Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2014, *8*, 155-162

Open Access

Adaptive Topic Tracking Research Based on Title Semantic Domain and Double-state Model

Qi Yincheng[1,*], Zhang Suxiang[2] and Wu Junna[1]

[1]*School of Electrical & Electronic Engineering, North China Electric Power University, Baoding, Hebei, 071003, P.R. China;* [2]*Network Technology Institute Network Management Research Center, Beijing University of Posts and Telecommunications, Beijing, 100876, P.R. China*

Abstract: Aiming at problems of sparse training corpora and topic excursion existing in topic detection and tracking, this paper examined twenty one most recent references and patents, and proposed an adaptive topic tracking strategy based on title semantic domain topic model and double-state model. Title semantic domain topic model can enhance the title-centric semantic domain cohesion of reports and reduce the dimensions of reports' feature space effectively. The double-state strategy is a tracking technology based on the combination of static model and dynamic model: static model uses a given number of training reports to construct the topic model, which is the basis of topic tracking; dynamic model uses the sliding text window mechanism to capture new contents of a topic, remove outdated ones and reflect the changes of topic's focus in a timely manner. Experimental results show that the combination of double-state model tracking strategy and title semantic domain topic model can improve the performance of adaptive topic tracking system.

Keywords: Adaptive topic tracking, Dynamic model, Sliding text window, Static model, Title semantic domain.

1. INTRODUCTION

The main task of topic tracking (TT) is to get customer-satisfied information from a myriad of messages on the Internet. Topic tracking is a sub-evaluation task which belongs to Topic Detection and Tracking (TDT); its specific work is the tracking of follow-up reports of known topics. The known topic is not clearly described; it is given implicitly by certain related reports detected before. Based on this, the TT system judges the correlation of each report and topic in the follow-up reports one by one and collects relevant reports in order to play the tracking function [1]. Topic tracking can bring decentralized information related with a certain topic together, which lets us know the topic completely. The basic idea of a topic tracking system is shown in Fig. (**1**).

Taking a report on flood as an example, the Linguistic Data Consortium (LDC) defines topic correlation as follows: "reports that describe the flood and weather changes which directly influence this flood are all related to the topic; also, disasters caused by this flood, the direct result of disasters and so on are all on-topic" [2]. From this definition we know that from "flooding reports" to "weather changes", "the number of casualties", "rescue work", and then "governments' response", it is a dynamically developing process and the report's focus of Internet news about the topic changes with the same trend. We call this "topic drift", which is one of the challenges for topic tracking study.

Patent US 6,104,989, entitled "Real time detection of topical changes and topic identification *via* likelihood based methods" [3], proposes a method for detecting topical changes and topic identification in texts in real time using likelihood ratio based methods. Topic identification is achieved by evaluating text probabilities under each topic, and a new topic is selected when one of those probabilities becomes significantly larger than the others. Patent US 7,577,654, entitled "Systems and methods for new event detection" [4], provides techniques for new event detection. For a new story and a corpus of stories, adjustments to the importance of terms are determined based on direct or indirect story characteristics associated with each story. Adjustments to the inter-story similarity metrics are determined based on story characteristics and/or a weighting function. New event scores and/or new event categorizations for stories are determined based on the inter-story similarity metrics. Patent US 8,386,240, titled "Domain dictionary creation by detection of new topic words using divergence value comparison" [5], discloses a method to identify topic words in a collection of documents that includes topic documents related to a topic. The candidate topic word is determined to be a topic word if the candidate topic word divergence value is greater than the reference topic word divergence value. The identification of such topics can improve the performance of a language model and/or a system using the language model for languages without boundaries in sentences.

The sparsity of training corpora is another problem to topic tracking study: according to the evaluation requirements, it is necessary to set up a topic model according to Nt (Nt is typically 1, 2, 4) reports, but a few reports neither could grasp the topic from the whole, nor could meet the requirements of topic tracking. To solve these two problems, many researchers point out that we should improve topic tracking's adaptive technology by unsupervised learning: on

*Address correspondence to this author at the School of Electrical & Electronic Engineering, North China Electric Power University, Baoding, Hebei, 071003, P.R. China; Tel: +86 312 7523140; E-mail: qiych@126.com

1874-110X/14 2014 Bentham Open

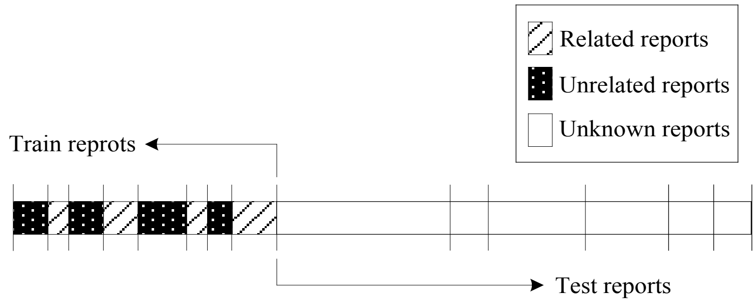

Fig. (1). The basic idea of the topic tracking system.

topic model, vector space model, latent semantic indexing model [6], HMM model, [7] etc. are all applied to the adaptive technology; in strategy, Rocchio method [8] and Pseudo-relevance feedback method [9] train the topic model again by reports relevant or irrelevant with the topic judged by the TT system, which effectively improves the performance of the TT system. A dynamic topic model is proposed based on the static model in reference [10]. It extends the initial topic model with the information from the incoming related stories and filters the noise using the latest unrelated story. Relevant model and irrelevant model are used to judge the current reports and have good effects in noise reduction. In reference [11], several weak topics are combined into a strong topic, which also improves the performance of the TT system. A double-centroid topic model is put forward [12]. It dynamically chooses a division point and the topics are expressed as an initial centroid and a current centroid, and updated with follow-up reports to adapt to the topics' dynamic evolution. A dynamic network model for the evolution of online public opinion is proposed [13], which focuses on the modeling theory of the dynamic network model such as its structure, evolution properties and description method and describes the dynamic process and microstructure of the evolution of online public opinion. Patent US 6,651,058, titled "System and method of automatic discovery of terms in a document that are relevant to a given target topic" [14], provides an automatic mining system to discover terms that are relevant to a given target topic from large databases of unstructured information such as the World Wide Web. The operation is performed in three stages: new terms discovery, candidate terms discovery, and relevant terms discovery. Patent US 7,024,624, titled "Lexicon-based new idea detector" [15], discloses a method and apparatus for detecting the occurrence of new ideas in documents or communications. The method is comprised of three processes: lexicalizing all words or symbols in a set of documents, comparing all words in a second set of documents to the words in the lexicon, measuring the spatial and temporal spread of said fad and computing metrics based on additional occurrences of said fad. Patent US 7,739,261, titled "Identification of topics for online discussions based on language patterns" [16], provides a topic identification system identifying topics of online discussions by iteratively identifying topic words or keywords of the online discussions and identifying language patterns associated with those keywords. The relevance of a sentence is just based on whether the sentence contains a word relating to a topic of the discussion.

The double-judged method of relevance and irrelevance can filter out noise, but can also filter out useful information; topics' combination and the double-centroid model both apply all the related reports to updating model and follow-up reports' judgment; this increases the amount of computation and may lead to accumulating errors. Once the judgment is wrong, erroneous reports may always have an impact on the follow-up judgments. In view of these defects, this paper presents a novel double-state model tracking strategy.

The main idea of the strategy is that static model is used to reserve topic's original contents, and dynamic model is used to capture new contents. They cooperate with each other to complete the task of tracking follow-up reports. In the process, static model plays an important role from start to finish. When dynamic model is updated, on one hand, it uses titles' related factors as auxiliary feedback conditions to improve feedback's accuracy; on the other hand, it removes old feedback reports and grasps the latest development of topic in time. The double-state tracking strategy takes full account of characteristics of topic evolution and makes up for the deficiencies of existing adaptive methods.

Semantic domain is a collection of a set of language structures that their semantics are similar; the feature space that describes semantics is called semantic space [17]. The semantic-based Keyword Extraction Algorithm for Chinese Text (SKE) puts words' semantic features into the process of keywords' extraction, builds words' semantics similarity network and uses intermediate density to measure the criticality of words' semantics [18]. The semantic-based focused crawling approach maps theme's ontology semantics to the list of keywords. Inference services about assertion set expanding and domain-range relation are defined. The semantic relation among keywords can be inferred by inference services [19].

This paper applied title semantic domain to topic tracking study for the first time. The sentence is considered as a unit to segment the news report, considering the similarity between the sentences, with the title as the core content of the report to condense sentence, thereby a large number of nonessential features are removed, and the core content of reports is refined. Experimental results show that the combination of the double-state strategy and title semantic domain achieves a nice topic tracking effect.

2. TITLE SEMANTIC DOMAIN

Semantic domain maintains the consistency of semantics by sentences, and feature words constitute the comprehensiveness of the semantic space. Semantic relation is an important part of natural language processing system. The combination of semantic relation emphasizes that non-adjacent sentences should build relations. These multi-sentence relations can be combined basally by two-connected mode.

Semantic domain is different from subtopic: semantic domain emphasizes on the meaning consistency of sentences; subtopic is a little "incident", it can comprehensively describe this "incident"; it focuses on integrity rather than consistency. By extracting sentences that have the same semantics with titles described by reports, it could agglutinate articles' significance and highlight the effect of center, weakening irrelevant characteristics' impact on topic description and reducing the topic drifts.

3. DOUBLE-STATE MODEL TRACKING STRATEGY

By analyzing the contents of topic tags in massive training corpora, characteristics found are as follows:

1) Some feature items appearing in topic's initial N_t reports will appear in the whole evolution process of topic, shared by different focuses of topic; we call them "relatively static contents" in the development of topic.

2) Some feature items appear in the dynamic process of topic which come from new focuses and development of topic; we call them "dynamic contents" of the topic.

In order to reduce the accumulating errors, in this paper, a fixed-length sliding text detector structure is used to calculate the similarity of feedback reports and current test reports within the window.

Based on the above, the double-state tracking strategy's specific steps are as follows:

(1) Building static model: building topic static model on the basis of N_t reports; these N_t reports are all early reports about topic, many important feature items (entity nouns like person's name, place's name, constitution's name and so on) and even short terms for the topic appearing in the N_t reports. Therefore, the topic model based on N_t reports is not changed or updated in the process of topic tracking; this paper has called it static model (SM).

(2) Judging relevance: judging relevant reports according to the relevant threshold *Tr* and extracting feedback reports from the relevant reports according to feedback conditions.

(3) Building dynamic model: it begins to train the topic dynamic model (DM) when reports meet the feedback conditions. With the increment of feedback reports d_1, feature items of DM are re-counted continuously and their weights are calculated, which update DM continuously.

(4) Adjusting dynamic model: when the number of feedback reports l in the dynamic model reaches the pre-set text window length L, the dynamic model is bi-directionally dealt. As a new feedback report d_0 is added, the earliest report d_L is removed immediately, so L remains unchangeable, $L=\{ d_0, d_1, \ldots, d_i, \ldots, d_{L-1}\}$, $0 \le l \le L-1$, and slides forward along with feedback reports.

Based on cooperation of the double-state model, this method determines the follow-up reports whether they belong to a topic or not. Its advantages are as follows:

(1) It solves, to some extent, the sparsity of training corpora by expanding the training corpora.

(2) The dynamic model can grasp the topic's new contents in time, which could solve the problem of topic drift effectively.

(3) If wrong feedback gets into the dynamic model, the static model plays a balancing role and reduces the accumulated errors.

Specific analysis can be seen in section 6.

4. BUILDING SYSTEM

4.1. Reports Segmenting

In fact, it appears messy if a single feature word is taken as the basic analyzing unit, and at the same time, there are many less important feature words. Then we need to find larger unit in higher level as basic unit of text analysis. Practice has proven that analyzing text in unit of sentence is feasible and efficient.

Segmenting report is the method that segments a complete news report into collections of sentence groups based on sentences and builds report model according to collections of the sentence groups. This paper has taken periods, question marks, exclamation points, etc. as the signs of end of sentences. Then, this paper has segmented training corpora and test corpora into collections of sentences relatively and sorted them by title, forming reports' sentences space model

$$\Psi(D) = \{S_1, S_2, \ldots, S_i\} \tag{1}$$

Where, D denotes a certain report; S denotes every sentence in D; i denotes the number of sentences in D, and i is different as the reports are different.

In this step, the title does not need to be segmented by punctuations, but it must be the first sentence of reports' sentence space model.

The current ceramic enterprises in ceramic product design process and production process of detailed study, the design process of ceramic products for further analysis and decomposition. Exploring the project design of ceramic needs human interaction steps in the 3d CAD system, and needs to be done by the system automatically in order to better improve enterprise's key steps in the efficiency of product design and in decomposition on the basis of the design process to classify modeling of ceramic products and finished components decomposition of the complex products. Ceramic products decomposition is different from the mechanical parts and components industry and other industry products; ceramic products decomposition lies mainly in the design of the components in the process of decomposition, and the final product in general is not an integral whole; for example, the design process of more complex products such as “pot” is broken down into parts such as the pot body, the pot, a spout, the lid and lid knob, a girder of the pot and an ear piece [10]. And all kinds of parts are set up in the three-dimensional model of the material library.

4.2. Sentences Modeling

According to the theory of anthropological linguistics, language sentence consists of key ingredients like subject, predicate and object, also attribute, adverbial modifier and complement. However, in some cases, some adjectives in reports such as customary words, phrases in news can also play iconic role of distinguishing reports. So, when sentences are segmented and similarity is calculated, all nouns, pronouns, verbs and adjectives of a sentence are taken as keywords and are considered in calculation only, and unused words like adverbs, conjunctions, modal particles are removed.

Sentence processing consists of segmentation and removing unused words. After processing, sentences are expressed as the collection of feature items:

$$L(S) = \{t_1, t_2, ..., t_n\} \tag{2}$$

Where, t is the feature item in sentences; n denotes the number of words in processed sentences S, namely the number of feature items.

Then we can build feature vector space of every sentence:

$$V(S) = \{t_1, w_1; t_2\ w_2; ...; t_n, w_n\} \tag{3}$$

Where, w_n denotes the weight of every feature item in sentence's collection and it can be calculated according to the proportion of a word's frequency in all words' frequencies.

4.3. Semantic Domain's Cohesion

The concrete steps of semantic domain cohesion strategy on a to-be-texted report D are as follows:

(1) Calculating the similarity $P(s_i,s_j)$ of all sentence pairs in $S\{(s_i, s_j)|s_i, s_j \in S\}$ according to the language model:

$$P(S_i, S_j) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{n} w^2{}_{ik}) \times (\sum_{k=1}^{n} w^2{}_{jk})}} \tag{4}$$

Where, s_i, s_j are sentences that their similarities are to be calculated; w is the weight of feature item; n is the dimension consisting of all features of two sentences.

(2) Taking every report's title as cohesion kernel sentence, and taking report's sentences that their correlation with the sentence is above the given threshold value T_h as candidate semantic domain and embedding them into the collection $L=\{s_1, s_2, \ldots, s_k\}$, k is the number of sentences. The collection is report's semantic domain. Then we can calculate the weight of every feature item of the collection and form the semantic domain feature vector space of reports.

Semantic domain cohesion strategy is an important part of building topic semantic domain model. There are two advantages of this treatment: the first is simplifying reports, removing a large number of irrelevant sentences and redundant feature words; the second is condensing the core, taking key contents from the report according to the important position of title in news report, highlighting the report's main content and providing a simple and effective model for following topic tracking.

4.4. Static Topic Model Building

We built the static topic model according to the semantic domain feature vector space of N_t pieces of training corpora. The training report d_i is treated as a normalized feature eigenvector: $V(d_i)=(term_1, w_1(d_i); \ldots; term_k, w_k(d_i); \ldots; term_n, w_n(d_i))$. $w_k(d_i)$ is the weight of $term_k$ in d_i, its abbreviation is w_{ik}. This paper used TF-IDF (term frequency-inverse document frequency) to calculate the weight of every feature item:

$$w_{ik} = \frac{tf_{ki} \times \lg(N / n_k + 0.01)}{\sqrt{\sum_{term_k \in d_i} [tf_{ki} \times \lg(N / n_k + 0.01)]^2}} \tag{5}$$

Where N is the number of verified news reports, n_k is the number of news reports that have feature item $term_k$ in verified news reports; tf_{ki} is the frequency of feature item $term_k$ in news reports d_i.

4.5. System Dynamic Topic Model Building

In the double-state model tracking strategy, two thresholds are set: relevant threshold (T_r) is used to judge whether the following reports are related with the topic or not; feedback threshold (T_f) is used to judge whether the topic should update the dynamic model or not. Usually, we take $T_f > T_r$. The specific procedure is shown in Fig. (**2**).

Selection rules for threshold value T_f and T_r are:

(1) The value of T_f and T_r is concerned with the performance index of the system and the performance index precision (P) and recall (R) are contradictory. In the process of determining the value of T_f and T_r, consideration should be given to both precision and recall. It should be tried to make the two performance indexes in balance that not only consider the higher precision or recall.

(2) Generally speaking, T_r controls the preliminary selection of the report text and T_f is used to choose the text of the feedback report carefully. Therefore, we demand T_f as greater than T_r.

Actual values of T_r and T_f are shown in section 6.3, and the analysis of feedback condition is shown in section 4.7.

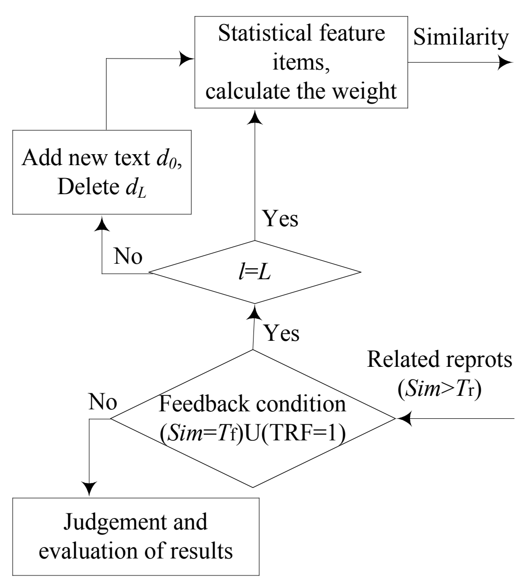

Fig. (2). Flowchart of dynamic topic model.

4.6. Process of Dynamic Tracking

This paper built the double-state model tracking strategy by combining the static model and the dynamic model and finished the tracking task of the topic adaptive in the strategy.

The calculation of similarity between the topic reports can use the cosine formula, as:

$$sim(m_i, m_j) = \frac{\sum_{k=1}^{M} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{M} w^2{}_{ik}) \times (\sum_{k=1}^{M} w^2{}_{jk})}} \quad (6)$$

Where, m_i is the feature vector of the to be tested news reports d_i, m_j is the feature vector of the *j*th tested news report d_j, M is the dimension of feature vector, w_{ik} and w_{jk} are the kth dimensions of the feature vector in news reports d_i and $d_{i,}$ respectively.

The dynamic model can not only grasp the development of topic in time, but it can also introduce wrong feedbacks. To reduce the impact of possible wrong feedbacks, the strategy endows the static model and the dynamic model with equal weights. The calculating method of weight is as follows:

(1) Prior to the establishment of the dynamic model, the number of feedback reports l is zero, *i.e.* l=0. The calculation of the comprehensive similarity is:

$$Sim = 1.0 \times Sim_{SM} + 0 \times Sim_{DM} \quad (7)$$

Where, Sim is the comprehensive similarity value; Sim_{SM} is the similarity between the current report and the static model; Sim_{DM} is the similarity between the current report and the dynamic model.

(2) In the process of the establishment of the dynamic model, the number of feedback reports l is less than the window length L, *i.e.* l<L. The calculation of the comprehensive similarity is:

$$Sim = (1 - \frac{l}{N_t + L}) Sim_{SM} + \frac{l}{N_t + L} Sim_{DM} \quad (8)$$

The window length L is determined by the experiments.

(3) After the dynamic model has been established, the number of feedback reports l equals to the window length L, *i.e.* l=L. The calculation of the comprehensive similarity is:

$$Sim = \frac{N_t}{N_t + L} Sim_{SM} + \frac{L}{N_t + L} Sim_{DM} \quad (9)$$

4.7. Feedback Conditions

In existing studies, the feedback condition may be the number of related documents [8]; it may also be the uncertainty of the samples [2] or some characteristics' selection rules [7]. In this paper, feedback conditions are set by feedback's threshold T_f.

It is hard to avoid misjudgment if feedback threshold is only relied on to choose the feedback reports. Adding some judgment rules appropriately can improve the accuracy. Internet news title usually contains some important information related with the topic; it plays a very important role in topic detection. So this paper used Title Related Factor (TRF) to set the feedback rules:

$$TRF = m \times | title_{model} \cap title_i | \quad (10)$$

Where, $title_{model}$ is the collection of title feature items of the double-state model; $title_i$ is the collection of report D_i's title feature items; |.| is the number of feature items; m is the adjustment factor.

One topic has different reports and the contents of titles are also different. So it is hard to ensure TRF's specific value. This paper has taken "Diaoyu Island Event" as an example to conduct the experiment. This topic model feature space has 34 features and 263 titles to be tested. We count the number of 34 topic model features appearing in 263 titles. Statistical result is shown in Fig. (**3**).

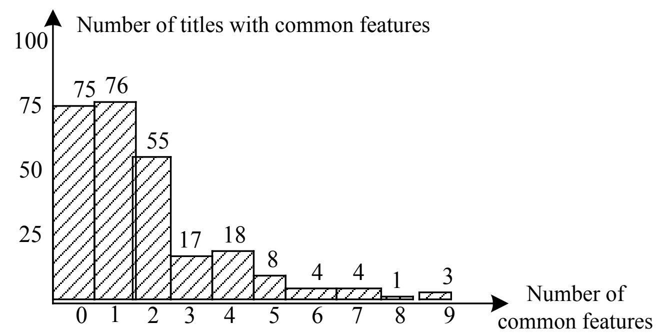

Fig. (3). Titles total number of random graph characteristics.

From the results, we can see that the number of common features shared by feature space of the title in the topic model and the title of the follow up reports has a peak at 0,1and 2. So, we set feedback condition as TRF≥1.

In summary, only when the tested report meets the requirement of (Sim≥Tf) ∪(TRF≥1), then can it be considered as feedback report and the dynamic model can be updated. When grasping new topic contents, it could reduce the accumulation of errors effectively.

5. SYSTEM FRAMEWORK

According to earlier parts' description, the flow chart of proposed adaptive topic tracking system is shown in Fig. (**4**).

The following factors are taken into consideration in calculating similarity between the topic reports. Static content of topic concerns the semantic domain feature vector space of Nt pieces of training corpora, dynamic contents coming from new focuses and development of the topic, and the tracking strategy by combining the static and the dynamic model.

6. EXPERIMENT AND RESULTS ANALYSIS

6.1. Evaluation Mechanism

This paper made use of traditional precision (P), recall (R) and overall classification rate (F_a) to evaluate the experimental results; the formula is as follows:

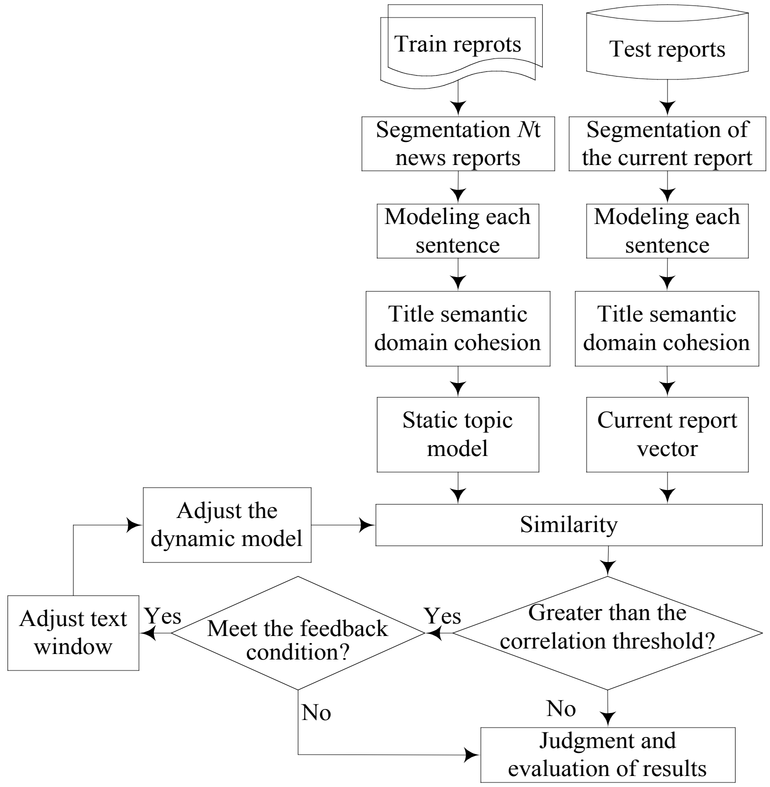

Fig. (4). The flow chart of adaptive topic tracking system.

(11)

(12)

Where, x is the number of reports that belong to the type actually and the system judges them according to the type; y is the number of reports that do not belong to the type actually but the system judges them according to the type; z is the number of reports that belong to the type actually but the system judges them not according to the type.

$$F_{\alpha} = \frac{(\alpha^2 + 1)PR}{\alpha^2 P + R} \tag{13}$$

Where, α is an adjusting parameter, $\alpha \geq 0$. It is used to balance the precision and recall in different weights. As α equals to 0, $F_{\alpha}=F_0=P$, the precision and the recall are equally important. As α approaches $+\infty$, $F_{\alpha} \to F_{+\infty} \to R$. As $\alpha=1$, $F_{\alpha} = F_1 = 2.P.R/(P+R)$. As $\alpha<1$, F_{α} puts emphasis on precision. As $\alpha > 1$, F_{α} puts emphasis on recall. This paper used the most common F_1 to evaluate the experimental results.

6.2. Experiment Design

This study reported several events that happened in the past two years from the Internet, obtaining 3600 news reports totally. Among them, 8 sample events and 1608 reports were chosen. The standard of choosing sample event is the report that has high attention, lasts a period of time and is necessary to be reported. Each topic has 50-200 counter-example reports and more than 800 counter-example reports in all topics. One topic's counter-example reports contain many other topics in order to ensure the generalization of counter-examples. Table **1** shows the sample events and the number of related reports.

The specific design of experiments is described as follows:

Experiment 1: measuring the influence of the slide text window's length L in the dynamic model on the tracking effect.

Experiment 2: using the basic topic tracking model built by experiment 1 [20] as the Baseline system. The system judges whether the following reports are related with the topic or not according to the relevant threshold T_r; it doesn't use self-learning technology. We use Baseline-Track to denote the system and the experimental results. In order to analyze the adaptive technology's influence on the performance of the topic tracking system, the double-state topic's adaptive method is compared with the traditional increment's adaptive method [21]. In the topic tracking system of traditional increment's adaptive method, we use two thresholds (T_r and T_f) to judge whether the following reports are related to the topic or not; reports which have similarity above T_f are all used to update the old topic model which is denoted by Increment-Track. In addition, the double-state tracking strategy and title semantic domain get experimental verification respectively: Doublestate-ATT is the topic tracking system of the double-state adaptive tracking strategy; this system does not use the title semantic domain technology to deal with the training corpus and the test corpus. It uses Title-domain-Track to stand for the TT system of the title semantic domain language model, i.e., it uses the title semantic domain space model to replace the TT system of the traditional vector space model in experiment 1 [20]. In the end, the semantic domain language model is combined with the double-state tracking strategy and the double-state strategy ATT system is put forward based on the title semantic domain, which is denoted by Title-Doublestate-ATT.

Table 1. Sample events.

Topic Number	Topic Description	Number of Related Reports
1	The US mid-term elections	217
2	Indonesia tsunami	148
3	Somali pirates event	31
4	President visit countries	29
5	Libyan military attacks	296
6	Bank exchange rate adjustment	244
7	Inter-Korean conflict	505
8	Diaoyu island incident	134
Total		1608

6.3. Experimental Results Analysis

In all experiments, N_t = 4, and the focuses of the study are the semantic domain topic model and the double-state tracking strategy. Firstly, to determine the best value of L, double-state topic tracking experiment is conducted with different values of L; the results are shown in Table **2**.

Table 2. Results of experiment 1 (T_f =0.44, T_r =0.3).

L	P (%)	R (%)	F_1 (%)
2	71.94	91.21	80.44
3	72.07	92.74	81.11
4	71.79	96.52	82.34
5	67.58	95.52	79.16
6	65.99	96.02	78.22

From the results of experiment 1 in Table **2,** we know: the precision and recall of double-state topic model's ATT system increase gradually when the length of the text window increases gradually from 2 and the value of comprehensive measure F_1 becomes better and better. But when the length is larger than 4, the precision and recall both begin to decrease and the value of F_1 also becomes worse. Experimental results show that good tracking effect can be achieved if the static model's text length gets close to the dynamic model's text length. So, L is selected as 4.

The results of experiment 2 are shown in Table **3**.

We can get such conclusions from experiment 2:

(1) Comparing with Baseline-Track, Increment-Track's recall is more than 99%, but the accuracy is only 54.86% and

Table 3. Results of experiment 2 (T_f=0.44, T_r=0.3/0.35/0.4, N_t=L=4).

Experimental System	P (%)	R (%)	F_1 (%)
Baseline-Track	76.48	87.89	81.78
Increment-Track	54.86	99.31	70.68
Doublestate-ATT	79.21	89.87	84.20
Title-Domain-Track	89.95	90.41	90.18
Title-Doublestate-ATT	90.91	94.81	92.82

F_1 decreases by 11.1%. By analyzing the testing results of the experimental corpus, we find a large number of noise data in the topic model: in the topic of "Indonesia's tsunami event", reports of topic model rise from 4 to 246, containing 98 misjudged irrelevant reports; in the topic of "Inter-Korean conflict ", the number rises from 4 to 665, containing 160 misjudged irrelevant reports. So, Increment-Track has high recall and low precision. The results tell us that the incremental topic's feedback [21] brings many irrelevant reports and affects the system's performance compared with the TT system in [20] for experiment 1; with topic tracking, a lot of relevant reports are put into the topic model, the topic model becomes enormous and the testing efficiency is low, which largely limits the application of incremental topic tracking model in practical.

(2) Comparing with Baseline-Track, Doublestate-ATT's precision increases by 2.73%, the recall increases by 1.98%, and F_1 increases by 2.42%. The results show that comparing with the Baseline-Track system in experiment 1, the double-state topic tracking model can improve the performance of the adaptive topic tracking system. The experiment was conducted many times and the thresholds were different, and the precision and recall both improved. Among the three systems, Doublestate-ATT has the best effect, which indicates that the double-state tracking strategy could get a good effect in topic tracking.

(3) Comparing with Baseline-Track, Title-Domain-Track's precision and recall both have substantial improvement: the precision increases by 13.74%, the recall increases by 2.52%, and F_1 increases by 8.4%. By analyzing every topic model feature space, feature dimensions of Title-Domain-Track decrease significantly. Taking "Indonesia's tsunami event" for example, the topic model's dimensions of Baseline-Track are 408, but Title-Domain-Track's are only 149. Features reserved can well reflect important contents of the topics. Comparing with the traditional TT system used in experiment 1 [20], topic title semantic domain model has the advantage of better expressing the topic and improving the performance of adaptive topic tracking.

(4) Comparing with Title-Domain-Track, Title-Doublestate-ATT's precision increases by 0.96%, the recall increases by 4.40% and F_1 increases by 2.64%. The double-state strategy is combined with semantic domain's topic model; they play a role together and improve ATT's performance. By analyzing the experimental data, we can find that the feature dimensions of topic model are few, but the features are all relevant and important; in collections of tracked

news reports, irrelevant reports are few and dispersed in time and contents with no serial wrong judgment reports. This tells us that the double-state tracking strategy can deal with topic drift effectively, even if wrong feedbacks get into dynamic model, while the static model inhibits the following reports' similarity and decreases the frequency of misjudgment. The fixed-length sliding window mechanism of dynamic model can also remove earlier feedback reports in time, reducing errors' accumulation and realizing dynamic tracking. Experimental results show that Title-Doublestate-ATT system proposed in this paper is the best. It can substantially improve the performance of the ATT system.

CONCLUSION

According to the problems of sparse training data and topic drift existing in adaptive topic tracking, this paper put forward the adaptive tracking algorithm based on the combination of title semantic domain and double-state model. This algorithm takes topic's two features into consideration: some topic contents exist throughout the topic's whole evolution, which are topic's static contents; some are new contents, which are topic's dynamic contents. In double-state strategy, the topic model leaves topic's static contents, grasps topic's changes in time and also uses sliding text window mechanism to prevent errors' accumulation. The title semantic domain can effectively reduce feature dimensions of the reports vector space. The experimental results show that either the double-state tracking strategy or the title semantic domain or even the combination of them can improve the performance of adaptive topic tracking system.

Adaptive topic tracking technology is an interdisciplinary key research of natural language processing, data mining and intelligent information processing. Moreover, it is an important means to provide convenient access to information in real life and has broad prospects.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This research has been partially supported by Grants from Fund of National Natural Science Foundation of China (Grant No: 61202248).

REFERENCES

[1] Y. Hong, Y. Zhang, T. Liu, and S. Li, "Topic detection and tracking review," *Journal of Chinese Information Processing*, vol. 21, pp. 71-87, 2007.

[2] H. Z. Wang, X. J. Zhang, J. B. Zhu and B. Zhang, "Adaptive topic tracking based on active learning," In: *Frontiers of Chinese information development-Chinese Information Processing Society of China's 25th Anniversary Academic Conference Proceedings,* pp. 373-382, 2006.

[3] D. Kanevsky, and E. Yashchin, "Real time detection of topical changes and topic identification via likelihood based methods", U.S. Patent 6,104,989, 2000.

[4] T. H. Brants, F. R. Chen, and A. O. Farahat, "Systems and methods for new event detection", U.S. Patent 7,577,654, 2009.

[5] J. Wu, T. X. Liu, F. Hong, Y. G. Wang, B. Yang, and L. Zhang, "Domain dictionary creation by detection of new topic words using divergence value comparison", U.S. Patent 8,386,240, 2013.

[6] X. F. Zhang, Z. G. Guo, and B. C. Li, "An Effective Algorithm of News Topic Tracking," In: *WRI Global Congress on Intelligent Systems*, pp. 510-513, 2009.

[7] J. P. Zeng, and S. Y. Zhang, "Incorporating topic transition in topic detection and tracking algorithms," *Expert Systems with Application*, vol. 36, pp. 227-232, 2009.

[8] G. A. Thomas, and Y. M. Yang, "Information filtering in TREC-9 and TDT-3: A comparative analysis," *Information Retrieval*, vol. 5, pp. 159-187, 2002.

[9] V. R. Shanks, and H. E. Williams. "TDT2001 topic tracking at RMIT university," In: *Proceedings of the Topic Detection and Tracking (TDT) Workshop*, 2001.

[10] X. Y. Zhang, and T. Wang, "Topic tracking with dynamic topic model and topic-based weighting method," *Journal of Software*, vol. 5, pp. 482-489, 2010.

[11] H. Z. Wang, J. B. Zhu, D. Ji, N. Ye, and B. Zhang, "Adaptive chinese topic tracking based on feedback learning," *Journal of Chinese Information Processings*, vol. 20, pp. 92-98, 2006.

[12] H. Zhao, T. J. Zhao, H. Yu, and S. Zhang, "Dynamic evolvement-oriented topic detection research," *High Technology Letters*, vol. 16, pp. 1230-1235, 2006.

[13] H. M. Zhu, X. N. Su, and X. B. Zhang, "The dynamic network model of Internet public opinions' evolution research," *Information Studies: Theory & Application*, vol. 33, pp. 75-78, 2010.

[14] N. Sundaresan, and J. Yi, "System and method of automatic discovery of terms in a document that are relevant to a given target topic", U.S. Patent 6,651,058, 2003.

[15] K. J. Hintz, "Lexicon-based new idea detector", U.S. Patent 7,024,624, 2006.

[16] H. J. Zeng, H. Li, J. Hu, Z. Chen, D. Zhang, and J. Wang, "Identification of topics for online discussions based on language patterns", U.S. Patent 7,739,261, 2010.

[17] Y. Hong, Y. Zhang, J. L. Fan T. Liu, and S. Li. "Chinese topic Link Detection Based on Semantic Domain Language Model," *Journal of Software,* vol. 19, pp. 2265-2275, 2008.

[18] L. X. Wang, and X. Y. Huai, "Semantic-based keyword extraction algorithm for Chinese text," *Computer Engineering*, vol. 38, pp. 1-4, 2012.

[19] Y. X. Ye, and D. T. OuYang, "Semantic-Based Focused Crawling Approach," *Journal of Software*, vol. 22, pp. 2075-2088, 2011.

[20] X. F. Xue, Y. K. Zhang, and X. D. Ren, "Methods research of New event detection based on news element," *Computer Applications*, vol. 28, pp. 2975-2977, 2008.

[21] J. P. Yamron, S. Knecht, and P. V. Mulbregt, "Dragon's Tracking and Detection Systems for the TDT2000 Evaluation," In: *Proceedings of Topic Detection and Tracking Workshop*, pp. 75-79, 2000.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014