# ORGANIZING AND SUMMARIES OF CONTENT USING TEMPORAL SIMILARITY

V.Narahari*1, S.Md.Ibrahim*2

M.Tech, Dept of CSE, SKDEC, Gooty, D.t: Anantapuram, A.P, India

M.Tech, Assistant Professor, Dept of CSE, SKDEC, Gooty, D.t: Anantapuram, A.P, India

## ABSTRACT

Temporal Text Mining (TTM) is concerned with discovering temporal patterns in text information collected over time. Since most text information bears some time stamps, TTM has many applications in multiple domains, such as summarizing events in news articles and revealing research trends in scientific literature. In this paper, we study a particular TTM task {discovering and summarizing the evolutionary patterns of themes in a text stream. We define this new text mining problem and present general probabilistic methods for solving this problem through (1) discovering latent themes from text; (2) constructing an evolution graph of themes; and (3) analyzing life cycles of themes. Our approach to this problem combines an extension of Factorial Hidden Markov models for topic intensity tracking with exponential order statistics for implicit data association. Experiments on text and email datasets show that the interplay of classification and topic intensity tracking improves the accuracy of both classification and intensity tracking. Even a little noise in topic assignments can mislead the traditional algorithms.

However, our approach detects correct topic intensities even with 30% topic noise.

**KEYWORDS:** Hidden Markov Model, Intensity Tracking, Topic Segmentation, Classification.

## I.INTRODUCTION

When following a news event, the content and the temporal information are both important factors in understanding the evolution and the dynamics of the news topic over time. When recognizing human activity, the observed person often performs a variety of tasks in parallel, each with a different intensity, and this intensity changes over time. Both examples have in common a notion of classification: e.g., classifying documents into topics, and actions into activities. Another common point is the temporal aspect: the intensity of each topic or activity changes over time. In a stream of incoming email for example, we want to associate each email with a topic, and then model bursts and changes in the frequency of emails of each topic. A simple approach to this problem would be to first consider associating each email with a topic using some supervised, semi-supervised or unsupervised (clustering) method; thus segmenting the joint stream into a stream for each topic. Then, using only data from each individual topic, we could identify bursts and changes in topic activity over time. In this traditional view (Kleinberg, 2003), the data association (topic segmentation) problem and the burst detection (intensity estimation) problem are viewed as two distinct tasks. However, this separation seems unnatural and introduces additional bias to the model. We combine the tasks of data association and intensity tracking into a single model, where we allow the temporal information to influence classification. The intuition is that by using temporal information the classification would improve, and by improved classification the topic intensity and topic content evolution tracking also benefit. Our approach combines an extension of Factorial Hidden Markov models (Ghahramani & Jordan, 1995) for topic intensity tracking with exponential order statistics for implicit data association. Additionally, we demonstrate the use of a switching Kalman Filter to track content evolution of the topic over time. Our approach is general in the sense that it can be combined with a variety of learning techniques; we demonstrate this flexibility by applying it in supervised and unsupervised settings. Experimental results show that the interplay of classification and topic intensity tracking improves accuracy of both classification and intensity tracking. More specifically, our contributions are:

- A suite of models, EDA–IT, IDA–IT and IDA–ITT, for simultaneous reasoning about topic labels and topic intensities, and extensions to topic drift tracking.
- A modeling trick which uses exponential order statistics to achieve implicit data association.

This idea allows us to make an intractable data association problem tractable for exact inference, and is of independent interest.

- The extensive empirical evaluation in the supervised and unsupervised setting on synthetic as well as two real world datasets.

Managing the explosion of electronic document archives requires new tools for automatically organizing, searching, indexing, and browsing large collections. Recent research in machine learning and statistics has developed new techniques for finding patterns of words in document collections using hierarchical probabilistic

These models are called "topic models" because the discovered patterns often reflect the underlying topics which combined to form the documents. Such hierarchical probabilistic models are easily generalized to other kinds of data; for example, topic models have been used to analyze images , biological data, and survey data. Department are assumed to be independently drawn from a mixture of multinomial's. The mixing proportions are randomly drawn for each document; the mixture components, or topics, are shared by all documents. Thus, each document reflects the components with different proportions. These models are a powerful method of dimensionality reduction for large collections of unstructured documents. Moreover, posterior inference at the document level is useful for information retrieval, classification, and topic-directed browsing. Treating words exchangeable is a simplification that it is consistent with the goal of identifying the semantic themes within each document. For many collections of interest, however, the implicit assumption of exchangeable *documents* is inappropriate. Document collections such as scholarly journals, email, news articles, and search query logs all reflect evolving content. For example, the *Science* article "The Brain of Professor Laborde" may be on the same scientific path as the article "Reshaping the Cortical Motor Map by Unmasking Latent Intracortical Connections," but the study of neuroscience looked much different in 1903 than it did in 1991. The themes in a document collection evolve over time, and it is of interest to explicitly model the dynamics of the underlying topics.

## II.RELATEDWORK

RED was firstly proposed and defined by Yang et al, and an agglomerative clustering algorithm (augmented Group Average Clustering, GAC) was proposed in that paper, but since then there are few right-on-the-target research work reported. But a similar topic, New Event Detection (NED), has been extensively studied. It is noted that some researchers use very similar algorithms to perform both NED and RED. Thus, we mainly review the previous work on NED in this section. The most prevailing approach of NED was proposed by Allan et al. and Yang et al. , in which documents are processed by an on-line system. In such on-line systems, when receiving a document, the similarities between the incoming document and the known events (sometime represented by a centroid) are computed, and then a threshold is applied to make decision whether the incoming document is the first story of a new event or a story of some known event. Modifications to this approach may be summarized from two aspects: better representation of contents and utilizing of time information. From the aspect of utilizing the contents, TF-IDF is still the dominant technique for document representation, and cosine similarity is the generally used similarity metric. However, many modifications have been proposed in recent years. Some work focus on finding new distance metrics, such as the Hellinger distance metric [5]. But more works focus on finding better representations of documents, i.e. feature selection. Yang et al. classified documents into different categories, and then removed stop words with respect to the statistics within each category. Significant improvements were reported by them. The usage of named entities have been studied, such as in Allan et al. [2], Yang et al. and Lam et al., but there are yet no generally acknowledged conclusions on whether named entities are useful. Reweighting of terms is another prevailing method, firstly proposed by Allan et al. in [2]. In, Yang et al. proposed to re-weight both named entities and non-named terms with respect to statistics within each category. A recent publication of Kumaran et al. [6] summarized the work in this direction and proposed some extensions. They exploited to use both text classification and named entities to improve the performance of NED. In their work, stop words are removed conditioned on categories, similar with the method of Yang et al., but they relaxed the constraint on document comparison: the incoming document were compared with all documents instead of only documents belonging to the same category. Then each document was represented by three vectors: the whole terms, named entities and no named entity terms. But there are no consistently best representations of documents for all categories. From the aspect of utilizing time information, generally speaking, there are two kinds of usages. Some approaches, such as the on-line nearest neighbor approach discussed above, only use the chronological order of documents. The other approaches, such as  and [5] use decaying functions to modify the similarity metrics of the contents. A unique thinking of NED is proposed by Zhang et al. , in which the authors distinguished the concepts of relevance and redundancy, and argue that relevance and redundancy should be modeled separately.

## III. PROPOSED SYSTEM EVOLUTION

### 3.1 Data Association for Topic Intensity Tracking

When following a news event, the content and the temporal information are both important factors in understanding the evolution and the dynamics of the news topic over time. When recognizing human activity, the observed person often performs a variety of tasks in parallel, each with a different intensity, and this intensity changes over time. Both examples have in common a notion of classification: e.g., classifying documents into topics, and actions into activities. Another common point is the temporal aspect: the intensity of each topic or activity changes over time.

In a stream of incoming email for example, we want to associate each email with a topic, and then model bursts and changes in the frequency of emails of each topic. A simple approach to this problem would be to first consider associating each email with a topic using

some supervised, semi-supervised or unsupervised (clustering) method; thus segmenting the joint stream into a stream for each topic. Then, using only data from each individual topic, we could identify bursts and changes in topic activity over time. In this traditional view (Kleinberg, 2003), the data association (topic segmentation) problem and the burst detection (intensity estimation) problem are viewed as two distinct tasks. However, this separation seems unnatural and introduces additional bias to the model. We combine the tasks of data association and intensity tracking into a single model, where we allow the temporal information to influence classification. The intuition is that by using temporal information the classification would improve, and by improved classification the topic intensity and topic content evolution tracking also benefit.

## IV. MULTIMODAL RESTOSPECTIVE NEWS EVENT DETECTION METHOD

As mentioned above, both news articles and events could be represented by two kinds of information: contents and timestamps. These two kinds of information have different characteristics, thus, we propose a multi-modal approach to incorporate them in a unified probabilistic framework.

**4.1 Representations of News Articles and News Events**

According to the knowledge about news, news articles can be further represented by four kinds of information: who (persons), when (time), where (locations) and what (keywords). Similarly, a news event also can be represented by persons, time(defined as the period between the first article and the last article), locations and keywords. For news article, the timestamp is a discrete value, while for news event, its time consists of two values. As a result, we define news article and event as: article = {persons, locations, keywords, time} event = {persons, locations, keywords, time} The keywords represent the remainder contents after removing named entities and stop words. The contents of news articles are divided into three kinds of information. In order to simplify our model, we assume the four kinds of information of a news article are independent:p(article) = p(persons)p(locations)p(keywords)p(time)

Usually, there are many named entities and keywords in news articles, and we generally term them as entity in this paper. As a result, there are three kinds of entities, and each kind of entity has its own term space.

**4.2 The Generative Model of News Articles**

According to the first characteristic of news articles and events, the generation processes of news articles can be modeled by a generative model. Since contents and timestamps of news articles are heterogeneous features, we model them with different types of models.

Contents The bag of words model is an effective representation of documents, and the Naive Bayes (NB) classifier basing on this model works very well on many text classification and clustering tasks [8]. Thus, just like in NB, we use mixture of unigram models to model contents. It is important to note that person and location entities are important information of news articles, but they only take a small part of the contents. If we model the whole contents with one model, this important information may be overwhelmed by keywords. Thus, we model persons, locations and keywords by three models, although as will cause extra computational cost.

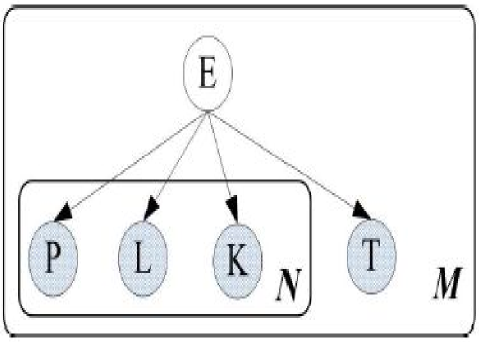
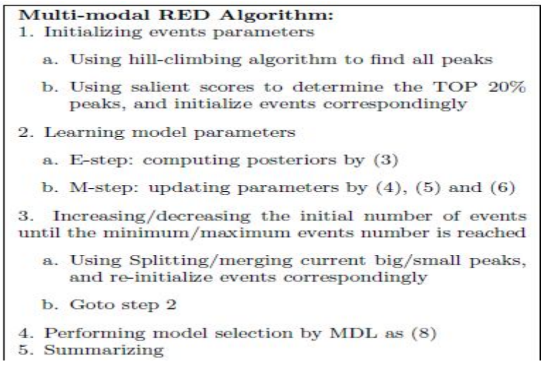


Figure 2: Graphical model representation of the generative model of news articles. $E$, $P$, $L$, $K$ and $T$ represent events, persons, locations, keywords and time respectively. Shadow nodes are observable; otherwise is hidden. $N$(entities) and $M$(articles) at the bottom-right corners represent plates.

Timestamps As mentioned in the previous section, each event corresponds to a peak on articles count-time distribution whether it can be observed or not. In other words, the distribution is a mixture of many distributions of events. A peak is usually modeled by a Gaussian function, where the mean is the position of the peak and the variance is the duration of event. As a result, Gaussian Mixture Model (GMM) is chosen to model timestamps. Consequently, the whole model is the combinations of the four mixture models: three mixture of unigram models and one GMM.

**4.3 Event Summarization**

In practice, we utilize two ways to summarize news events. On the one hand, we can choose some features with the maximum probabilities to represent event. For example, for event j, the ’protagonist’ is the person with the maximum p(personp|ei). Locations and keywords can be chosen similarly. However, the read abilities of such summarizations are bad. Thus, as an alternative way, we choose one news article as the representative for each news event.



```
Multi-modal RED Algorithm:
1. Initializing events parameters
    a. Using hill-climbing algorithm to find all peaks
    b. Using salient scores to determine the TOP 20%
       peaks, and initialize events correspondingly
2. Learning model parameters
    a. E-step: computing posteriors by (3)
    b. M-step: updating parameters by (4), (5) and (6)
3. Increasing/decreasing the initial number of events
   until the minimum/maximum events number is reached
    a. Using Splitting/merging current big/small peaks,
       and re-initialize events correspondingly
    b. Goto step 2
4. Performing model selection by MDL as (8)
5. Summarizing
```

**Figure 4: Summary of the proposed multi-modal RED algorithm**

## V.APPLICATION: HISCOVERY SYSTEM

Based on the proposed event detection algorithm, we build a research system, HISCOVERY (History discovery), in which we provide two useful functions: Photo Story and Chronicle. In HISCOVERY, news articles come from 12 news sites, such as MSNBC, CNN and BBC. We run a web crawler once to get old news articles, and from then on, only trace the front pages of these sites to get the latest news articles.

### 5.1 Photo Story

Photo story is a rich representation of the past news even ts belonging to certain topic(e.g. “Halloween” is a topic, but each year’s Halloween is an event). Usually, there are informative images embedding in news articles, which are very helpful to illustrate news events. By the proposed RED approach, news articles and their images are associated with found events. Figure 5 illustrates the user interface of Photo Story. Events and summaries are shown by their temporal order. We also use computer vision technologies to detect attention attracting areas (e.g. human faces), and then make a slides-show which emphasize on these areas.

### 5.2 Chronicle

Chronicles (e.g. chronicle of George W. Bush) provide very useful information, which are made manually by editors or history researchers nowadays. In HISCOVERY, the generation of a chronicle is constituted by three steps: i) user enters a topic, just like the query in Web search engine, Figure 5: User interface of Photo Story. The bottom area shows events arranged in temporal order(each event is represented by a cluster of images), and the circled event is current event; the slide show of current event is provided at the top left area; and corresponding summary is presented at the top right area

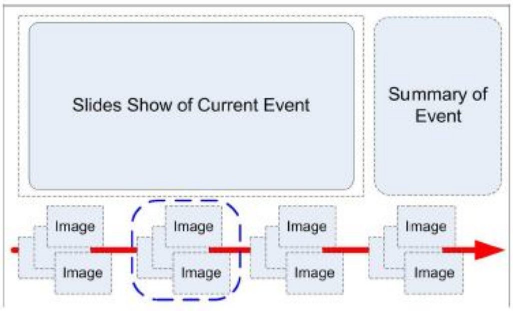
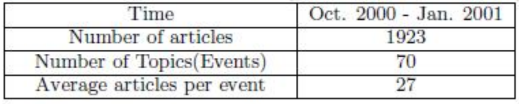


**Figure 5: User interface of Photo Story. The bottom area shows events arranged in temporal order(each event is represented by a cluster of images), and the circled event is current event; the slide show of current event is provided at the top left area; and corresponding summary is presented at the top right area**

**Table 1: Details of dataset 1(part of TDT 4 dataset)**



| Time | Oct. 2000 - Jan. 2001 |
|---|---|
| Number of articles | 1923 |
| Number of Topics(Events) | 70 |
| Average articles per event | 27 |

ii) HISCOVERY searches our news corpus to gather related articles, and iii) the system utilizes the proposed RED approach to detect events belonging to this topic, and then sort summaries of events in chronological order. Since we have the images of the events, some representative images can also be shown in the final report.

## VI.CONCLUSIONS

Text streams often contain latent temporal theme structures which reject how different themes inuence each other and evolve over time. Discovering such evolutionary theme patterns can not only reveal the hidden topic structures, but also facilitate navigation and digestion of information based on meaningful thematic threads. In this paper, we propose general probabilistic approaches to discover evolutionary theme patterns from text streams in a completely unsupervised way. To discover the evolutionary theme graph, our method would first generate word clusters (i.e., themes) for each time period and then use the Kullback-Leibler divergence measure to discover coherent themes over time. Such an evolution graph can reveal how themes change over time and how one theme in one time period has inuence other themes in later periods. We also propose a method based on hidden Markov models for analyzing the life cycle of each theme. This method would first discover the globally interesting themes and then compute the strength of a theme in each time period. This allows us to not only see the trends of strength variations of themes, but also compare the relative strengths of different themes over time.

## VII REFERENCES

[1] Topic detection and tracking(tdt) project. homepage:http://www.nist.gov/speech/tests/tdt/.2012

[2] J. Allan, H. Jin, M. Rajman, C. Wayne, G. D., L. V., R. Hoberman, and D. Caputo. Summer workshop final report. In Center for Language and Speech Processing, 2009.

[3] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In Proc. of SIGIR Conference on Research and Development in Information Retrieval, 2008.

[4] D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. Machine Learning, 2009.

[5] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In Proc. of the SIGIR conference on Research and development in information retrieval, 2003.

[6] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In Proc. of the SIGIR Conference on Research and Development in Information Retrieval, 2004.

[7] W. Lam, H. Meng, K. Wong, and J. Yen. Using contextual analysis for news event detection. International Journal on Intelligent Systems, 2001.

[8] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. Machine Learning, 2000.

[9] A. Strehl, J. Ghosh, and R. Mooney. Impact of the similarity measures on web-page clustering. In Proc. of the AAAI 2000 Workshop on AI for Web Search, 2000.

[10] J. F. Trevor Hastie, Robert Tibshirani. The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, 2001.

V.Narahari M.Tech , Department of CSE
SRI KRISHNA DEVARAYA ENGINEERING COLLEGE,
Gooty, Ananthapuram (Dt).

S.Md.Ibrahim M.Tech,
Assistant Professor, Department of CSE, SRI KRISHNA DEVARAYA ENGINEERING COLLEGE,
Gooty, Ananthapuram (Dt).