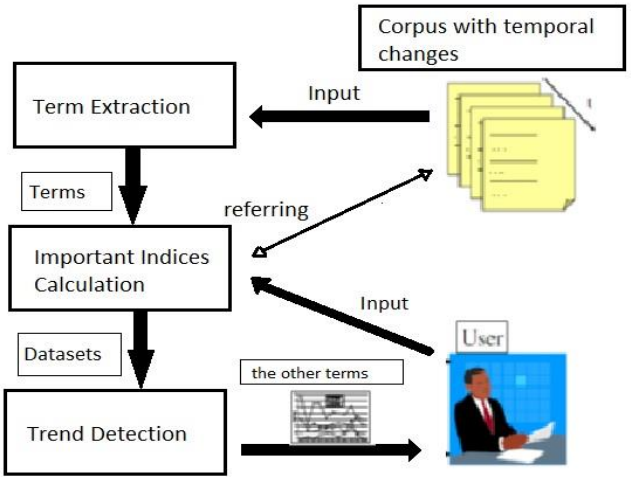

Fig.3. Architecture of trend detection from corpus

Term extraction method is based on the adjacent frequency of compound nouns CN. This method involves the detection of technical terms by using the following values for each candidate CN:

$$FLR(CN) = f(CN) \times (\prod_{i=1}^{L} (FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{L}}$$

Where f(CN)means frequency of the candidates CN, and FL(Ni) and FR(Ni)indicate the frequencies of the right and the left of each noun Ni. After determining terms in the given corpus, the system calculates multiple importance indices of the terms for the documents of each period. Some well-known indices are TF-IDF and Jaccard's coefficient.

In the following experiments, the threshold for FLR is set up as FLR >1.0 to extract terms from the whole set of documents. Then, the linear regression analysis technique is applied in order to detect the degree of existing trends based on the two importance indices.

Degree of term is calculated as follows

$$Deg(t) = \frac{\sum_{i=1}^{M}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{M}(x_i - \bar{x})^2}$$

Where $\bar{x}$ is the average of the M time points and $\bar{y}$ is the average of each importance index for the period. Simultaneously we calculate the intercept Int(t) of each term t as follows

$$Int(t) = \bar{y} - Deg(t)\,\bar{x}$$

Results for automatically extracted terms are calculated by using the degree and the intercept of each term. The following two trends are determined

• Emerging
– sorting the degree in ascending order
– sorting the intercept in descending order
• Subsiding
– sorting the degree in descending order
– sorting the intercept in ascending order

D. Sensing Trending Topics in Twitter

Rich and timely information about real-world events of all kinds is generated by Online social and news media. However, due to the huge breadth of user base and the large amount of data available it requires a substantial effort of information filtering to successfully drill down to relevant topics and events. Trending topic detection is therefore a fundamental building block to monitor and summarize information originating from social sources. The paper presented by Luca Maria Aiello et al. compares six topic detection methods on three Twitter datasets related to major events [11]. One of the novel topic detection methods proposed based on n-grams co-occurrence and df-idft topic ranking consistently achieves the best performance compared to the conventional state of the art techniques. The methods tested in this paper cover three different classes: probabilistic models (Latent Drichlet Allocation), classical Topic Detection and Tracking (a common document-pivot approach) and feature-pivot methods. Along this series of methods, four novel approaches are developed; including methods that use the concept of frequent item set mining. In particular, it is showed that a method that leverages n-gram co-occurrences (instead of uni-grams) and df-idft topic ranking is consistently the best per-forming method among the ones tested. The proposed df-idft is a score for burstiness detection that can significantly assist in determining the most rapidly emerging topics. The 3 selected datasets cover the domains of politics (the US Super Tuesday primaries of

March 2012 and the US Presidential elections of November 2012) and sports (the English FA Cup Final).

The significant contributions can be summarized as follows:
• A comparative study of a wide range of topic detection methods across three large Twitter datasets on a real-world event sensing scenario is presented. The main idea of using different datasets is to compare the performance of the algorithms in different domains which have their own special features.
• Analysis is done on how factors such as the type of input data (e.g. time span, topic breadth) and pre-processing techniques can affect the quality of topic detection results.
• Among the tested methods, the proposed novel algorithm combining n-grams with df-idft ranking performs best, outperforming other state-of-the-art techniques.

The six methods compared are
1) Latent Dirichlet Allocation (LDA)
2) Document-Pivot Topic Detection (Doc-p)
3) Graph-Based Feature-Pivot Topic Detection (GFeat-p)
4) Frequent Pattern Mining (FPM)
5) Soft Frequent Pattern Mining (SFPM) and
6) BNgram

Using n-grams makes particular sense for Twitter, since a large number of status updates are just copies or retweets of previous messages, so important n-grams tend to be frequent. A new feature selection method is used in which changing frequency of terms is taken into account over time as a useful source of information to detect emerging topics. The main goal of this approach is to find emerging topics in post streams by comparing the term frequencies from the current time slot with those of preceding time slots. A new metric is proposed which introduces time to the classic score. Historical data is used to penalize those topics that began in the past and are still popular in the present, and that therefore do not define new topics. This approach indexes all keywords from the posts of the collection. The keyword indices, implemented using Lucene are organized into different time slots. In addition to single keywords, the index also considers bigrams and trigrams. Once the index is created, the df-idft score is computed for each n-gram of the current time slot i based on its document frequency for this time slot and penalized by the logarithm of the average of its document frequencies in the previous t time slots.

$$df - idf_t = \frac{df_i + 1}{\log\left(\frac{\sum_{j=i}^{t} df_{i-j}}{t} + 1\right) + 1}$$

In addition, a factor is considered to raise the importance of proper nouns (persons, locations and organizations, using a standard named entity recognizer, as they are essential keywords in most discussed stories. As a result of this process, a ranking of n-grams is created based on their df-idft scores. A single n-gram is often not very informative, but a group of them often offers interesting details of a story. Therefore, a clustering algorithm is used to group the most representative n-grams into clusters, each representing a single topic. The clustering is based on distances between n-grams or clusters of n-grams. From the set of distances, those not exceeding a distance threshold are assumed to represent the same topic. Similarity between two –grams is defined as the fraction of posts that contain both of them. Initially every n-gram is assigned to its own singleton cluster and then follow a standard "group average" hierarchical clustering algorithm to iteratively find and merge the closest pair of clusters. When an n-gram cluster is joined to another, the similarities of the new cluster to the other clusters are computed as the average of the similarities of the combined clusters. The clustering is repeated until the similarity between the nearest un-merged clusters falls below a fixed threshold, producing the final set of topic clusters for the corresponding time slot. In the experiments, a similarity threshold of .5 is used, which means that two n-grams must appear in more than 50%of the same tweets in order to belong to the same topic. This assumption is stronger in this case because only the post for a specific time slot is considered. So it is more likely that the n-gram clusters whose similarities are higher than the threshold represent the same topic. Finally, the clusters are ranked according to the highest score of the n-grams contained in the cluster. This ranking criterion is based on the assumption that each cluster score should be associated with the score of the most representative n-gram in the cluster, as the cluster is mainly composed of posts containing it.

E. Early detection of twitter trends-A Non parametric approach

Trend detection in twitter is done mostly based on time series analysis. In the parametric approach, a jumpiness parameter say 'p' is estimated from a window of activity and trend detection is based on whether p exceeds a threshold. Here a model is created for the type of activity say a pattern that can occur before a topic is declared as trending.Ths type of method is called parametric because it estimates a parameter 'p' from data. But such a model often fails to capture all the type of patterns that can precede before the topic becomes trending. Sometimes the preceding pattern can be a gradual rise, sometimes constant and a big jump and it may also be a series of small jumps which leads to a big jump. So in these cases different models are to be built for each expected trend causing pattern which is an infeasible solution.

A Novel efficient non parametric data driven method is proposed by Snikolov and Shah for the early detection of twitter trends [12].This method is based on the assumption that people acting in social network are predictable. Usually in a social networking site, if a person 'x' has friends which are all friends of a person 'y' then there is a chance that 'y' is friend of 'x'. If many of a person x's friends are talking about some event then there

is high chance that the person 'x' also may talk about the same. So here it is assumed that there can be only a few frequent patterns that can occur before a topic becomes trending. In this method, data defines the model .Data is learned about the patterns that precede trends and patterns that don't. Then instead of creating a model from data, data itself is used to determine whether it is going to be a trend or not.

First of all the activity of a new topic is tracked. An observed pattern 's' is found out for a time period and it is compared with example activity patterns 'r' from topics that became trending in the past and which didn't. This means r can be set R- and set R+ consisting of say 500 trending and non-trending patterns each. The comparison is based on a voting scheme. Each of the example activity patterns 'r' vote for the observed pattern 's' based on a voting function. Weight of each vote depends upon the Euclidian distance between r and s. Finally all the trending and non-trending votes are summed up and ratios of these are calculated. If the ratio is greater than '1' the topic is trending and if it is less than '1', it's not.

The advantages of this method are

a) Computations are simple since only distance needs to be calculated
b) Distance Computations can be parallelised
c) No model is used for detection. This algorithm was able to detect Twitter trends ahead of Trending algorithm provided by twitter almost 79% of the time with an error rate of 95% (95 % true positive rate and 4% false positive rate).

F. Trend detection from Facebook

At present Facebook is the largest online social networking site with more than 800 million active users followed by twitter with an estimated 280 million users. Lot of contents are shared online in different formats. Real time news, opinions and statuses are shared online through Facebook. Public nature of these status messages from Facebook is used in the technique proposed by Irena Pletikosa Cvijikj et al. for trend detection from Facebook which is claimed to be the first of its kind [13]. With the help of a novel algorithm they categorise trending topics into a) disruptive events b) popular topics and c) daily routines. Finally the characteristics of the proposed categories are analysed and compared in terms of distribution and information diffusion in order to increase the understanding of emerging trends on Facebook.

Facebook GRAPH API provides access to the Facebook social graph via a uniform representation of objects in the graph (e.g., people, posts, pages, etc.). In this study the Post objects were of interest. Each Post object contains the following information: (1) content details for the post (message, name, caption, description), (2) Facebook user who posted the message, (3)Type of the message as defined by Facebook, i.e. status, photo, link, etc., (4) Time of creation, (5) Application used for sharing the post, etc. All of these elements were stored in a relational database for further investigation. Linguistic analysis was done on extracted 3 sets of posts. The average number of sentences in a post is approximately 1.4. At the same time, the average number of words is approximately 18, a bit higher compared to the 16 words in a tweet. However, looking at the full dataset, while the average post length in character didn't significantly differ from our results (108), the maximum length was found to be 754 characters, which on average corresponds to approximately 10 sentences and 122 words. Trend detection is commonly based on (1) topic identification, and (2) cluster detection.TF-IDF formula is not applicable for the content shared on Facebook because of the limited length of the Facebook posts, which would reduce the value of the term frequency component in the equation. Hence the concept of a hybrid document was used in this method. The notion of a hybrid document represents a collection of posts P = {p1, p2, …, pK}, obtained within a timeframe T, which corresponds to the interval for fetching posts from Facebook in a near real-time system. Each time frame T represents a separate dataset described by a separate weighted list.TF-IDF calculation is modified to suit for this type of hybrid document approach.

Post topic identification results in an ordered list of the most significant terms in the corpus. The next step is to cluster together terms that belong to the same topic. Post clustering is performed in two steps (1) clustering by distribution, and (2) clustering by co-occurrence. Clustering by distribution is a combination of the comparison of the term weight and the intersection of the related documents. The goal is to eliminate the multiple occurrences of the similar n-grams with different lengths belonging to the same posts. Once the grouping is done, we replace the groups with the n-gram with the maximal length since it contains maximum information regarding the topic. Clustering by co-occurrence is based on the assumption that terms that appear frequently in similar posts belong to the same topic. This step is used to further group the terms that are not semantically similar and belong to different posts, but still refer to the same topic.

Evaluation was done based on precision and recall. 10 experiments were conducted, each collecting and processing 1000 posts from different time intervals and review of the results were obtained. Approach tried to identify differences in terms of distribution through the shape and volume of the shared information. Approach also used measuring the speed and scale of information distribution on Facebook as an indicator of the possibility to use Facebook as a news media. The distribution in terms of the volume of the posts shared on Facebook regarding a certain topic is a clear indicator of a level of interest of users for the related topic. In addition, the shape of the distribution is an indicator of a topic belonging to the category of 'daily routines' that is always present in the conversation at some relatively equalized level, or if it relates to an event occurring at a particular point in time. Understanding the differences between distributions that relate to the 'daily routines' and 'popular topics' on one side, and the distributions related to 'disruptive events' on the other, gives us the possibility to train the systems for automatic trend detection in order to distinguish between these different types of trends.

Comparison to the cumulative distribution of posts over time of day shows that the peaks on the 'popular topics' correspond with the peaks on the daily post distribution graph, while peaks for 'daily routine' are the opposite. On a more general level, daily routines are usually related to a certain period of time in a day, for example, "good night" appears as a trending topic only in the evenings. Furthermore, these two topics are present and trending during the whole time interval, indicating a popular topic, but not something new.

III. CONCLUSION AND FUTURE WORK

Trend detection techniques are regularly performed on static text data and real time stream data. This paper has presented a study on recent methods for trend detection in text data and real time streams. First an overview of an existing survey was presented providing a larger glimpse of existing ETD methods. Two recent methods for detecting trend from unstructured text are described in detail. The network method creates a keyword network to detect trends whereas the other method calculates trends based on important indices. 3 methods for trend detection from social media data were also discussed. These methods detect trends of live streaming data from online social networks such as Twitter and Facebook. Here the techniques differ from that of static text because mostly trends are based on events in social networks. The concepts provided by different methods were concisely described along with the relevancy of each method. The important contribution is that the survey focussed on enlisting existing methods used for trend detection under the categories of both static documents and real time streams. This paper also provides an efficient and quick reference for those who are interested to do research in the field of trend detection in texts. Future work is to present a complete study for ETD in real time streaming data and proposing an integrated framework.

REFERENCES

[1] April Kontostathis,Leon M Galitsky,William M Pottenger, SomaRoy, Daniel J Phelps, "A survey of emerging trend detection in textual Data Mining "

[2] Arjun Duvvuru , Sagar Kamarthi , Sivarit Sultornsanee, "Under covering research trends: Network analysis of keywords in scholarly articles ." Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE) 2012.

[3] Hidenao Abe,Shusaku Tsumoto "Detecting Temporal Patterns of Technical Phrases by using Importance Indices in a Research Documents" ,Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA - October 2009.

[4] Masahiro Terachi, Ryosuke Saga and Hiroshi Tsuji, "Trends Recognition in Journal Papers by Text Mining ", IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan

[5] Daehoon Kim, Daeyong Kim, Seungmin Rho and Eenjun Hwang, "Detecting Trend and Bursty Keywords Using Characteristics of Twitter Stream Data " International Journal of Smart Home Vol. 7, No. 1, January, 2013

[6] Levent Bolelli,Şeyda Ertekin,C. Lee Giles "Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation" , Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval Springer-Verlag Berlin, Heidelberg ,2009 pp 776 - 780

[7] Xiaodong Wang , Juan Wang, "A Method of Hot Topic Detection in Blogs Using N-gram Model " JOURNAL OF SOFTWARE, VOL. 8, NO. 1, JANUARY 2013

[8] Teng-Kaifan ,Chia-Huichang , "Exploring Evolutionary Technical Trends From Academic Research Papers " Journal of information science and engineering 26, 2010, pp 97-117

[9] Manish Gupta ,Jing Gao,ChengXiang Zhai,Jiawei Han "Predicting Future Popularity Trend of Events in Microblogging Platforms" ,Proc of ASIST, , Baltimore, MD, USA. October 28-31, 2012

[10] Aleksandre Lobzhanidze, Wenjun Zeng, Paige Gentry and Angelique Taylor, "Mainstream Media vs. Social Media for Trending Topic Prediction – An Experimental Study " The 10th Annual IEEE CCNC

[11] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba,Ayse Göker, Ioannis Kompatsiaris, Senior Member, IEEE, and Alejandro Jaimes, "Sensing Trending Topics in Twitter." IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 6, OCTOBER 2013

[12] http://snikolov.wordpress.com/2012/11/14/early-detection-of-twitter-trend

[13] Irena Pletikosa Cvijikj, Florian Michahelles , "Monitoring Trends on Facebook " Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011