Incremental Novelty Learning in Adaptive Topic Tracking

Yu Hong, Yu Zhang, Ting Liu and Sheng Li
School of Computer Science and Technology, Harbin Institute of Technology
P.O.Box 321, Harbin Institute of Technology, No.27, Jiaohua Street, Nangang District,
Harbin, Heilongjiang, P.R.China, 150001
{hy, zhangyu, tliu}@ir.hit.edu.cn, lisheng@hit.edu.cn

Abstract

As an important task in Topic Detection and Tracking (TDT), Topic tracking aims to monitor story stream arranged in temporal order and identify relevant stories to given topics. On this basis Adaptive Topic Tracking focuses on modifying topic descriptions by learning feedbacks in real time to improve the accuracy of tracking. Many researches have achieved substantial performances in this field, however most of them attempt to resolve the issues of tracking in use of existing techniques in Information Retrieval which often ignore the characteristics of TDT corpus such as timeliness, novelty and burst of news stories. In this paper we propose an incremental novelty learning algorithm (INL) which learns the evolution of topic based on the incremental distribution of novel terms and modifies the topic description by combining seminal terms with novel ones in real time. The algorithm focuses on increasing the efficiency of updating the kernel of topic description and improving the ability of tracking novel events suddenly occurred in story stream. We compare our method with some existing tracking systems on TDT4 corpus, which demonstrates INL can substantially improve the accuracy of adaptive topic tracking.

Keywords

Adaptive topic tracking; novelty learning; novel event; temporal order; topic model; event;

1. INTRODUCTION

Topic Detection and Tracking (TDT) aim at identifying and organizing information from a stream of news stories. It provides a new multilingual platform for evaluating techniques of information retrieval, data mining and natural language processing (James, Jaime et al. 1998). Topic tracking task focuses on detecting on-topic stories (viz. relevant stories to given topics) in the story stream, and a basic tracking system includes three main components, namely topic model establishment, determination mechanism and threshold estimation. The topic model is used to describe the kernel of a topic, and it is the most important frame of reference to determine on-topic stories newly reported. The determination mechanism is constructed by rules, statistical models or machine learning algorithms which measure the relevance between the topic model and stories. And the threshold is the decisive boundary dividing on-topic stories and off-topic ones (viz. irrelevant stories to given topics), that is, stories whose relevance sores are higher than the threshold will be determined as on topic, otherwise off topic.

Topic tracking is very similar to Information Filtering (IF) which retrieves relevant information of static profiles from dynamic data stream. So it has the same difficulty with Information Filtering in identifying the evolving relevant information on the condition that its initial topic model never update. Therefore most recent researches focus on developing adaptive learning models which modify the topic model in real time by learning feedback in the process of tracking. A tracking system which includes adaptive learning model is named adaptive topic tracking. It can be divided into two subtasks, respectively supervised adaptive tracking and unsupervised adaptive tracking. The supervised one can adopt relevant feedbacks, which are the on-topic stories labeled by human in TDT corpus, to carry out its adaptive learning. On the contrary, the unsupervised one is only permitted to use pseudo-relevant feedbacks that means the on-topic stories determined by machine may be actually irrelevant to topic model but still used in learning process without supervision of human (Margaret et al. 2004).

The core issue of the adaptive learning model is to explore laws of evolvement of information and mine the real-time kernel of relevant information. It is difficult for Information Filtering to resolve the issue because the data it processed are always irregular. But it is different for Topic Tracking who processes the data, as news stories, arranged in temporal order (Li et al. 2006). This characteristic of story offers optimum advantages for analyzing the tendency of topic shifting. In TDT, topic is defined to be a seminal event along with all directly related events. And a story is an audio or textual file discussing events happening at some specific time. Thus topic shifting actually represents the phenomenon that the kernel of topic focuses on different relevant events at various times. Therefore it is a significant attempt for adaptive learning in TDT to explore the laws of topic shifting along with events occurred in temporal order.

Some researches have achieved substantial improvement of Topic Tracking by adopting adaptive learning mechanism, such as the incremental learning algorithm of James Allan. However these works ignore the features of novelty and burst of event which often cause topic shifting. In this paper we attempt to demonstrate the kernel of topic will shift when novel events are reported. On the basis, we propose an incremental novelty learning model which modifies topic model by integrating seminal terms with novel terms, especially the terms burst occurred in short period of time.

The remainder of this paper is organized as follows. In section 2, we firstly introduce a basic topic tracking system; secondly it is illustrated that the static topic model in the basic system can not absolutely capture the kernels of some on-topic stories at the late stage of tracking; at last we verify whether the static topic model can be improved by the existing statistical strategies. In section 3, we firstly demonstrate adaptive learning can improve the ability of tracking system capturing kernels of on-topic stories when topic shift; secondly it is illustrated that the novelty and burst of events can be used to improve the performance of adaptive learning; at last we gives the incremental novelty learning model. Experiments and results are given in Section 4. Section 5 concludes the paper.

2. UNRECONSTRUCTED STATIC TOPIC MODEL

2.1 A Basic Topic Tracking

A basic topic tracking system includes three components: topic model establishment, determination mechanism and threshold estimation as mentioned in last section. Here, we adopt a simple tracking system proposed by James Allan as our basic system which achieves the best performance in the pilot TDT study (James, Jaime et al. 1998).

Any topic in the system is represented as a vector space model (VSM) including 50 most frequent terms occurring in training stories (Ron et al. 1999). Specifically, the foremost four (N_t=4) on-topic stories are used as training data for a topic and a vector of the 50 most frequent terms in the training data are used as the topic model. In fact, there are only 10 terms used to build the topic model of the system in pilot TDT study because it is believed that capturing several "killer terms" is sufficient to tracking relevant events with high accuracy (James, Ron et al. 1998). However it has been illustrated that tracking performance can be improved by adding a little more terms into topic model, such as 50 terms in TDT 2002 (James, Victor et al. 2002).

In the determination mechanism of the system, a story is described as the VSM of 50 most weighted terms. The weighting function is denoted as follows:

$$d_i = \frac{tf}{tf+2} \cdot (1 - \log_N df_i) \qquad (1)$$

where d_i is the weight of term i in the story, tf is the number of times the term occurs in the story, df_i is the total number of the times the term occurs in training corpus, and N is the number of stories in the corpus. The function is a simplification of the more complex weighting scheme used in InQuery (Ellen et al. 1998); it assumes that all stories are of roughly the same length, and that df_i is not zero. On this basis, the relevance between the story and the topic model can be measured as follows:

$$r(T,D) = \frac{\sum_{i=1}^{n} t_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} t_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}} \qquad (2)$$

where t_i is the weight of term i in the topic model, and it equals to the frequency of the term occurring in training stories. And n is the number of terms in the topic model. This function calculates the cosine similarity between two term vectors. So if two vectors include many overlapping terms and the weights of terms are high, the relevance score between the vectors will be high.

The basic tracking system simply skirt the threshold estimation by using a Detection Error Tradeoff curve (Alvin et al. 1997) to show how false alarm and miss rates vary with respect to each other at various threshold values, and the threshold corresponding to the minimum tradeoff is the optimum one.

2.2 Incomplete Novel Kernel Tracking

A novel event can be defined as a new relevant event to the seminal event of a topic. For example, "*Terrorists attacked the World Trade Center Twin Towers on September 11, 2001*" is the seminal event of topic "*9/11 Terrorist Attacks*", and "*suspect investigation*" is a novel event of the topic. It is imaginable that the static topic model can not absolutely capture the kernels of the on-topic stories which discuss novel events. That is because the model focuses on describing the seminal event which is often the kernel of initial on-topic stories, and it never updates in despite of the fact that the kernel of later relevant stories shifts to novel events. Here, this phenomenon is named Capturing Kernel Attenuation (CKA). The CKA can be illustrated by the degression of the number of overlapping terms between the static topic model and kernels of later on-topic stories. As an example, the CKA of the Topic 40003 in TDT 2002 is shown in Figure 1. In the figure, the horizontal axis denotes the stream of on-topic stories arranged in temporal order; the vertical axis denotes the percent of overlapping terms between the static topic model and the kernel of a on-topic story; each gray dot on the dashed curve denotes the percent of a story; and the thick solid line denotes the approximation to the dashed curve by the method of polynomial smoothing with 5 degree constraint. Although the dashed curve has repeated indentations, it has been on the decrease on the whole. It demonstrates the kernel of later on-topic stories gradually deviate from the initial topic model. Thus it will be difficult for tracking system to distinguish the stories from off-topic ones.

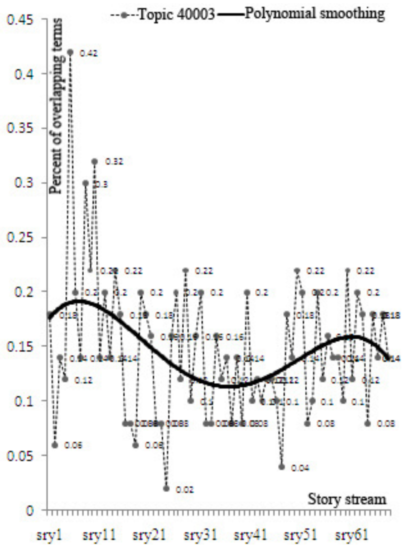


Figure 1. CKA curve of topic 40003 in TDT 2002

It is necessary to clarify the universality of the CKA. So we verified that of all topics except the ones including less than 4 relevant stories in TDT 2002. The results of the verification are shown in Figure 2. It can be found that most topics, which are arranged on the left side of the thick gray line in the figure, have the phenomenon of CKA. But it is different for the topics on the other side whose smoothing curves obviously rise at the late stages. That is because the news stories in the corpus of TDT 2002 only discuss the events happening in a short time (from October 2000 to January 2001). So the corpus doesn't include all the on-topic stories to the long-time topics, thus the topics can not achieve their complete CKA curves. In other words, the curves on the right side of the thick gray line in Figure 2 just denote the partial distributions of CKA of the long-time topics. It can be illustrated by their less number of on-topic stories in the corpus.

An interesting phenomenon is that all of the CKA curves in Figure 2 aren't linear but undulating. So it may be asked why there exist additional peaks, which denote high rate of overlapping terms between the static topic model and the kernels of on-topic stories, since the later stories focus on discussing novel events. That is caused by the characteristic of topical retrospection which means if a story discusses a novel event it will review the old relevant events especially the seminal event. For example, when a story discusses the novel event “Suspect investigation of 9/11 terrorism”, it also reviews the seminal event “Terrorists attacked the World Trade Center”. Topical retrospection is an important manner of news report which provides the backgrounds of novel events for readers. However it isn’t indispensable as illustrated by the troughs of CKA curves in Figure 2. That is because if a novel event happens near to its background events then its stories seldom redundantly discuss the events just as the story predetermines readers have known the backgrounds.

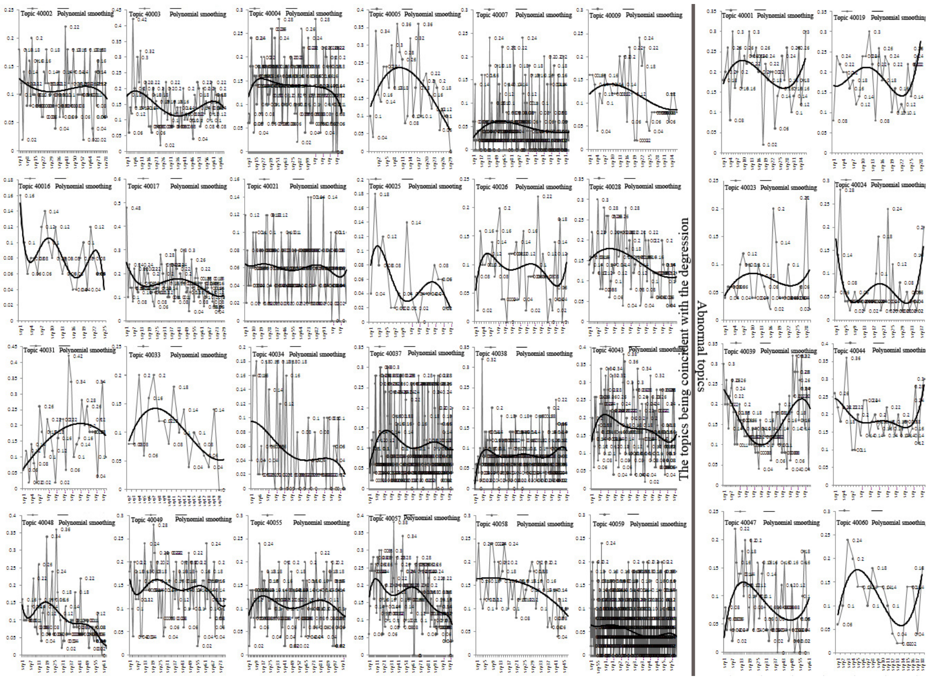


Figure 2. CKA curves of all the topics in the corpus of TDT 2002

In brief, the CKA curves demonstrate the deficiency of the static topic model in capturing kernels of stories discussing novel events. But whether other tracking researches based on static topic model also have the phenomenon of CKA? In next section, we attempt to verify the generalization of CKA in different static topic models.

2.3 CKA of Other Static Topic Models

It is necessary for topic tracking systems to capture the kernels of on-topic stories. So if the systems don’t contain adaptive learning models, they have to rely on the ability of the static topic model mastering the “killer terms” occurred in the kernels of on-topic stories. So it may be helpful to extend the topic model by inserting more terms. However it is inevitable for the extended model to include more redundancies and noises which often cause high relevance score between topic models and off-topic stories. Therefore, it is necessary for static topic models to availably extract the “killer terms”, which have far-reaching effects in whole process of tracking, from the initial training corpus. Thus it is reasonable to dispute whether CKA in section 2.2 is a specific phenomenon caused by the term extraction of the basic tracking system.

In this section, we attempt to demonstrate that CKA of static topic model is a general phenomenon. So we build other six static topic models based on some existing NLP (Natural Language Processing) and statistical strategies, and then we evaluate their distributions of CKA. The models are described as follows: 1) Nominal Static Topic Model (***N-STM***): Terms are weighted by their frequencies in four training stories; and a topic is described as a vector space model (VSM) including 50 most frequent nouns. 2) Verbal Static Topic Model (***V-STM***): It is similar with *N-STM* except that only verbs are used to describe topics. 3) Adjectival Static Topic Model (***A-STM***): It is also similar with *N-STM* except that only adjectives are used to describe topics. 4) Static Topic Model based on Widrow-Hoff algorithm (***WH-STM***) (David et al. 1996): Terms are weighted based on *WH* weight learning algorithm; and a VSM which includes 50 most weighted terms is used to describe topic. 5) Static Topic Model based on Exponentiated-Gradient (EG) algorithm (***EG-STM***) (David et al. 1996): It adopts the same topic model with *WH-STM* except that terms are weighted based on *EG* weight learning algorithm. 6) Static Topic Model based on Rocchio algorithm (***Rocchio-STM***) (Yiming et al. 2000): It also adopts VSM to represent topics, but terms are weighted based on TFIDF scheme; Its initial topic model is constructed by 50 most weighted terms in four training on-topic stories, and it retrieves the four most similar off-topic stories in training corpus based on cosine coefficient between the initial topic model and the stories; And then all the terms in the four training stories are reweighted based on linear weighted Rocchio algorithm which decreases the weights of the terms occurred in the off-topic stories. At last the 50 most weighted terms, which are newly achieved by Rocchio algorithm, are adopted to establish the static topic model.

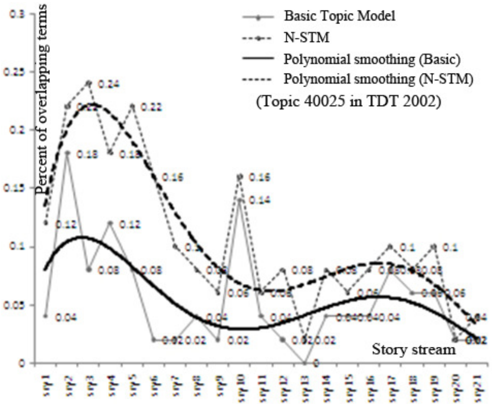


Figure 3. Similar CKA trends (*Basic topic model* vs *N-STM* at topic 40025 in TDT 2002)

We adopt all topics in TDT 2002 (except the ones including less than 4 on-topic stories) to generate the CKA curves of the six static topic models and verify the similarity of the curves with that achieved by the basic topic model mentioned in section 2.2. We find that most of them have very similar distribution trends. For example, Figure 3 shows two CKA curves of topic 40025 achieved respectively by the basic topic model and *N-STM*. It is obvious that the curves both attenuate at the late stage of story stream although *N-STM* achieves more percent of overlapping terms at every point. It is difficult to illustrate all CKA curves of the six static topic models in the limited pages of this paper, so we design a numerical evaluation method of their similar trends, namely Attenuation Approximation Analysis (abbr. A^3). Given a topic, the process of A^3 for two CKA curves includes three steps as follows:

Step 1: For a point dot_i (The point corresponds to a story on horizontal axis) on a CKA curve, an attenuation vector $v_i = \{a_{i,1}, \cdots, a_{i,i-1}\}$ is calculated; each feature $a_{i,j}$ in the vector denotes whether dot_i is higher than dot_j that corresponds to an earlier story on the CKA curve, if yes then $a_{i,j}$ equal to 1 else 0, such as the vector *$v_3(X)$* of *$dot_3(Y)$* in Figure 4;

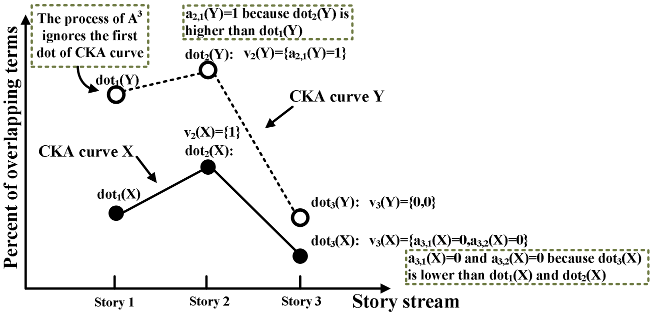

The process of A^3 ignores the first dot of CKA curve
$a_{2,1}(Y)=1$ because $dot_2(Y)$ is higher than $dot_1(Y)$
$dot_2(Y)$: $v_2(Y)=\{a_{2,1}(Y)=1\}$
$dot_1(Y)$
CKA curve Y
$v_2(X)=\{1\}$
$dot_2(X)$:
CKA curve X
$dot_1(X)$
$dot_3(Y)$: $v_3(Y)=\{0,0\}$
$dot_3(X)$: $v_3(X)=\{a_{3,1}(X)=0,a_{3,2}(X)=0\}$
$a_{3,1}(X)=0$ and $a_{3,2}(X)=0$ because $dot_3(X)$ is lower than $dot_1(X)$ and $dot_2(X)$
Percent of overlapping terms
Story 1
Story 2
Story 3
Story stream

Figure 4: An example of *step 1* of A^3

Step 2: It measures the attenuation similarity between each pair of dots corresponding to the same story but respectively on two different CKA curves, such as the similarity between *$dot_2(X)$* and *$dot_2(Y)$* in Figure 4; the similarity is calculated by cosine coefficient between their attenuation vectors;

Step 3: The attenuation approximation score between two CKA curves is calculated by the average score of all the attenuation similarities, such as the score between the CKA curve *X* and *Y* in Figure 4 is calculated as follows (The first dot on CKA curve is ignored in *A^3* process because there isn't any other dot corresponding to earlier story):

$$A^3(X,Y) = \frac{\sum_{i=2}^{3} sim(v_i(X), v_i(Y))}{2}$$

The A^3 score measures the similarity between the attenuation trends of two CKA curves, that is, the higher the score is, the more similar the trends are. The table 1 shows all the results of A^3 evaluation between the CKA curves of the six static topic models and that of the basic topic model mentioned in section 2.1. Given one of the six models, the table records the numbers of the topics whose A^3 scores of CKA curves are within a certain range of numerical quantity. For example, given the model *N-STM*, there are 25 topics in TDT 2002 whose A^3 scores are within the range between 0.7 and 0.8. It is necessary to develop a threshold to determine whether two CKA curves have similar attenuation trends when their A^3 score is higher than it. Here, a threshold of 0.6 is appropriate to the determination by observing. As a comparison the A^3 score between the CKA curves in Figure 3 is 0.7034.

As shown in Table 1 below, the models *N-STM*, *WH-STM*, *EG-STM* and *Rocchio-STM* all achieve the high A^3 scores at most of the topics in TDT 2002. It demonstrates their CKA curves have very similar attenuation trends with that of the basic topic model, so they have the same difficulty in capturing the kernels of the stories discussing novel events. On the contrary, the models *V-STM* and *A-STM* achieve low A^3 scores at most of the topics. It demonstrates their CKA curves are dissimilar with that of the basic topic model. However all of the curves are very close to the horizontal axis at every point, in other words the two models can not effectively capture the kernels of all on-topic stories. In brief, the researches can not substantially improve the performances of topic models capturing kernels of novel events.

A^3 scopes / Models	A^3 *score <0.5*	*0.5<*A^3 *score <0.6*	*0.6<*A^3 *score <0.7*	*0.7<*A^3 *score <0.8*	*0.8<*A^3 *score <0.9*	*0.9<*A^3 *score*
N-STM	**0**	**1(+)**	**6(+)**	**25(+)**	**3(+)**	**3(+)**
V-STM	**9(-)**	**13(-)**	**9(-)**	**2(-)**	**2(-)**	**3(-)**
A-STM	**14(-)**	**14(-)**	**5(-)**	**0**	**2(-)**	**3(-)**
WH-STM	**0**	**0**	**1(±)**	**1(±)**	**20(±)**	**16(±)**
EG-STM	**0**	**11(±)**	**9(±)**	**12(±)**	**4(±)**	**2(±)**
Rocchio-STM	**0**	**0**	**1(±)**	**0**	**15(±)**	**22(±)**

(+): All points on the CKA curve of the evaluated model are higher than that of the basic topic model
(-): All points on the CKA curve of the evaluated model are lower than that of the basic topic model
(±): The CKA curve of the evaluated model tangles with that of the basic topic model

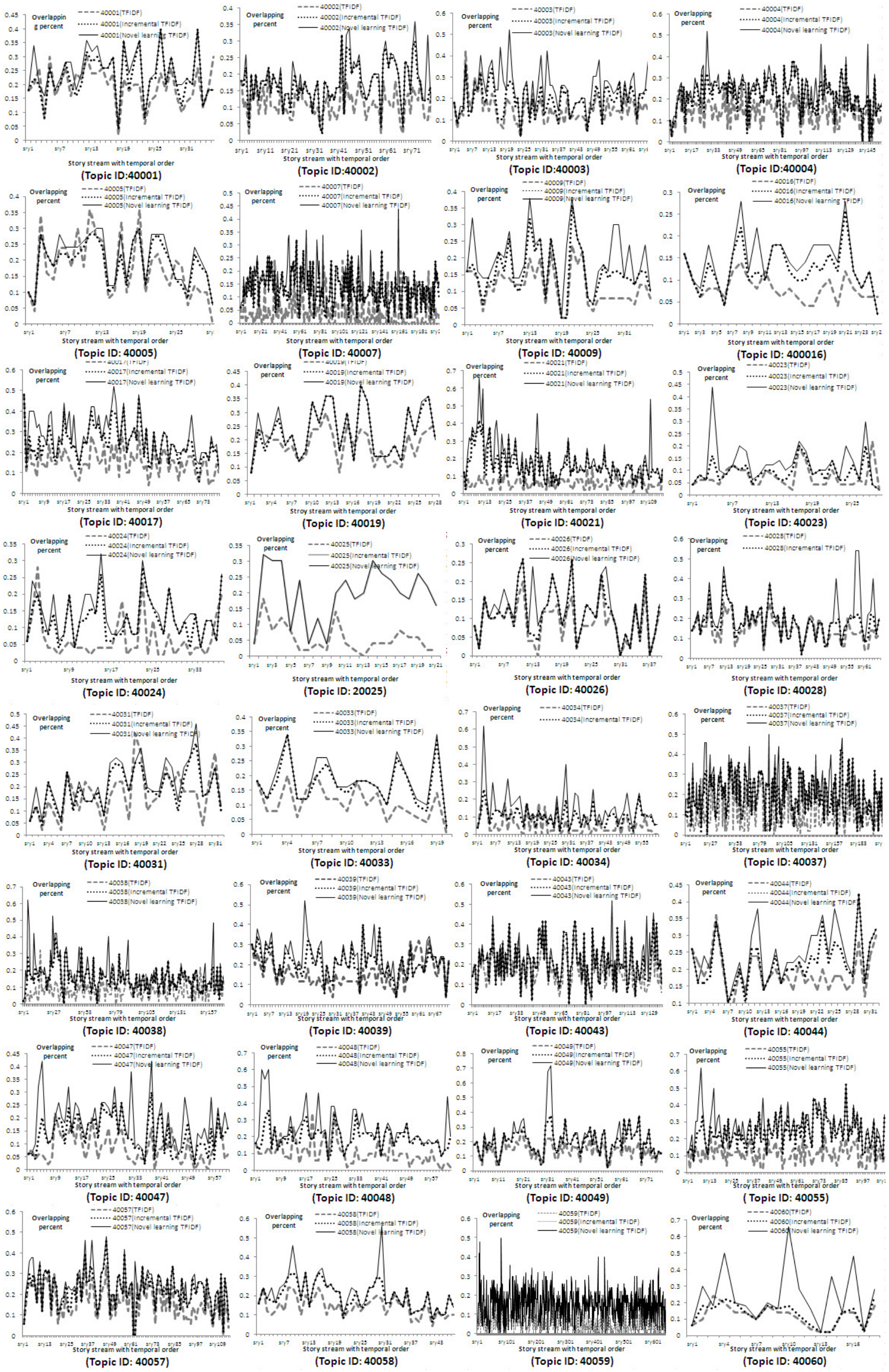
Table 1. Results of the evaluation A^3 for the six static topic models

3 INCREMENTAL NOVELTY LEARNING MODEL

3.1 Incremental Learning Model

The basic tracking system has difficulty in identifying some later on-topic stories because its static topic model always focuses on seminal events but ignores novel ones. A method to solve the problem is to involve adaptive learning models in the process of tracking. Most of the models aim to learn the trend of topic shifting and establish dynamic topic models adapting to the trend. In this section, we adopt an incremental learning model (James, Victor et al. 2002) to improve the basic tracking system and verify whether it is beneficial to capture kernels of stories when they shift from seminal events to novel ones.

The improved tracking system, which involves the incremental learning model, doesn’t adopt the static topic model but the dynamic one. The topic model is initially trained by the similar method with that of the basic tracking system mentioned in section 2.1. But when an on-topic story has been retrieved, the learning model would reweight the terms based on their frequencies in all retrieved relevant stories and the training ones. At the same time the dynamic topic model would be rebuilt by using 50 highest weighted terms. It is obvious that the terms in the topic model will be changed on every time when the learning model operates because of the influence of term distribution in kernels of newly retrieved on-topic stories. Therefore the learning model improves the adaptability of topic model to the trend of topic shifting by continuously involving novel information in it.



(Topic ID:40001)
(Topic ID:40002)
(Topic ID:40003)
(Topic ID:40004)
(Topic ID: 40005)
(Topic ID: 40007)
(Topic ID: 40009)
(Topic ID: 400016)
(Topic ID: 40017)
(Topic ID: 40019)
(Topic ID: 40021)
(Topic ID: 40023)
(Topic ID: 40024)
(Topic ID: 20025)
(Topic ID: 40026)
(Topic ID: 40028)
(Topic ID: 40031)
(Topic ID: 40033)
(Topic ID: 40034)
(Topic ID: 40037)
(Topic ID: 40038)
(Topic ID: 40039)
(Topic ID: 40043)
(Topic ID: 40044)
(Topic ID: 40047)
(Topic ID: 40048)
(Topic ID: 40049)
(Topic ID: 40055)
(Topic ID: 40057)
(Topic ID: 40058)
(Topic ID: 40059)
(Topic ID: 40060)

Figure 5: Improvement of adaptive learning models

It is necessary to illustrate the improvement achieved by the learning model. So we adopt the dynamic topic model to generate CKA curves of the topics in TDT 2002 and compare them with that of the static topic model mentioned in section 2.2. All the curves are exhibited in Figure 5, where the dashed curves are generated by the basic topic model; the dotted ones are generated by the dynamic topic model; and the solid ones are generated by our incremental novelty learning model which will be introduced in section 3.3.

It can be found in Figure 5 that the dynamic topic model can achieve higher percent of overlapping terms with on-topic stories. But as illustration in section 2.3, the static topic model *N-STM*, which is constructed by nouns, has the similar achievement, although its CKA curves have obvious attenuating trends. So we should focus on verifying whether the dynamic topic model improves the ability of capturing kernels of later on-topic stories by incrementally learning novel information. The figure 5 shows that most CKA curves of the dynamic topic model don't have declining tails. Additionally the tails are more far from that of the CKA curves achieved by the static topic model than other segments. It demonstrates the incremental learning model improve the ability of topic model adapting to the trend of topic shifting.

3.2 Burst of News Event

Topic shifting is mainly caused by novel events. Specifically, when on-topic stories focus on discussing novel events, their kernels will shift from the primary centroid of the topic. At this time the kernels involve more terms describing the novel events. Therefore the improved adaptability of the dynamic topic model benefits from its novel terms achieved by the incremental learning model.

However the learning model ignores the burst characteristic of news event. The characteristic often influences the kernel distribution of news story. Specifically, when paroxysmal events happened, most of newly reported stories will focus on discussing them, which often causes substantial increase of the number of story kernels shifting from the primary centroids of topics to that of the paroxysmal events. Thus the novel terms, as particular descriptions of the novel events, will frequently appear in a short period of time. Several examples about the burst of novel terms are shown in Figure 6. In order to indicate clearly, there is only one kind of terms, the time expressions[1], being considered as target terms. The *T* axis in the figure denotes the temporal order of time expressions occurring in a topic and the rightmost time expression on the axis is the earliest one; the *S* axis denotes the temporal order of relevant stories of a topic occurred in news stream, and the rightmost story on the axis is the first story discussed the topic; the *N* axis denotes the amount of time expressions. Thus the same color columns in the three-dimensional space describe how the number of a time expression increases along with on-topic stories being successively reported in news stream, and each column is the number of a time expression in on-topic stories since the expression firstly occurs in the topic.

It can be observed that the frequencies of some time expressions burst forth in a short time since they first appear in novel events (The novel events are labeled by human), such

[1] A unified format is applied to describe the time expressions, e.g. "*20000812*" denotes "*August 12, 2000*". We extract time expression using a rule-based tagger released by TERN organization (http://timex2.mitre.org/taggers/timex2_taggers.himl) which covers many types of time expressions contained in the TIMEX2 2001 guidelines.

as the time expressions corresponding to red columns in Figure 6. In fact, the symbolic terms of novel events, such as person, place, time and activity all have the phenomenon of burst frequency. It is useful for dynamic topic model to involve the terms by learning their burst frequencies. However it is obvious that the incremental learning model, mentioned in section 3.1, ignore the burst frequency of novel term. In detail the learning model weights terms based on their frequencies in all the retrieved on-topic stories in spite of some novel terms only frequently occur in several latest ones. Thus it is difficult for the learning model to identify novel terms because their initial frequencies are always lower than that of the terms existed in topic model over a long period of time, especially the terms of seminal event which always have high frequencies because they nearly occur in all on-topic stories. For example, the frequencies of the time expressions corresponding to seminal events, as indicated by the blue columns in Figure 6, continuously increase along story stream (S axis). Therefore it is necessary to improve the incremental learning model based on the burst frequency of novel term.

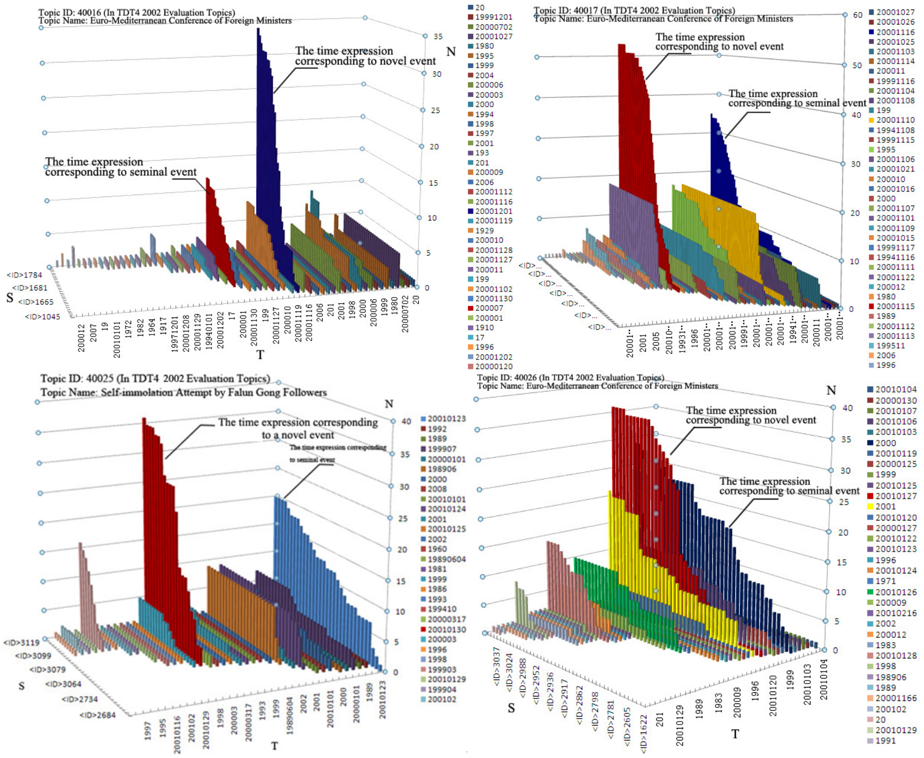


Figure 6: Burst of novel terms (Time expressions)

3.3 Incremental Novelty Learning

We propose an Incremental Novelty Learning model (abbr. INL) to further improve the adaptability of dynamic topic model to the trend of topic shifting. The INL focus on learning the burst frequencies of terms which can be calculated as follows:

$$BF(t) = \frac{tf}{l} \tag{3}$$

where *tf* is the number of times the term *t* occurs in the retrieved on-topic stories, and *l* is the number of the retrieved on-topic stories after the term *t* firstly occurs in the topic. Thus a term which has a low frequency may have a high burst frequency. The Figure 7 shows an example of the calculation of burst frequency, in which the frequency of term *i* in all retrieved on-topic stories is higher than that of term *j*, but the burst frequencies of them are exactly opposite.

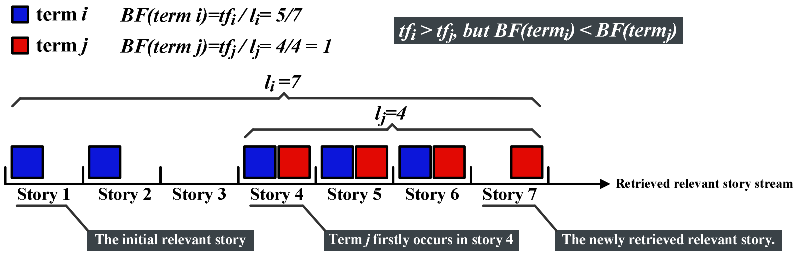

term i
term j
BF(term i)=tfi / li= 5/7
BF(term j)=tfj / lj= 4/4 = 1
tfi > tfj, but BF(termi) < BF(termj)
li =7
lj=4
Story 1
Story 2
Story 3
Story 4
Story 5
Story 6
Story 7
Retrieved relevant story stream
The initial relevant story
Term j firstly occurs in story 4
The newly retrieved relevant story.

Figure 7: An example of burst frequency

On the basis, we build a supervised adaptive topic tracking system whose dynamic topic model holds two lists of terms, viz. L_{inc} and L_{nvl}, weighted respectively by the frequency *TF* and the burst frequency *BF*. When the system has retrieves a new on-topic story, the incremental learning model will insert the novel terms of the story into the list L_{inc} and reweight all terms of the list based on their *TF* in the retrieved on-topic stories (including the newly retrieved one). At the same time the model INL will insert the novel terms into the list L_{nvl} and reweight all terms of the list by their burst frequency *BF*. Then the dynamic topic model extracts 50 highest weighted terms from the list L_{inc} to build a vector V_{inc} which focuses on describing the slow trend of topic shifting, additionally the topic model extracts 50 highest weighted terms from the list L_{nvl} to build a vector V_{nvl} which focuses on describing the burst of topic shifting caused by a novel event. In fact the two vectors have some overlapped terms. So the model INL deletes that from the vector V_{nvl} to achieve a pure description of burst. In the subsequent process of tracking, the determination mechanism measures the relevance between the dynamic topic model and kernels of stories by following formula:

$$R(T,D) = \alpha \cdot r(V_{inc}, D) + \beta \cdot r(V_{nvl}, D) \tag{4}$$

where *D* denotes the kernel of a story which is described by a vector of 50 highest weighted terms in the story, and the weight of a term is calculated by the formula (1); the $r(*,*)$ denotes the formula (2) which calculates the cosine similarity between two term vectors; the parameters α and β are used to adjust the influences of V_{inc} and V_{nvl} to the relevance score *R*(*T*,*D*). Here we simply set the parameters both equal to 0.5 based on hypothesis the vectors V_{inc} and V_{nvl} have the same importance for relevance determination.

It is necessary to illustrate the validity of INL in improving the adaptability of dynamic topic model to the trend of topic shifting. Therefore we generate the CKA curves of the improved topic model to verify its ability of capturing kernels of on-topic stories. The curves are indicated by solid lines in Figure 5. It can be observed that most of the curves achieve higher percent of overlapping terms; especially some segments of them have abrupt increase at the troughs of the dotted CKA curves achieved by incremental learning model. It demonstrates that the terms, which have burst frequencies in a short time, are helpful for dynamic topic model to capture kernels of novel events. Thus INL improve the capacity of the topic model adapting to the trend of topic shifting.

However it is believable that query expansion can be used to increase the number of overlapping terms between topic model and relevant stories. For example if we use 100 highest frequent terms in all retrieved on-topic stories to build dynamic topic model, we may also achieve some terms that have burst frequencies. So why can query expansion improve the ability of topic model capturing kernels of novel events. In fact the answer is in the affirmative. But the query expansion is harmful to topic tracking because it causes the topic model involves much more noises than the helpful terms. Such being the case, we have to question whether the novel terms achieved by INL also include additional noises which cause topic tracking system retrieves more off-topic stories than on-topic ones. We will demonstrate the model INL can improve the performance of tracking system in the experiments in section 4.

4. EXPERIMENTS

4.1 Corpus

We used LDC dataset TDT4 in our experiments. TDT4 contains 98,245 news stories from October, 2000 to January, 2001. The sources of these stories include ABC, CNN, and Associated Press, etc. Their languages include English, Chinese, and Arabic. Both English stories and translation versions (English) of Chinese and Arabic stories are considered. TDT4 also contains stories from NBC and MSNBC TV broadcasts, etc. We also used the transcribed versions of these TV and radio broadcasts besides textual news.

There are 33821 stories in TDT4 are used for the evaluation of TDT hold in 2002 (abbr. TDT 2002). In TDT 2002, there are 3085 stories labeled to be relevant to 40 topics, and 30736 stories are labeled to be irrelevant to the topics. We adopt the corpus of TDT 2002 as training data in our experiments. Additionally there are 18916 stories in TDT4 used for the evaluation of TDT hold in 2003 (abbr. TDT 2003). In TDT 2003, there are 3083 stories labeled to be relevant to other 40 topics, and 15833 stories are labeled to be irrelevant to the topics. We adopt the corpus of TDT 2003 as testing data in our experiments.

4.2 Evaluation

Based on the TDT evaluation guideline, a detection error tradeoff cost (C_{Det}), which combines probabilities of miss and false alarms, is used for our evaluations[2]. The C_{Det} is calculated by the function as follows:

$$C_{Det} = C_{Miss} P_{Miss} P_{t\arg et} + C_{FA} P_{FA} P_{non-t\arg et}\,. \tag{5}$$

[2] http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm. The 2003 TDT task definition and evaluation.

where, P_{Miss} denotes the probability of miss alarms, and P_{FA} denotes that of false alarms. The miss alarms are relevant stories that aren't retrieved by topic tracking system, and false alarms are irrelevant stories that are mistakenly retrieved. The C_{Miss} and C_{FA}, as costs of miss and false alarms, equals to 1.0 and 0.1 respectively. And P_{target} and $P_{non\text{-}target}$, denotes priori target and non-target probability which equals to 0.02 and 0.98 respectively. Additionally the cost C_{Det} can be normalized as follows:

$$Norm(C_{Det}) = \frac{C_{Det}}{\min(C_{Miss} \cdot P_{t\arg et}, C_{FA} \cdot P_{non-t\arg et})} \tag{6}$$

Detection error tradeoff curve[3] (abbr. *DET* curve) in two-dimensional graph is another method to measure the performance of topic tracking system. It's a visualization tool of the trade-off between P_{Miss} and P_{FA}. Each point on the curve corresponds to a pair of P_{Miss} and P_{FA} with a threshold θ. The closer the curve is to the origin of the graph; the better performance the tracking system achieves. The minimum value of *Norm*(C_{Det}) on the curve, namely *Min Norm* (C_{Det}), is the optimal value that a system could achieve at the best threshold θ.

4.3 Main Results

We divide our experiments into four parts. The first part aims to verify whether the learning model INL can farther improve the performance of topic tracking than the incremental learning model. So we compare the performances of following tracking systems:

①Basic tracking system (abbr. ***B-system***). The system adopts a static topic model to describe the kernel of a topic; the model is constructed by a vector of 50 most frequent terms in 4 training stories (Ron et al. 1999). And the kernel of a story is described as a vector of 50 most weighted terms in the story; the weight is calculated by equation (1). The relevance between the topic model and the kernel of a story is measured by their cosine similarity as equation (2). If the similarity is higher than a threshold $\theta_{B-system}$, then the system determines the story to be on topic, else off topic.

②Incremental learning tracking system (abbr. ***I-system***). The system adopts the same initial topic model training, relevance measurement and determination as that of *B-system*. But it is a supervised adaptive tracking system which adopts a dynamic topic model to describe the trend of topic shifting, and modifies the model based on the incremental learning algorithm every time when an on-topic story being retrieved. Here we indicate the threshold by $\theta_{I-system}$.

③Incremental Novelty Learning tracking system (abbr. ***N-system***). The system is similar with *I-system* except that the modification of its dynamic topic model involves the learning algorithm INL as mentioned in section 3.3. Here we indicate the threshold by $\theta_{N-system}$.

A necessary step in the process of relevance measurement is to normalize relevance scores across all topics. The measurement function above results in a ranking of stories where the higher in the rank the more likely a story is to discuss the topic in question.

[3] Topic weighted curves were in use, which were introduced in 1999 evaluation overview paper (http://www.nist.gov/speech/tests/tdt/1999).

However, a score of 0.3 could mean "very likely to match" for one topic and "very unlikely" for another topic. Thus it is difficult to determine relevance across all topics based on uniform thresholds. Our goal is to normalize the scores so that 0.3 (and every other score) has roughly the same meaning no matter what topic and story are being measured. This should result in more "meaningful" scores for stories and more appropriately matches the assumptions behind the DET curve discussed above.

To normalize scores, we calculate the similarity of the topic model against the 4 training stories and find the average of those similarities. That average value is used as a normalization factor, and all scores (for that topic) are divided by it. As a result, although scores can range from zero through well above 1.0, a particularly "good" story (for any topic) should score 1.0 or higher. That is, its un-normalized score will be near those of the 4 training stories, so dividing by that average score, will result in a normalized score near to 1.0. The interpretation of 1.0 as "very like the training stories" is more likely to be true across all topics than before normalization.

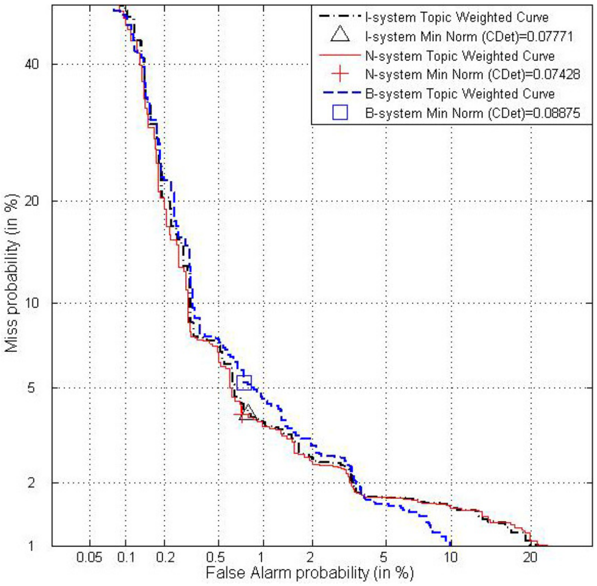
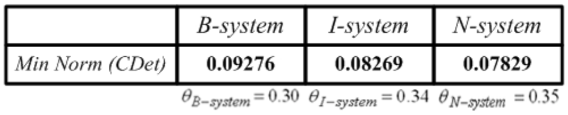


Figure 8: Performance Comparison of the tracking systems (*B-system, I-system and N-system*) in TDT 2002 corpus

We adopt the training corpus (viz. the 40 topics in TDT 2002 along with their labeled on-topic stories and off-topic ones) to generate the DET curves of the three tracking systems (viz. *B-system*, *I-system* and *N-system*). The curves are shown in Figure 8 where the geometrical icons on the curves respectively indicate the minimum-normal DET costs of the systems. It can be observed that the curve of *I-system* is closer to the origin than that of *B-system*, and *I-system* achieves smaller value of *Min Norm (CDet)* than *B-system*. It demonstrates that the incremental learning model improves the performance of tracking by inserting novel terms into the dynamic topic model. In fact the improvement benefits from the lower probability of false alarm. That is because the learning model enhances relevance scores between the topic model and on-topic stories, especially the stories discussing relevant novel events, by which the tracking system can adopt higher threshold to filter more off-topic stories without missing on-topic ones.

The figure 8 also illustrates the farther improvement achieved by our *N-system* whose DET curve is closest to the origin and *Min Norm (CDet)* is smallest. The improvement benefits from the lower miss alarm. In fact the dynamic topic model of *I-system* can not effectively capture the kernels of some on-topic stories which discuss paroxysmal events. It causes so low relevance scores between the stories and the topic model that *I-system* will miss the stories at the time when it increases the threshold to filter off-topic stories. Therefore it is believable that the incremental novelty learning model (INL) in *N-system* improves the ability of the dynamic topic model to describe paroxysmal events by involving the terms burst occurred in a short period of time, by which *N-system* can increase the relevance scores of the stories discussing the paroxysmal events and distinguish the stories from off-topic ones. Thus *N-system* achieves lower miss alarm without influencing false alarm. We verify the performances of the three systems in testing corpus (viz. the 40 topics in TDT 2003 along with their labeled on-topic stories and off-topic ones) as shown in Table 2, where the $\theta_{B-system}$, $\theta_{I-system}$ and $\theta_{N-system}$ respectively indicate the optimal thresholds corresponding to *Min Norm (CDet)* of the systems. We found the *N-system* achieves a relatively stable performance. Its *Min Norm (CDet)* grows less than 0.1 percent in testing corpus (Its minimum normal cost of DET achieved in training corpus is shown in Figure 8).



	B-system	*I-system*	*N-system*
Min Norm (CDet)	**0.09276**	**0.08269**	**0.07829**

$\theta_{B-system} = 0.30$ $\theta_{I-system} = 0.34$ $\theta_{N-system} = 0.35$

Table 2. Performance comparison in testing corpus (*B-system*, *I-system* and *N-system*)

In the second part of our experiments, we verify whether query expansion can take the place of INL to improve the performance of tracking. So we additionally establish a supervised adaptive tracking system, namely ***I-system (QE)***, which involves query expansion in the learning process of *I-system*. Specifically, after *I-system (QE)* retrieves an on-topic story, it reweights all terms in the retrieved stories based on their frequencies and adopts n_t (n_t >50) most weighted terms to generate the expanded topic model. In fact, *I-system (QE)* is very similar with *I-system* except the number of terms in topic model. The process of reweighting terms results in a rank of terms in descending order of frequency. Therefore *I-system (QE)* aims to improve the ability of dynamic topic model adapting to topic evolution by involving more high-frequent terms.

The INL of *N-system* also attempts to improve the adaptability of dynamic topic model by involving some novel terms. But it differs from the query expansion in the aspect of term selection. The INL selects the terms burst occurred in a short period of time in every learning process. The terms often have relatively low frequencies when they initially appear in the retrieved on-topic stories, thus they nearly can not be immediately selected as expansions of dynamic topic model in the learning process of *I-system (QE)*. Therefore, it is difficult for the topic model of *I-system (QE)* to capture some burst events. To demonstrate this conclusion, we compare the performance of *I-system (QE)* with that of *N-system* achieved above.

For a fair comparison, the two systems need to involve the same number of terms in every learning process. Specifically, when the two systems retrieve the same on-topic story given a threshold, the dynamic topic model of *N-system* will be rebuilt by the 50 terms in the vector V_{inc} and n (n<50) terms in the vector V_{nvl} (the two vectors are discussed in section 3.3); the number n is less than 50 because the vector V_{nvl} deletes the terms which have been involved in the vector V_{inc}; on the other side the dynamic topic model of *I-system (QE)* will be rebuilt by n_t (n_t =50+ n) most weighted terms in the list L_{inc} (the list L_{inc} is also discussed in section 3.3). In fact, the dynamic topic model of *I-system* is just rebuilt by the 50 most weighted terms in the list L_{inc} in every learning process, so the topic model is the same with the vector V_{inc} (the 50 terms in the vector V_{inc} are the most weighted ones in the list L_{inc} as mentioned in section 3.3). Thus the main difference between *N-system* and *I-system (QE)* is the dynamic topic model of *N-system* additionally involves n terms burst occurred in a short period of time, on the contrary that of *I-system (QE)* additionally involves n terms frequently occurred in a long period of time. Therefore the comparison between the two systems focuses on verifying how the two groups of terms respectively influence the performance of tracking.

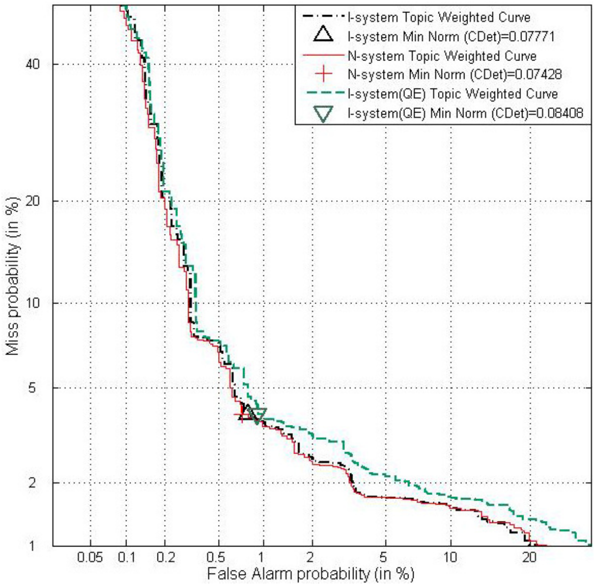
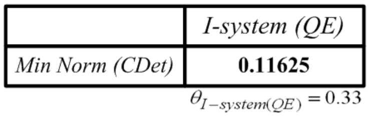


Figure 9: Performance verification of *I-system* after involving Query Expansion

We generate the DET curve of *I-system (QE)* from training corpus and compare it with that of *I-system* and *N-system* as shown in Figure 9. It can be observed that the curve of *I-system (QE)* is much farther from the origin than that of *N-system*. In other words, *N-system* achieves better performance of tracking than *I-system (QE)*. It demonstrates that the n terms supplied by query expansion can not achieve the same effect as that by INL in improving adaptability of dynamic topic model. Additionally the DET curve of *I-system* is closer to the origin than that of *I-system (QE)*. It illustrates the query expansion reduces the performance of *I-system*. The phenomenon may be caused by the noises in the n terms supplied by the query expansion. We also verify the performance of *I-system (QE)* in testing corpus. The *Min Norm (CDet)* of the system is shown in Table 3, where $\theta_{I-system(QE)}$ indicates the optimal threshold. It can be observed that *I-system (QE)* achieves a very unstable performance. Its *Min Norm (CDet)* grows more than 3 percent in testing corpus (Its minimum normal cost of DET achieved in training corpus is shown in Figure 9).

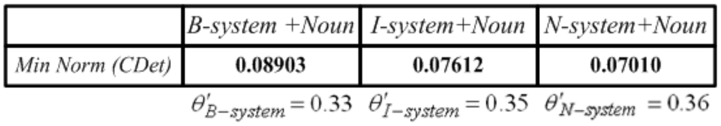


	I-system (QE)
Min Norm (CDet)	**0.11625**

$\theta_{I-system(QE)} = 0.33$

Table 3. The performance of *I-system (QE)* in testing corpus

In the third part of our experiments, we verify the influences of part of speech on the tracking performances of *B-system*, *I-system* and *N-system*. We have demonstrated in section 2.3 that the static topic model, which is constructed by nouns, can achieve higher percent of overlapping terms with most of on-topic stories. Well then, whether can the dynamic topic models in *I-system* and *N-system* also benefit from the virtue of nouns? Therefore we modify the three systems by only adopting nouns to construct their topic models. Additionally the learning models of *I-system* and *N-system* also select nouns to update their dynamic topic models. On the basis, we compare the performances of the three modified systems in testing corpus. Their *Min Norm (CDet)s* are shown in Table 4. It can be observed that all the three modified tracking systems achieve improvements. Especially the *Min Norm (CDet)* of the modified *N-system* decreases approximately 1 percent. That is because nouns include main features of event, such as human, place and time, etc. So nouns can highlight discriminations among different events. Thus the nouns, which are extracted by INL of *N-system*, more substantially improve the ability of the dynamic topic model capturing kernels of novel events when the events suddenly occur.

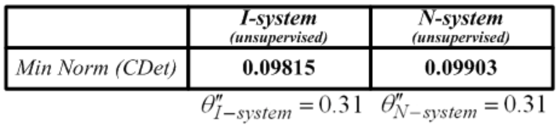


	B-system +Noun	*I-system+Noun*	*N-system+Noun*
Min Norm (CDet)	**0.08903**	**0.07612**	**0.07010**

$\theta'_{B-system} = 0.33$ $\theta'_{I-system} = 0.35$ $\theta'_{N-system} = 0.36$

Table 4. Influence of Nouns on tracking performance

In the last part of our experiments, we verify whether INL can improve the performance of unsupervised adaptive topic tracking. The unsupervised tracking system differs from the supervised one in the aspect of feedback (Martin et al. 2001). Specifically, the learning process of the unsupervised tracking system adopts pseudo-relevant feedback, which not only include on-topic stories but also off-topic ones, to modify the dynamic topic model; on the contrary, that of supervised tracking system adopt relevant feedback, which only include on-topic stories, to modify topic model. Therefore, the *I-system* and *N-system* in the verification perform their learning process under the restriction that pseudo-relevant feedback can be used. We compare the performances of the two modified systems in testing corpus. The *Min Norm (CDet)s* of them are shown in Table 5.



	I-system ***(unsupervised)***	***N-system*** ***(unsupervised)***
Min Norm (CDet)	**0.09815**	**0.09903**

$\theta''_{I-system} = 0.31$ $\theta''_{N-system} = 0.31$

Table 5. Performances of unsupervised *I-system* and *N-system*

It can be observed that the two modified systems achieve worse performance than the *B-system* (The minimum normal cost of DET of *B-system* is shown in Table 2). It demonstrates that dynamic topic model has negative effects on unsupervised adaptive topic tracking. That is caused by noises in pseudo-relevant feedback. Specifically, when an off-topic story in the feedback is used to modify dynamic topic model in learning process, lots of noises (viz. irrelevant terms) will be involved into the model. It results in the updated topic model will mislead the subsequent relevance determination of tracking system. Therefore the challenge in the research of unsupervised adaptive topic tracking is how to automatically shield dynamic topic model from noises.

6. CONCLUSION

In this paper we focus on discussing the features of events, viz. temporal order, novelty and burst, can influence the performance of topic tracking. Therefore we firstly propose a Capturing Kernel Attenuation (CKA) curve to confirm static topic model can not continuously capture kernels of on-topic stories through passage of time. In other word, static topic model can not adapt to topic shifting. And we additionally propose an Attenuation Approximation Analysis (A^3) to prove most current term weighted algorithm can not repair the deficiency of static topic model. Secondly we adopt the CKA curve to verify dynamic topic model has superior adaptability to the trend of topic shifting that caused by novel events. But it has difficulty in tracking the events burst occurred in a short period of time. At last we propose an incremental novelty learning model (INL) to remedy the defect. We compare our method with some existing tracking systems on TDT4 corpus, which demonstrates INL can substantially improve the accuracy of supervised adaptive topic tracking.

However we found INL can not improve the performance of unsupervised adaptive topic tracking. That is because the off-topic stories in pseudo-relevant feedback mislead the learning process. Therefore the key issue is how to automatically shield the dynamic topic model from the noises that caused by the off-topic stories. We found that the noises always suddenly occur and immediately disappear in pseudo-relevant feedback. Thus the learning process INL can adopt the characteristic to filter noises from dynamic topic model in unsupervised adaptive topic tracking. We will implement this work and verify its feasibility in future.

7. REFERENCES

Alvin, Martin, George, Doddington, Terri, Kamm, Mark, Ordowski and Mark, Przybocki, 1997, The Det Curve in Assessment of Detection Task Performance, Proceedings of EuroSpeech'97, volume 4, pp 1895-1898.

David D, Levis, Robert E, Schapire, James P, Callan and Ron Papka, 1996, Training Algorithm for Linear Text Classifiers, Proceedings of the 19th Annual International ACM SIGIR, pp 298-306.

Ellen M,Voorhees and Donna, Harman, 1998, Overview of the sixth text retrieval conference, The Sixth Text Retrieval Conference (TREC-6).

James, Allan, Jaime, Carbonell, George, Doddington, Jonathan, Yamron and Yiming Yang, 1998, Topic Detection and Tracking Pilot Study: Final Report, Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.

James, Allan, Ron, Papka, Victor, Lavrenko, 1998, On-Line New Event Detection and Tracking, Proceedings of SIGIR '98:21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 37–45.

James, Allan, Victor, Lavrenko and Ramesh, Nallapati,2002, UMass at TDT 2002, Proceedings of the Topic Detection and Tracking.

Li, Baoli, Li, Wenjie and Lu Qin, 2006, Enhancing Topic Tracking with Temporal Information, Proceedings of the 29th Annual International ACM SIGIR, pp 667-668.

Margaret, Connell, Ao,Feng, Giridhar Kumaran, Hema, Raghavan, Chirag, Shah and James, Allan, 2004, UMass at TDT 2004, TDT2004 System Description Workshop.

Martin, Franz, J Scott, McCarley, Todd, Ward and Wei-Jing, Zhu, 2001, Unsupervised and Supervised Clustering for Topic Tracking, Proceedings of the 24th Annual International ACM SIGIR, pp 310-317.

Ron, Papka, James, Allan, and Victor, Lavrenko, 1999, UMASS Approaches to Detection and Tracking at TDT2, Broadcast News Workshop'99 Proceedings.
Yiming, Yang, Tom, Ault, Thomas, Pierce and Charles W, Lattimer, 2000, Improving Text Categorization Methods for Event Tracking, Preceedings of the 23th Annual International ACM SIGIR