Volume 2, Issue 5, May 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey of Topic Tracking Techniques

Kamaldeep Kaur
UIET, Panjab university, Chandigarh
Vishal Gupta
UIET, Panjab University, Chandigarh

***Abstract*— Text mining is a field that automatically extracts previously unknown and useful information from unstructured textual data. It has strong connections with natural language processing. NLP has produced technologies that teach computers natural language so that they may analyze, understand and even generate text. Topic tracking is one of the technologies that has been developed and can be used in the text mining process. The main purpose of topic tracking is to identify and follow events presented in multiple news sources, including newswires, radio and TV broadcasts. It collects dispersed information together and makes it easy for user to get a general understanding. In this paper, a survey of recent topic tracking techniques is presented.**

***Keywords*— Text Mining, Topic detection, topic tracking**

I. INTRODUCTION

Text mining is a new area of computer science which fosters strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. It seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. The techniques employed usually do not involve deep linguistic analysis or parsing, but rely on simple 'bag-of-words' text representations based on vector space. Several approaches exist for the identification of patterns including dimensionality reduction, automated classification and clustering [1]. The field of text mining has received a lot of attention due to the always increasing need for managing the information that resides in the vast amount of available documents [2]. The goal is to discover unknown information, something that no one yet knows.

[3] The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle.

[2]A typical text mining process can be shown as:

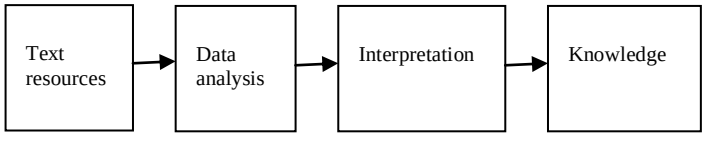

Fig. 1 Typical Text Mining Process

Taking a collection of text resources, a text mining tool would proceed with the data analysis. During this analysis many sub processes could take place such as parsing, pattern recognition, syntactic and semantic analysis, clustering, tokenization and application of various other algorithms.

II. TOPIC TRACKING

A topic tracking system works by keeping user profiles and based on the documents the user views, predicts other documents of interest to the user [4].

The task of topic tracking is to monitor a stream of news stories and find out what discuss the same topic described by a few positive samples [12]. The main purpose is to identify and follow events presented in multiple news sources, including newswires, radio and TV broadcasts [8]. With the fast development of internet, topic related information is often isolated and scattered in different time periods and places. TDT techniques are used to organize news pages from a lot of news websites into topics [9]. It collects dispersed information together and makes it easy for user to get a general understanding [13].

[4] There are many areas where topic tracking can be applied in industry. It can be used to alert companies anytime a competitor is in the news. This allows them to keep up with competitive products or changes in the market. Similarly businesses might want to track news on their own company and products. It could also be used in the medical industry by doctors and other people looking for new treatments. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest.

A typical topic tracking system can be illustrated as: [11]

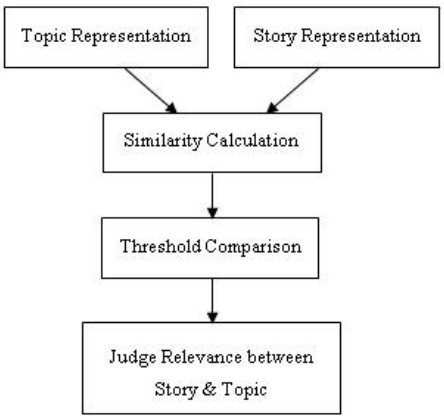

Fig. 2 Architecture of a topic tracking system

When performing topic tracking, the topic tracker needs to represent topic/ story by some inner format, select similarity function for measuring topic-story similarity, and finally do threshold comparison. If the similarity ishigher than predefined threshold, then the story is judged to be on-topic; otherwise, off-topic.

A. Background

[14][29][30]Topic Detection and tracking is fairly a new area of research in IR, developed over the past 8 years. Began during 1996 and 1997 with a Pilot study conducted to explore various approaches and establish performance baseline. The research began in 1996 with DARPA funded pilot study.
Quite soon the traditional methods were found more or less inadequate for online detection purposes. In recent years, TDT techniques have been developed to identify the issues discussed in a large collection text.
Very brief descriptions of the existing TDT tasks are given:

- TDT-1: deals with the three major tasks: (1) segmenting a stream of data, especially recognized speech, into distinct stories; (2) identifying those news stories that are the first to discuss a new even occurring in the news; (3) given a small sample news stories about the event. The TDT corpus includes approximately 16,000 stories about half collected from Reuters newswire and half from CNN broadcast news transcripts during the period July 1, 1994 to June 30, 1995. An integral and key part of the corpus is the annotation in terms of news events discussed in the stories. Twenty-five events were defined that span a variety of event types and that cover a subset of the events discussed in the corpus stories. Annotation data for these events are included in the corpus and provide a basis for training TDT systems.
- TDT-2: they ran tracking experiments on just the Mandarin stories in the development corpus. The TDT2 English Corpus has been designed to include six months of material drawn on a daily basis from six English news sources. The period of time covered is from January 4 to June 30, 1998. The six sources are the New York Times News Service, the Associated Press World stream News Service, CNN "Headline News", ABC "World News Tonight", Public Radio International's "The World", and the Voice of America.
- TDT-3: the tracking part of the corpus consists of 71,388 news stories from multiple sources from English and Mandarin (AP, NYT, CNN, ABC, NBC, MSNBC, Xinhua, Zaobao, Voice of America and PRI the World) in the period of October to December 1998. Machine-translated versions of the non-English stories (Xinhua, Zaobao and VOA Mandarin) are provided as well.
- TDT-4: In TDT-4, LDC defined 60 topics based upon a stratified, random sample of the eight English and seven Chinese news sources collected from October 2000 through January 2001. The seed stories that generated the final 60 topics are equally divided between English and Mandarin.
- TDT-5: Corpus was newly collected from English, Chinese and Arabic from April-September 2003. Unlike previous TDT corpora, TDT-5 does not contain any broadcast news data; all sources are newswire.

To solve the TDT challenges, researchers are looking for robust, accurate, fully automatic algorithms that are source, medium, domain, and language independent.

B. Techniquess

Many techniques have been proposed for topic/ event tracking by researchers. The main focus in this paper is on the latest research being done in this field.

1) Based on Hidden Markov Models: [19] [20] This approach makes use of hidden Markov modeling and clustering techniques. A stream of unsegmented text **(**as might be generated from automatic transcription of broadcast news, for example) is regarded as being composed of a series of "topics" in something like the same way that a stream of speech consists of a series of phonemes. A story on a particular topic can then be viewed as analogous to an utterance of a particular phoneme, and a stream of text can be decoded into a series of topics in the same way that a speech recognizer decodes a stream of speech into a series of phonemes. The boundaries of these topics are identified with story boundaries.
Suppose that there are k topics T(1), T(2), . . . , T(k). There is a language model associated with each topic T(i), 1< i< k, using which one can calculate the probability of any sequence of words. In addition, there are transition probabilities among the topics, including a probability for each topic to transition to itself (the "selfloop" probability), which implicitly specifies an expected duration for that topic. Given a text stream, a probability can be attached to any particular hypothesis about

the sequence and segmentation of topics in the following way: 1) Transition from the start state to the first topic and accumulate a transition probability. 2) Stay in topic for a certain number of words or sentences, and, given the current topic, accumulate a self-loop probability and a language model probability for each. 3) Transition to a new topic, accumulate the transition probability, and go back to step 2.
Because a segmenter operating in this way assigns a topic label to each story, it is possible (with some modifications) to use the same engine to "track," or find successive stories on, an event of special interest. These ideas have been applied in some experiments on the Pilot Study Corpus. The results are promising and suggest that this general methodology is likely to be quite successful both at recovering story boundaries and at identifying instances of stories about the same event over time.

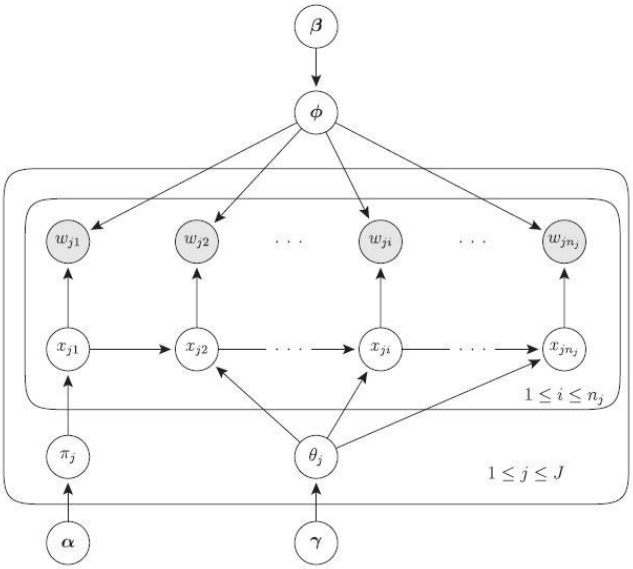

Fig. 3 Architecture of a typical HMM

[31]The above diagram explains Hidden Markov Model.
Each observed variable wji is assumed to be sampled from one of K topic distributions, collectively specified
by Φ which belong to the probability distribution β. Likewise, which one of these distributions is chosen is determined by the value of the probability variable x_{ji}. In the Hidden Markov Model, $x_{j1} \ldots x_{jnj}$ are drawn from a Markov chain specific to the jth text. The parameters of this chain are θ_j (probability of text transitioning to a new topic) and πj (initial probability of a word belonging to a particular topic) which belong to probability distribution γ and α respectively.

2) Probabilistic Model: [21] The probabilistic models for use in detecting and tracking topics in broadcast news stories is presented.
Two different fundamental models for comparing a story to a group on a topic are proposed:
1. The group is the model and the words in the story were generated according to the word distribution of the group, i.e. we are trying to calculate p(T|S) where S is the story and T represents the group of stories on the same group(*BBN topic spotting*).

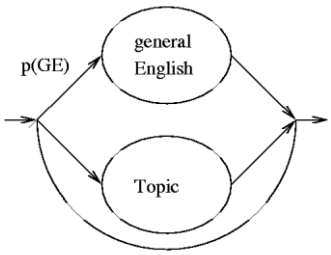

Fig. 4 Probabilistic model

It is represented by the two state model, i.e. for every word it either belong to the topic or is "back-off" to General English Model, Stop words are not considered.
2. The story is the model and the words in the group were generated according to the word distribution of the story, i.e. we calculate p(SisR|T), which is the probability that S is relevant given the topic model(*BBN IR metric*).
Again we use two state model where one state is a unigram distribution estimated from the story S, and the other is the unigram distribution from the whole corpus

Apart from above two, Relevance Feedback and Word feature are used which are similar to BBN IR metric with a difference that only those words are used which are common to two or more of the topic stories.

The tracking system utilizes scores from all four methods:
TS (topic spotting metric), WF (word feature), IR (information retrieval), and RF (relevance feedback). An automatic procedure is utilized to normalize the scores within topic and combine different methods to achieve better results.

Score normalization
Because one threshold is used for all topics, score normalization across topics is important for optimizing system performance. Therefore, statistics on the scores are collected by using the training stories as the test stories, then normalize the test scores based on these statistics for each topic.

Model combinations
Different systems focus on different features of the stories.
Thus, it seems reasonable to combine the probability scores from many tracking systems with a time-decayed prior probability score. This reflects that a test story is less likely to be on-topic as its age increases. A linear combination of the log scores is used from the above four systems and the time decay to form the BBN tracking system. The experiments show a significant reduction of both miss and false alarm rates with a combined system.

3) ETTS: [22] Users or professionals would like to be always updated with the latest hot topics emerging in the particular

information area of their interest. However, due to the fact that the information in the Web is overwhelming and changing dynamically, updating ourselves by browsing through some particular Websites of interest manually and regularly is both a difficult and time consuming job. User can register multiple pages with a tracking system in order to keep watching in a wider area, but the users have to bear in mind that, in one single day, they may receive many emails of acknowledgement just because of some uninteresting changes. And the users would not know this until they go and look for the changes on the pages registered.

An Alternative approach would be that, a system is designed that track the changes to a particular area of user's interests on the Web and generate a summary of emerging topic back to the user [32].

This system consists of three main components, which are Area View System, Web Spider and Summary Generator. Area View System as a Meta-search engine directs the user keyword to a commercial search engine, get the hits, do further analysis and derive a number of most relevance domain sites. Their, Web Spider will dispatch and scan all these domains at a certain time interval to collect all the modified and newly added html pages. Lastly, Summary Generator will first extract all the newly added sentences or Changes from the collected html pages and then count the term weight in the Changes by adapting a newly innovated algorithm TF*PDF (Term Frequency * Proportional Document Frequency). Terms that seem to explain the emerging topic will be heavily weighted. Sentences with the highest average weight will be extracted to form a summary of emerging topic and summary will be returned to the users.
The architecture of ETTS can be shown as:

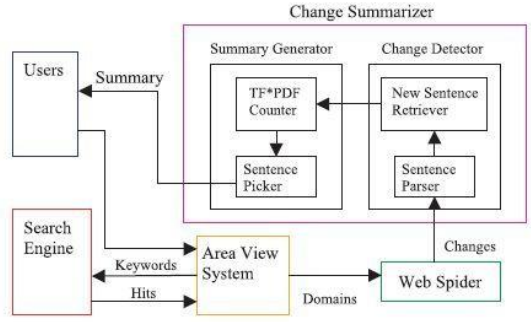

Fig. 5 Architecture of ETTS

4) Lexical Chains: [23] The use of lexical chains to build effective topic tracking systems has been proposed. Lexical chaining is a method of grouping lexically related terms into lexical chains, using simple NLP techniques. [33] A *lexical chain* is a sequence of related words in the text, spanning short (adjacent words or sentences) or long distances (entire text). A chain is independent of the grammatical structure of the text and in effect it is a list of words that captures a portion of the cohesive structure of the text. For example, the following lexical chain, *{mud pie, dessert, mud pie, chocolate, it}* could be constructed given the sentences: *John had mud pie for dessert. Mud pie is made of chocolate. John really enjoyed it*. The word *it* in the third sentence refers back to *dessert* in the first sentence. Before proceeding to the Lexical chain all the stop words are removed and a noun database is used for checking the synonyms or related terms.

[34]Algorithm to form Lexical chain is as follows:
• Candidate terms are selected from the text.
• An appropriate chain is chosen for each candidate word depending on how the word is related to other words in the chain and the word is added to it.
• Otherwise if no appropriate chain is found then the word is used to start a new chain.

In one of the technique for comparing different Lexical chains thus formed following steps are used:
• Calculate the similarity score between two chains, e.g. a very simple score would correspond to the number of term repetitions and/or the number of terms with shared synonym sets.
• Calculate the overall similarity score between two sets of chains based on the pair-wise similarities of chains from the two sets, e.g. the sum of the similarity scores of the most similar pairs of chains.
The resulting score is checked against the threshold and hence tagged as a new event if below threshold or related to an already existing event if above the threshold value.

Following Diagram shows the process of Lexical chain for topic tracking [35]:

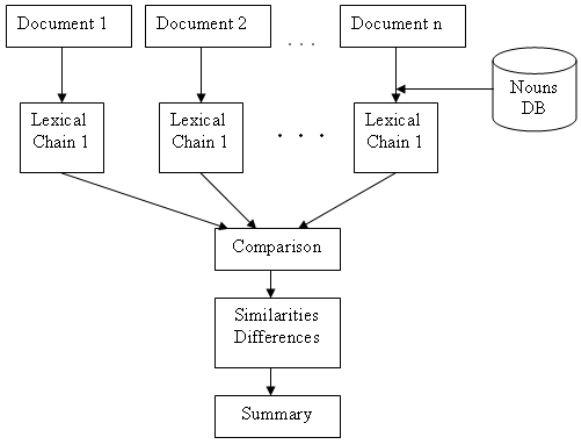

Fig. 6 Lexical chain for topic tracking

5) Time Adaptive Boosting Model: [24]To overcome the existing problems in topic tracking and the shortcomings of current adaptive methods, a new adaptive method, called time adaptive boosting (TAB) model has been proposed. This model adopts the idea of boosting and presents new algorithm to the adaptive learning mechanism in the task of topic tracking. It can solve the problem of topic excursion, and remedy the deficiency of current adaptive methods. Time

sequence is also considered, using the sigmoid functions to express it. Experimental results show that the adaptive method based on TAB can improve the performance of topic tracking.

6) Based on adding semantic role to Dependency Structure language model: [25]Jing Qiu and LeJian Liao propose an idea of adding semantic role to the dependency structure language model. Compare the verbs of the sentences in the stories with a list of verbs related with the verb of the topic. Then, annotate the verbs with semantic roles. This can enable to establish a relation between topics and semantic roles. So, only stories whose sentences containing the right semantic roles are selected. Using this semantic information as an extension of the dependency structure language model in order to reduce the number of stories retrieved by the system, and get a high precision in topic detection and tracking.
First a topic is done. Then the dependency structure language model identifies a set of stories which are more or less relative to this event. Then the verbs of the sentences of these stories are compared with the verb of the topic, and a list of verbs related to it, in order to select only the sentences containing a verb from this list. Next, the selected sentences are annotated with semantic roles making use of semantic role labeling method presented previous. Finally, a set of semantic relationships are applied. These relationships establish a relation between topics and semantic roles. So, only stories whose sentences containing the right semantic roles are selected.

7) Dual Center Event Description Model: [26] Event evolvement and topic shifting can affect the accuracy of event tracking. A dual-center event model is proposed to improve the event tracking process by adjust some attributes of event dynamically. The experiment results show that the dual-center event can improve event tracking effectively under the condition of event evolvement and topic shifting.

Event Center: Event center is defined by all the common attributes owned by all related stories, for event *e*, its center is denoted by C_e.
Center Intensity: Center intensity is defined by the number of stories supporting the event. For a Center *C*, its center intensity is denoted by I_c.
Assume the number of related stories of event *e* is n_e, the event center C_e is represented by $C_e = \{\overline{w}_{e_1}, \overline{w}_{e_2}, ..., \overline{w}_{e_n}\}$ and $I_{c_e} = n_e$. Here $\{\overline{w}_{e_1}, \overline{w}_{e_2}, ..., \overline{w}_{e_n}\}$ is the weight of n features of event *e*. In order to standardize the computation of C_e, the result of normalization of C_e, denoted by $\tilde{C}_e$, is represented by following:

$$\tilde{C}_e = \{\tilde{w}_{e_1}, \tilde{w}_{e_2}, ..., \tilde{w}_{e_n}\}$$

$$\tilde{w}_{e_k} = \frac{\overline{w}_{e_k}}{\sqrt{\sum_{i=1}^{n} \overline{w}_{e_k}^2}}, \text{(k=1,2...,n)}$$

DCEM description
The event is formed by a series of stories. The relationship between stories in an event is analyzed first.
1. When an event just appears, the related stories, called core stories, are familiar with each other. Core stories include some secondary information, called related stories.
2. After a period of time, the development and evolution of the core stories may be gradually formed a number of branches.
3. Around the various branches there will be gradually generate some new event centers, so the multi-center event is formed.

The analysis results show that in spite of an event may include multiple centers, only two centers is enough to be considered. Because multiple centers of events can be seen as the result of development of two centers case, which has the same idea with the TDT Hierarchical Topic Detection (HTD). DCEM can simplify the model to describe the events and its nature.

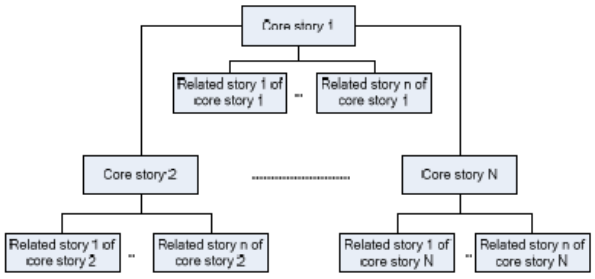

Fig. 7 Relationship between core and related stories

8) Based on Hierarchical Clustering: [27]The common agglomerative hierarchical clustering algorithm is improved based on average-link method, which is then used to implement the retrospective topic detection and the online topic detection of news stories of the stocks. Additionally, the improved single pass clustering algorithm is employed to accomplish topic tracking.
It is considered that the feature terms which occur in the title of a news story contribute more during the similarity calculation and increase their corresponding weights. The results show that the proposed method can effectively detect and track the online financial topics.
The improved single-pass clustering method is used to implement the system. The process of clustering is described as follows:
1. Process the stories every time interval Δt using the improved agglomerative hierarchical clustering algorithm, and all those processed stories are the ones which arise in Δt. We can get the set of candidate topics, i.e. CTS.

2. Get a topic *ct* from CTS, and calculate the similarity algorithm, and all those processed stories are the ones which arise in Δt . We can get the set of candidate topics, i.e. CTS.
2. Get a topic *ct* from CTS, and calculate the similarity between *ct* and each single previous topic within the latest period of time ΔT . If the maximum similarity is smaller than the threshold θ *n* , we consider *ct* a new topic.
3. Delete *ct* from CTS, if CTS is empty, then the algorithm terminates. Otherwise, the algorithm goes to step 2.

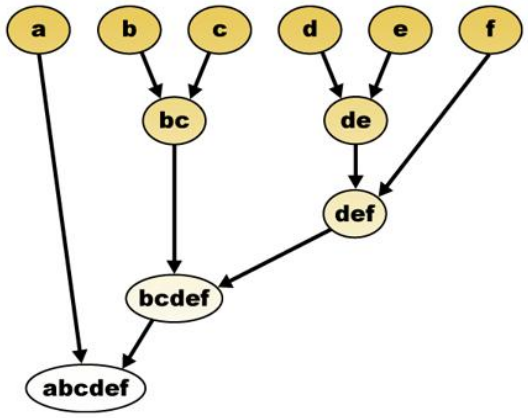

Fig. 8 Agglomerative hierarchical clustering model

9) Based on KeyGraph: Masaki Mori, Takao Miura and Isama Shioya proposed a new approach to observe, summarize and track events from a collection of news web pages. They obtained valid timestamps from a set of temporal web pages and detect events by means of clustering. The valid timestamps can be the creation time (birth time) and the last update time (transaction time) of a web page. Then track the events by using KeyGraph based on the complete link clustering technique, in which two clusters are combines into one if the least similarity values go beyond a given threshold, that is, the two clusters share a topic and basic concepts too. [5]

10) Based on Keyword Extraction: The keyword extraction technique can be used for tracking the topics over time. Keywords are the set of significant words in an article that gives high level description of its contents to readers. But manual keyword extraction is extremely difficult and time consuming task. This problem has been addressed by Sungjick Lee and Han-joon Kim[6]. They proposed an automatic unsupervised keyword extraction technique that includes several variants of the conventional TF-IDF model.

The conventional model evaluates the degree of importance of a word in a single document, but the proposed variants evaluate the degree of importance of a word in a whole document collection. They have also proposed cross-domain filtering for stop word removal and a new measure for term frequency, called table term frequency. It is a two stage system. In the first stage, candidate keywords are extracted. In the second stage, meaningless words are removed by comparing candidate keywords in terms of ranking results.

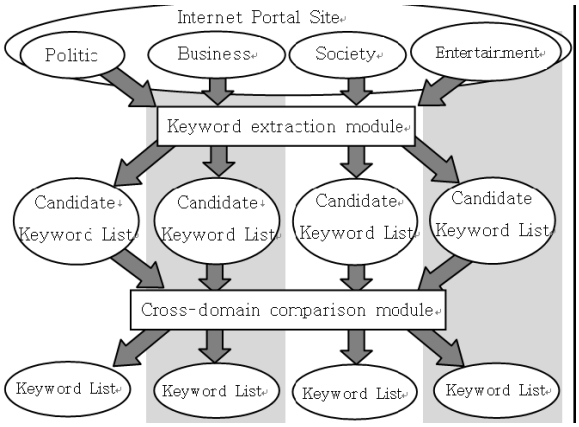

Fig 9. Keyword Extraction System

11) Based on KeyPhrase Extraction: Canhui Wang, Min Zhang, Liyun Ru and Shaoping Ma introduced an automatic online News Topic Keyphrase Extraction System [7]. News topics are related to a set of keywords or keyphrases. News stories are gathered from many websites and organized into news topics by practical web applications. Topics are constructed and updated online automatically using the techniques of burstiness of terms and the aging theory. The proposed system first extracts keyword candidates from single news stories, filters them with topic information and then combines them into phrase candidates using position information. Finally, the phrases are ranked and top ones are selected as topic keyphrases.

12) Based on NER: Wang Xiaowei, JiangLongbin, MaJialin and Jiangyan came up with a new improved approach for topic tracking [9]. Due to the high dimensionality of Vector Space Model, some important characteristics of the text are usually submerged by a lot of weak ability characteristics. They proposed multi vector model that extracts NER features from text and make it into a separate vector. It first selects the features and classify in accordance with characteristics of different tasks, then calculates the vector, then finally selects the combination of model and optimize the parameters. Their experimental result shows that the tracking performance is improved by using multi vector model.

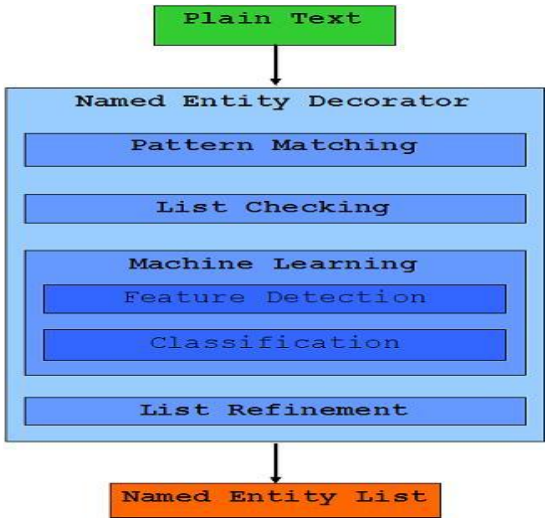

Fig. 10 NER System Architecture

13) Based on Statistical Language Modeling: Maximilian, Michal, Cai-Nicolas and Dietmar proposed an approach which allows to monitor news wire on different levels of temporal granularity, extracting key phrases that reflect short term topics as well as longer term trends by means of statistical language modeling [10]. The focus is on observing the development of topics over time. Modeling these developments over time lend tools to track and analyze topics in a method independent of time slices by themselves. The approach uses several tiers of sliding windows in order to capture topics of varying longevity. Representative keyword vectors are established by discovering salient terms within a topic, such that common topic terms are uncommon within the text corpus in general.

14) Based on LSI-SVM: All these above methods assume the independence of words. These methods use only the frequency information of words to track News stories rather than make a deep semantic analysis from human understanding. Xianfei Zhang, Zhigang Guo and Bicheng Li proposed a new method for News topic tracking [12]. The LSI-SVM (Latent Semantic Analysis- Support Vector Machine) method makes an in depth analysis of the co-occurrence of words and provides a way of dealing with synonymy automatically without the need for a manually constructed thesaurus. It is based on the assumption that there is an underlying or latent structure in the pattern of word usage across document. The experimental results show that LS-SVM outperforms conventional methods and reduces fault and fail rate of topic tracking.

The topic tracking system based on this technique is shown [12]

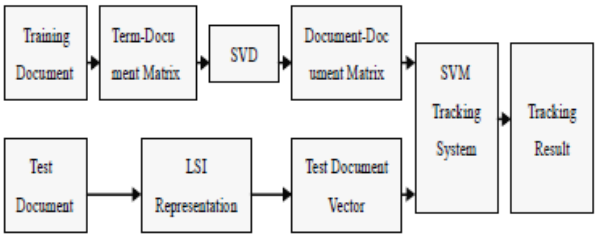

Fig. 11 Architecture of topic tracking system

15) Subtopic based Evolution: Using the present topic tracking models, which is a three level model(topic level, event level and story level), one can't easily catch the development process of the whole topic [13]. Concept Subtopic is put forward and four levels topic model is constructed. Traditional topic evolution analysis uses event as the unit, but we could step deep into the evolution and the analyzing unit to subtopic, find out subtopics aggregated ina short time slice in each event and compute the similarity between different subtopics. Topic structural model can be shown as:

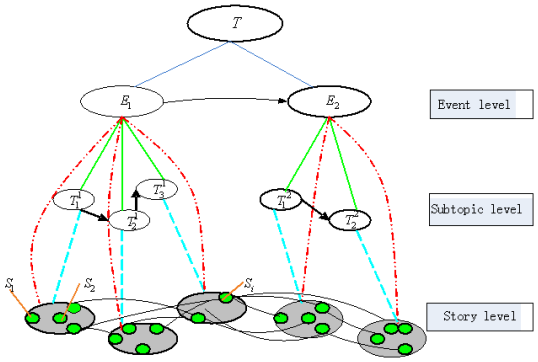

Fig. 12 Topical Structural Model

Leave parameter and feedback parameter are used for topic model renewal, which state that a topic tends to pay more attention to new and relate aspects as compared to old. Subtopic detection is done by using a KNN based single pass cluster algorithm.

16) Monolingual Approach: Anup Kumar Kolya, Asif Ekbal and Sivaji Bandyopadhyay proposed a simple approach for monolingual event tracking system in Bengali [14]. The goal of the system is to determine whether two news documents within a range of dates describe the same event. An event vector is created using the NER(Name Entity Recognition) system to tag the news stories with the person, location, organization and miscellaneous NE tags, such as date, time, number, percentages, monetary expressions and measurement expressions. It considers both the language dependent and language independent features, which helps in improving the performance of NER system. A particular news document is described as a collection of such event vectors. A particular

threshold value is considered to check whether the number of event vectors of two separate news documents match at least by this threshold by measuring similarity between the event vectors.

17) Proposed Incremental Algorithm: The problem of decreasing readership for Egyptian newspapers websites which failed to represent the most inetersting topics has been addressed by Laila Mohamed ElFangary by applying an enhanced algorithm for mining incremental updates on an Egyptian Newspaper Website for improved topic tracking. [15]. The Proposed Incremental Algorithm(PIA) is designed to efficiently update large itemsets by taking set of previously discovered rules into account using some hypothesis to remove some of the old large itemsets or to add new large itemsets without doing much work. The target data set are collected related to news paper visitors and their navigation through the website news, their opinion about different topics. Associations are discovered between different kinds of topics, frequent itemsets, interesting and non interesting rules. He concluded in his work that the site should be dynamically changing according to the user interests to increase the number of visitors.

18) Based on VSM: [16] [17] A new topic tracking prototype system has been proposed based on VSM that consists of topic representation model, KNN algorithm for text classification and TDT evaluation method.

It uses VSM as one of the better performing topic representation methods. Texts are seen as vector space composed of a set of orthogonal term vectors. The words which can not well represent text classification information are removed and only those words are retained that are useful for classification. It also compares information gain and chi square and experimental results show that chi square in VSM has better performance for topic tracking algorithm than information gain algorithm.

17) Related Topic Network: To improve the performance of existing topic tracking methods, Chao Chang, Daniel Zeng and Huimin Zhao proposed a new more efficient model named Related Topic Network with a new term weighting method [18]. It represents the structure of topics with its nodes representing topics and its edges representing the connections between topics. It does not impose restrictive rules and more efficient as it only needs term matching and term co-occurrence analysis in extracting related topics. A story is represented as a vector consisting of weights of terms, weights are calculated using a combined TF-IDF and TF-IWF approach, called TF-WF/DF. Clustering methods are applied to detect topics based on pair wise similarities between stories.

20) Based on ensemble of all existing systems: Most of the state-of-art study on topic tracking are based on document/ topic representation, incremental tracking, adapted tracking, feature selection, calculation of term weight, threshold selection, model to calculate similarity, classification algorithm. Xiangju Qin and Yang Zhang improved the performance of topic tracking by ensembling several topic tracking systems by using majority voting [11]. After integrating many tracking systems with large individual performance variances, can average the individual performance variances to produce a tracking system with much less variance. The experimental results show that this ensemble topic tracking system performs better than each individual topic tracking system.

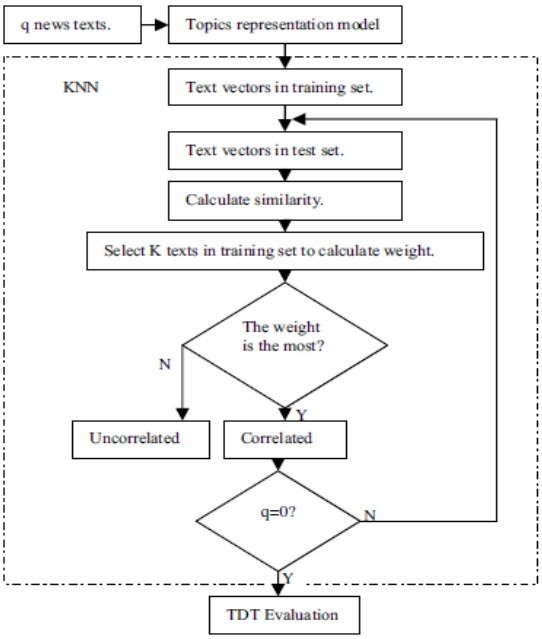

Fig. 13 Architecture of topic tracking prototype system

21) Improved K-Nearest neighbor classification in topic tracking: [28] This approach makes effective improvement on traditional KNN taxonomy and applies it to topic tracking; besides, it adds time window strategy to the process of topic tracking, which effectively reduces calculation complexity. The final experimental result also proves that this method is superior to traditional KNN topic tracking method.

Improved KNN Topic Tracking Method:

(1) Adopt named entity as feature item; conduct feature selection on training text set using evaluation function; Calculate weight information of each feature vector and represent the extracted feature words with vector;

(2) After the new text x arrives, decide whether this text's "life" time is within the time window of topics we are interested in. If so, carry out next step; otherwise, judge this text as "not belonging" and deal with the next new text;

(3) Conduct word segmentation on new text according to feature words and determine the vector representation of new text;

(4) Calculate the cosine similarity values cos(x, di)

between new text x and training text set di and select k texts most similar to new text;
(5) Extract kl texts belonging to positive examples from k texts to construct a positive example set P (x, k 1); find sum of these texts and similarity value of new text cos(x, di) as similarity value between new text x and positive example set P (x, kl). Similarly, extract k2 texts belonging to negative examples to form a negative example set N (x, k2); find sum of these texts and similarity value of new text x cos(x, di) as similar value between new text x and negative example N (x,k2);
(6) Compare the similarity value of new text with values of positive and negative examples;
(7) When p (x, E) >O, it's judged that new text x belongs to this topic; when p (x, E) < 0, it's decided that new text x does not belong to this topic.

III. CONCLUSIONS

In this paper, we have discussed topic tracking as an important issue in text mining. Topic tracking monitors a stream of news stories and find out what discuss the same topic described by a few positive samples. We have also discussed some of its applications and the historical background which details its development through five TDT tasks. The main focus in this paper is on different techniques that have been designed for topic tracking such as hidden markov models, VSM based, KNN classification, ETTS, hierarchical clustering, subtopic similarity, DCEP model, NER based, probabilistic models, lexical chaining, time adaptive boosting model, ensemble method, keygraph approach, LS-SVM. The algorithms of some of these techniques have also been presented. Out of these, related topic network, improved KNN, hierarchical clustering are some of the recent techniques which are improvements over the previous ones.

REFERENCES

[1] Milos Radovanovic, Mirjana Ivanovic, (2008), "Text Mining: Approaches and Applications",Novi Sad J. Math, Vol. 38, No. 3: 227-234
[2] Anna Stavrianou, Periklis Andritsos, Nicolas Nicoloyannis, (2007), "Overview and Semantic Issues of Text Mining", Sigmod Record, Vol. 36, No. 3
[3] Navathe, Shamkant B. and Elmasri Ramez,(2000), "Data Warehousing and Data Mining",in "Fundamentals of Database Systems", Pearson Education pvt Inc, Singapore, 841-872
[4] Vishal Gupta, G.S. Lehal, (2009), "A Survey of Text Mining Techniques and Applications",in Journal of Emerging Technologies in Web Intelligence
[5] Masaki Mori,Takao Miura, Isamu Shioya, (2006), "Topic Detection and Tracking for News Web Pages", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence
[6] Sungjick Lee, Han-joon Kim,(2008), "News Keyword Extraction for Topic Tracking", 4th International Conference on Networked Computing and Advanced Information Management, IEEE
[7] Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma,(2008), " An Automatic Online News Topic Keyphrase Extraction System",IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
[8] JingQiu, LeJian Liao, XiuJie Dong, (2008), " Topic Detection and Tracking for Chinese News Web Pages", International conference on Advanced Language Processing and Wen Information Technology, IEEE
[9] Wang Xiaowei, JiangLongbin, MaJialin, Jiangyan, (2008), "Use of NER Information for Improved Topic Tracking", Eighth International Conference on Intelligent Systems Design and Applications, IEEE
[10] Maximilian Viermetz, Michal Skubacz, Cai-Nicolas Ziegler, Dietmar Seipel, (2008),"Tracking Topic Evolution in News Environment", 10th IEEE Conference on E-Commerce Technology and 5th IEEE Conference on Enterprise Computing, E-Commerce and E-Services
[11] Xiangju Qin, Yang Zhang, (2008), "Improving the performance of Topic Tracking System by Ensemble", International Conference on Computer Science and Software Engineering, IEEE
[12] Xianfei Zhang, Zhigang Guo, Bicheng Li, (2009), " An Effective Algorithm of News Topic Tracking", Global Congress on Intelligent Systems,IEEE
[13] Yan Liu, Nan Lv, Junyong Luo, Huijie Yang, (2009), "Subtopic Based Topic Evolution Analysis", International Conference on Web Information Systems and Mining, IEEE
[14] Anup Kumar Kolya, Asif Ekbal, Sivaji Bandyopadhyay, (2009), "A Simple Approach for Monolingual Event Tracking System in Bengali", 8th International Symposium on Natural Language Processing, IEEE
[15] Laila Mohamed ElFangray, (2009), " Applying an Enhanced Algorithm for Mining Incremental Updates on an Egyptian Newspaper Website", 5th International Joint Conference on INC, IMS and IDC, IEEE
[16] Shengdong Li, Xueqiang Lv, Yuqin Li, Shuicai Shi,(2009), " Study on Feature Selection Algorithm in Topic Tracking"
[17] Shengdong Li, Xueqiang Lv, Qiang Zhou, Shuicai Shi,(2010), " Study on Key Technology of Topic Tracking Based on VSM", Proceedings of the 2010 IEEE International Conference on Information and Automation, June 20-23, Harbin, China
[18] Chao Chang, Daniel Zeng, Huimin Zhao,(2010), " Related Topic Network", IEEE
[19] J.P.Yamron, Carp L. Gillick, S.Lowe, P. Van Mulbregt, (1997), "Event Tracking and Text Segmentation via Hidden Markov Models",IEEE
[20] J.P.Yamron, Carp L. Gillick, S.Lowe, P. Van Mulbregt, (1998), "A Hidden Markov Approach to Segmentation and Event Tracking",IEEE
[21] F. Walls, H. Jin, S. Sista and R. Schwartz, (1999), "Probabilistic Models for Topic Detection and Tracking",IEEE

[22] Khoo Khyou Bun and Mitsuru Ishizuka, (2001), "Emerging Topic Tracking System",IEEE
[23]Joe Carthy, Micheal Sherwood, (2002), "Lexical chains for topic tracking", IEEE
[24] Huizhen Wang, Jingbo Zhu, Duo ji, Na Ye and Bin Zhang, (2005), "Time Adaptive Boosting Model for Topic Tracking", Proceeding of NLP-KE'05,IEEE
[25] Jing Qiu and LeJian Liao, (2007), "Add Semantic Role to Dependency Structure Language Model for Topic Detection and Tracking", 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, IEEE
[26] Wei Wang, JunZheng, Wu Yang and Yongtian Yang, (2008), "A Dual Center Event Description Model Used in Event Tracking", 2009 World Congress on Computer Science and Information Engineering,IEEE
[27]Xiang Ying Dai, Qing Cai Chen, Xiao Long Wang and Jun Xu, (2010), "Online Topic Detection and Tracking of financial News based on Hierarchical Clustering", Proceedings of the 9th International Conference on Machine Learning and Cybernetics, IEEE
[28] Hongxiang Diao, Zhansheng Bai and Xilin Yu, (2010), "The Application of Improved K-Nearest Neighbor Classification in topic tracking", International Conference on Educational and Information Technology, IEEE
[29] Omid Dadgar,"Topic Detection and tracking", Available: www.tcnj.edu/~mmmartin/.../TDT/TopicDetectionTracking04 .ppt
[30] Topic Detection and Tracking, Available: www.projects.ldu.upenn.edu
[31] Mark Andrews, Gabriella Vigliocco, (2009), "The Hidden Markov Topic Model: A Probabilistic Model of Semantic Representation"
[32] Khoo Khyou Bun , Mitsuru Ishizuka, (2005), " Emerging topic tracking system in www", Knowledge-Based Systems164–171
[33] Joe Carthy, Alan F. Smeaton, "The Design of a Topic Tracking System"
[34] Paula Hatch, Nicola Stokes and Joe Carthy, "Topic detection, a new application for Lexical Chaining"
[35]Yan-Min Chen, Xiao-Long Wang and Bing-Quan Liu,"Multi-Document Summarization based on Lexical Chains"

