Story Link Detection based on Dynamic Information Extending

Xiaoyan Zhang Ting Wang Huowang Chen
Department of Computer Science and Technology, School of Computer,
National University of Defense Technology
No.137, Yanwachi Street, Changsha, Hunan 410073, P.R.China
{zhangxiaoyan, tingwang, hwchen}@nudt.edu.cn

Abstract

Topic Detection and Tracking refers to automatic techniques for locating topically related materials in streams of data. As the core technology of it, story link detection is to determine whether two stories are about the same topic. To overcome the limitation of the story length and the topic dynamic evolution problem in data streams, this paper presents a method of applying dynamic information extending to improve the performance of link detection. The proposed method uses previous latest related story to extend current processing story, generates new dynamic models for computing the similarity between the current two stories. The work is evaluated on the TDT4 Chinese corpus, and the experimental results indicate that story link detection using this method can make much better performance on all evaluation metrics.

1 Introduction

Topic Detection and Tracking (TDT) (Allan, 2002) refers to a variety of automatic techniques for discovering and threading together topically related material in streams of data such as newswire or broadcast news. Such automatic discovering and threading could be quite valuable in many applications where people need timely and efficient access to large quantities of information. Supported by such technology, users could be alerted with new events and new information about known events. By examining one or two stories, users define the topic described in them. Then with TDT technologies they could go to a large archive, find all the stories about this topic, and learn how it evolved.

Story link detection, as the core technology defined in TDT, is a task of determining whether two stories are about the same topic, or topically linked. In TDT, a topic is defined as "something that happens at some specific time and place" (Allan, 2002). Link detection is considered as the basis of other event-based TDT tasks, such as topic tracking, topic detection, and first story detection. Since story link detection focuses on the streams of news stories, it has its specific characteristic compared with the traditional Information Retrieval (IR) or Text Classification task: new topics usually come forth frequently during the procedure of the task, but nothing about them is known in advance.

The paper is organized as follows: Section 2 describes the procedure of story link detection; Section 3 introduces the related work in story link detection; Section 4 explains a baseline method which will be compared with the proposed dynamic method in Section 5; the experiment results and analysis are given in Section 6; finally, Section 7 concludes the paper.

2 Problem Definition

In the task definition of story link detection (NIST, 2003), a link detection system is given a sequence of time-ordered news source files $S = \langle S_1, S_2, S_3, \ldots, S_n \rangle$ where each S_i includes a set of stories, and a sequence of time-ordered story pairs $P = \langle P_1, P_2, P_3, \ldots, P_m \rangle$ where $P_i =$

$(s_{i1}, s_{i2}), s_{i1} \in S_j, s_{i2} \in S_k, 1 \leq i \leq m, 1 \leq j \leq k \leq n$. The system is required to make decisions on all story pairs to judge if they describe a same topic.

We formalize the procedure for processing a pair of stories as follows:

For a story pair $P_i = (s_{i1}, s_{i2})$:

1. Get background corpus B_i of P_i. According to the supposed application situation and the custom that people usually look ahead when they browse something, in TDT research the system is usually allowed to look ahead N (usually 10) source files when deciding whether the current pair is linked. So $B_i = \{S_1, S_2, S_3, \ldots, S_l\}$, where
$$l = \begin{cases} k+10 & , s_{i2} \in S_k \text{ and } (k+10) \leq n \\ n & , s_{i2} \in S_k \text{ and } (k+10) > n \end{cases}.$$

2. Produce the representation models (M_{i1}, M_{i2}) for two stories in P_i. $M = \{(f_s, w_s) \mid s \geq 1\}$, where f_s is a feature extracted from a story and w_s is the weight of the feature in the story. They are computed with some parameters counted from current story and the background.

3. Choose a similarity function F and computing the similarity between two models. If t is a predefined threshold and $F(M_{i1}, M_{i2}) \geq t$, then stories in P_i are topically linked.

3 Related Work

A number of works has been developed on story link detection. It can be classified into two categories: vector-based methods and probabilistic-based methods.

The vector space model is widely used in IR and Text Classification research. Cosine similarity between document vectors with $tf*idf$ term weighting (Connell et al., 2004) (Chen et al., 2004) (Allan et al., 2003) is also one of the best technologies for link detection. We have examined a number of similarity measures in story link detection, including cosine, Hellinger and Tanimoto, and found that cosine similarity produced outstanding results. Furthermore, (Allan et al., 2000) also confirms this conclusion among cosine, weighted sum, language modeling and Kullback-Leibler divergence in its story link detection research.

Probabilistic-based method has been proven to be very effective in several IR applications. One of its attractive features is that it is firmly rooted in the theory of probability, thereby allowing the researcher to explore more sophisticated models guided by the theoretical framework. (Nallapati and Allan, 2002) (Lavrenko et al., 2002) (Nallapati, 2003) all apply probability models (language model or relevance model) for story link detection. And the experiment results indicate that the performances are comparable with those using traditional vector space models, if not better.

On the basis of vector-based methods, this paper represents a method of dynamic information extending to improve the performance of story link detection. It makes use of the previous latest topically related story to extend the vector model of current being processed story. New dynamic models are generated for computing the similarity between two stories in current pair. This method resolves the problems of information shortage in stories and topic dynamic evolution in streams of data.

Before introducing the proposed method, we first describe a method which is implemented with vector model and cosine similarity function. This straight and classic method is used as a baseline to be compared with the proposed method.

4 Baseline Story Link Detection

The related work in story link detection shows that vector representation model with cosine function can be used to build the state-of-the-art story link detection systems. Many research organizations take this as their baseline system (Connell et al., 2004) (Yang et al., 2002). In this paper, we make a similar choice.

The baseline method represents each story as a vector in term space, where the coordinates represent the weights of the term features in the story. Each vector terms (or feature) is a single word plus its tag which is produced by a segmenter and part of speech tagger for Chinese. So if two tokens with same spelling are tagged with different tags, they will be taken as different terms (or features). It is notable that in it is independent between processing any two comparisons the baseline method.

4.1 Preprocessing

A preprocessing has been performed for TDT Chinese corpus. For each story we tokenize the text, tag the generated tokens, remove stop words, and then get a candidate set of terms for its vector model. After that, the term-frequency for each token in the story and the length of the story will also be acquired. In the baseline and dynamic methods, both training and test data are preprocessed in this way.

The segmenter and tagger used here is ICTCLAS [1] . The stop word list is composed of 507 terms. Although the term feature in the vector representation is the word plus its corresponding tag, we will ignore the tag information when filtering stop words, because almost all the words in the list should be filtered out whichever part of speech is used to tag them.

4.2 Feature Weighting

One important issue in the vector model is weighting the individual terms (features) that occur in the vector. Most IR systems employed the traditional $tf * idf$ weighting, which also provide the base for the baseline and dynamic methods in this paper. Furthermore, this paper adopts a dynamic way to compute the $tf * idf$ weighting:

$$w_i(f_i, d) = tf(f_i, d) * idf(f_i)$$

$$tf = t/(t + 0.5 + 1.5dl/dl_{avg})$$

$$idf = log((N + 0.5)/df)/log(N + 1)$$

where t is the term frequency in a story, dl is the length of a story, dl_{avg} is the average length of stories in the background corpus, N is the number of stories in the corpus, df is the number of the stories containing the term in the corpus.

The tf shows how much a term represents the story, while the idf reflects the distinctive ability of distinguishing current story from others. The dynamic attribute of the $tf * idf$ weighting lies in the dynamic computation of dl_{avg}, N and df. The background corpus used for statistics is incremental. As more story pairs are processed, more source files could be seen, and the background is expanding as well. Whenever the size of the background has changed, the values of dl_{avg}, N and df will update accordingly. We call this as incremental $tf*idf$ weighting. A story might have different term vectors in different story pairs.

[1] http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/

4.3 Similarity Function

Another important issue in the vector model is determining the right function to measure the similarity between two vectors. We have firstly tried three functions: cosine, Hellinger and Tanimoto, among which cosine function performs best for its substantial advantages and the most stable performance. So we consider the cosine function in baseline method.

Cosine similarity, as a classic measure and consistent with the vector representation, is simply an inner product of two vectors where each vector is normalized to the unit length. It represents cosine of the angle between two vector models $M_1 = \{(f_{1i}, w_{1i}), i \geq 1\}$ and $M_2 = \{(f_{2i}, w_{2i}), i \geq 1\}$.

$$\cos(M_1, M_2) = (\Sigma(w_{1i} \times w_{2i}))/\sqrt{(\Sigma w_{1i}^2)(\Sigma w_{2i}^2)}$$

Cosine similarity tends to perform best at full dimensionality, as in the case of comparing two stories. Performance degrades as one of the vectors becomes shorter. Because of the built-in length normalization, cosine similarity is less dependent on specific term weighting.

5 Dynamic Story Link Detection

5.1 Motivation

Investigation on the TDT corpus shows that news stories are usually short, which makes that their representation models are too sparse to reflect topics described in them. A possible method of solving this problem is to extend stories with other related information. The information can be synonym in a dictionary, related documents in external corpora, etc. However, extending with synonym is mainly adding repetitious information, which can not define the topics more clearly. On the other hand, topic-based research should be real-sensitive. The corpora in the same period as the test corpora are not easy to gather, and the number of related documents in previous period is very few. So it is also not feasible to extend the stories with related documents in other corpora. We believe that it is more reasonable that the best extending information may be the

story corpus itself. Following the TDT evaluation requirement, we will not use entire corpus at a time. Instead, when we process current pair of stories, we utilize all the stories before the current pair in the story corpus.

In addition, topics described by stories usually evolve along with time. A topic usually begins with a seminal event. After that, it will focus mainly on the consequence of the event or other directly related events as the time goes. When the focus in later stories has changed, the words used in them may change remarkably. Keeping topic descriptions unchanged from the beginning to the end is obviously improper. So topic representation models should also be updated as the topic emphases in stories has changed. Formerly we have planed to use related information to extend a story to make up the information shortage in stories. Considering more about topic evolution, we extend a story with its latest related story. In addition, up to now almost all research in story link detection takes the hypothesis that whether two stories in one pair are topically linked is independent of that in another pair. But we realize that if two stories in a pair describe a same topic, one story can be taken as related information to extend another story in later pairs. Compared with extending with more than one story, extending only with its latest related story can keep representation of the topic as fresh as possible, and avoid extending too much similar information at the same time, which makes the length of the extended vector too long. Since the vector will be renormalized, a too big length means evidently decreasing the weight of an individual feature which will instead cause a lower cosine similarity. This idea has also been confirmed by the experiment showing that the performance extending with one latest related story is superior to that extending with more than one related story, as described in section 6.3. The experiment results also show that this method of dynamic information extending apparently improves the performance of story link detection.

5.2 Method Description

The proposed dynamic method is actually the baseline method plus dynamic information extending. The preprocessing, feature weighting and similarity computation in dynamic method are similar as those in baseline method. However, the vector representation for a story here is dynamic. This method needs a training corpus to get the *extending threshold* deciding whether a story should be used to extend another story in a pair. We split the sequence of time-ordered story pairs into two parts: the former is for training and the later is for testing. The following is the processing steps:

1. Preprocess to create a set of terms for representing each story as a term vector, which is same as baseline method.

2. Run baseline system on the training corpora and find an optimum *topically link threshold*. We take this threshold as *extending threshold*. The *topically link threshold* used for making link decision in dynamic method is another predefined one.

3. Along with the ordered story pairs in the test corpora, repeat a) and b):

   (a) When processing a pair of stories $P_i = (s_{i1}, s_{i2})$, if s_{i1} or s_{i2} has an extending story, then update the corresponding vector model with its related story to a new dynamic one. The generation procedure of dynamic vector will be described in next subsection.

   (b) Computing the cosine similarity between the two dynamic term vectors. If it exceeds the *extending threshold*, then s_{i1} and s_{i2} are the latest related stories for each other. If one story already has an extending story, replace the old one with the new one. So a story always has no more than one extending story at any time. If the similarity exceeds *topically link threshold*, s_{i1} and s_{i2} are topically linked.

From the above description, it is obvious that dynamic method needs two thresholds, one for making extending decision and the other for making link decision. Since in this paper we will focus on the optimum performance of systems, the first threshold is more important. But *topically link threshold* is also necessary to be properly defined to approach a better performance. In the baseline method, term vectors are dynamic because of the incremental $tf * idf$

weighting. However, dynamic information extending is another more important reason in the dynamic method. Whenever a story has an extending story, its vector representation will update to include the extending information. Having the extending method, the representation model can have more information to describe the topic in a story and make the topic evolve along with time. The dynamic method can define topic description clearer and get a more accurate similarity between stories.

5.3 Dynamic Vector Model

In the dynamic method, we have tried two ways for the generation of dynamic vector models: increment model and average model. Supposing we use vector model $M_1 = \{(f_{1i}, w_{1i}), i \geq 1\}$ of story s_1 to extend vector model $M_2 = \{(f_{2i}, w_{2i}), i \geq 1\}$ of story s_2, M_2 will change to representing the latest evolving topic described in current story after extending.

1. Increment Model: For each term f_{1i} in M_1, if it also occurs as f_{2j} in M_2, then w_{2j} will not change, otherwise (f_{1i}, w_{1i}) will be added into M_2. This dynamic vector model only takes interest in the new information that occurs only in M_1. For features both occurred in M_1 and M_2, the dynamic model will respect to their original weights.

2. Average Model: For each term f_{1i} in M_1, if it also occurs as f_{2j} in M_2, then $w_{2j} = 0.5 * (w_{1i} + w_{2j})$, otherwise (f_{1i}, w_{1i}) will be added into M_2. This dynamic model will take account of all information in M_1. So the difference between those two dynamic models is the weight recalculation method of the feature occurred in both M_1 and M_2.

Both the above two dynamic models can take account of information extending and topic evolution. Increment Model is closer to topic description since it is more dependent on latest term weights, while Average Model makes more reference to the centroid concept. The experiment results show that dynamic method with Average Model is a little superior to that with Increment Model.

6 Experiment and Discussion

6.1 Experiment Data

To evaluate the proposed method, we use the Chinese subset of TDT4 corpus (LDC, 2003) developed by the Linguistic Data Consortium (LDC) for TDT research. This subset contains 27145 stories all in Chinese from October 2000 through January 2001, which are gathered from news, broadcast or TV shows.

LDC totally labeled 40 topics on TDT4 for 2003 evaluation. There are totally 12334 stories pairs from 1151 source files in the experiment data. The answers for these pairs are based on 28 topics of these topics, generated from the LDC 2003 annotation documents. The first 2334 pairs are used for training and finding *extending threshold* of dynamic method. The rest 10000 pairs are testing data used for comparing performances of baseline and the dynamic methods.

6.2 Evaluation Measures

The work is measured by the TDT evaluation software, which could be referred to (Hoogma, 2005) for detail. Here is a brief description. The goal of link detection is to minimize the cost due to errors caused by the system. The TDT tasks are evaluated by computing a "detection cost":

$$C_{det} = C_{miss} \cdot P_{miss} \cdot P_{target} + C_{fa} \cdot P_{fa} \cdot P_{non-target}$$

where C_{miss} is the cost of a miss, P_{miss} is the estimated probability of a miss, P_{target} is the prior probability under which a pair of stories are linked, C_{fa} is the cost of a false alarm, P_{fa} is the estimated probability of a false alarm, and $P_{non-target}$ is the prior probability under which a pair of stories are not linked. A miss occurs when a linked story pair is not identified as being linked by the system. A false alarm occurs when the pair of stories that are not linked are identified as being linked by the system. A target is a pair of linked stories; conversely a non-target is a pair of stories that are not linked. For the link detection task these parameters are set as follows: C_{miss} is 1, P_{target} is 0.02, and C_{fa} is 0.1. The cost for each topic is equally weighted (usually the cost of topic-weighted is the mainly evaluation parameter) and normalized so that for a given system, the normalized value $(C_{det})_{norm}$ can be no less than

one without extracting information from the source data:

$$(C_{det})_{norm} = \frac{C_{det}}{\min(C_{miss}P_{target}, C_{fa}P_{non-target})}$$

$$(C_{det})_{overall} = \Sigma_i(C^i_{det})_{norm}/\#topics$$

where the sum is over topics i. A detection curve (DET curve) is computed by sweeping a threshold over the range of scores, and the minimum cost over the DET curve is identified as the minimum detection cost or min DET. The topic-weighted DET cost is dependent on both a good minimum cost and a good method for selecting an operating point, which is usually implemented by selecting a threshold. A system with a very low min DET cost can have a much larger topic-weighted DET score. Therefore, we focus on the minimum DET cost for the experiments.

6.3 Experiment Results

In this paper, we have tried three methods for story link detection: the baseline method described in Section 4 and two dynamic methods with different dynamic vectors introduced in Section 5. The following table gives their evaluation results.

metrics	baseline	dynamic 1	dynamic 2
P_{miss}	0.0514	0.0348	0.0345
P_{fa}	0.0067	0.0050	0.0050
$Clink_{min}$	0.0017	0.0012	0.0012
$Clink_{norm}$	0.0840	0.0591	0.0588

Table 1: Experiment Results of Baseline System and Dynamic Systems

In the table, $Clink_{min}$ is the minimum $(C_{det})_{overall}$, DET Graph Minimum Detection Cost (topic-weighted), $Clink_{norm}$ is the normalized minimum $(C_{det})_{overall}$, the dynamic 1 is the dynamic method which uses *Increment Model* and the dynamic 2 is the dynamic method which uses *Average Model*. We can see that the proposed two dynamic methods are both much better than baseline method on all four metrics. The $Clink_{Norm}$ of dynamic 1 and 2 are improved individually by 27.2% and 27.8% as compared to that of baseline method. The difference between two dynamic methods is due to different in the P_{miss}. However, it is too little to compare the two dynamic systems. We also make additional experiments in which a story is extended with all of its previous related stories. The minimum (Cdet)overall is 0.0614 for the system using Increment Model, and 0.0608 for the system using Average Model. Although the performances are also much superior to baseline, it is still a little poorer than that with only one latest related story, which confirm the ideal described in section 5.1.

Figure 1, 2 and 3 show the detail evaluation information for individual topic on Minimum Norm Detection Cost, P_{miss} and P_{fa}. From Figure 1 we know these two dynamic methods have improved the performance on almost all the topic, except topic 12, 26 and 32. Note that detection cost is a function of P_{miss} and P_{fa}. Figure 2 shows that both two dynamic methods reduce the false alarm rates on all evaluation topics. In Figure 3 there are 20 topics on which the miss rates remain zero or unchange. The dynamic methods reduce the miss rates on 5 topics. However, dynamic methods get relatively poorer results on topic 12, 26 and 32 . Altogether dynamic methods can notably improve system performance on evaluation metrics of both individual and weighted topic, especially the false alarm rate, but on some topics, it gets poorer results.

Further investigation shows that topic 12, 26 and 32 are about Presidential election in Ivory Coast on October 25, 2000, Airplane Crash in Chiang Kai Shek International Airport in Taiwan on October 31, 2000, and APEC Conference on November 12-15, 2000 at Brunei. After analyzing those story pairs with error link decision, we can split them into two sets. One is that two stories in a pair are general linked but not TDT specific topically linked. Here general linked means that there are many common words in two stories, but the events described in them happened in different times or different places. For example, Airplane Crash is a general topic, while Airplane Crash in certain location at specification time is a TDT topic. The other is that two stories in a pair are TDT topically linked while they describe the topic from different perspectives. In this condition they will have few common words. These may be due to that the information extracted from stories is still not accurate enough to represent them. It also may be because of the

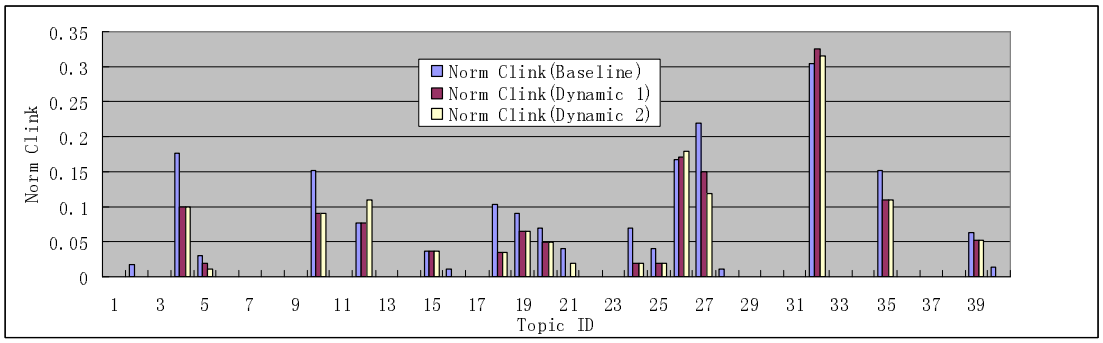

Figure 1: Normalized Minimum Detection Cost for individual topic

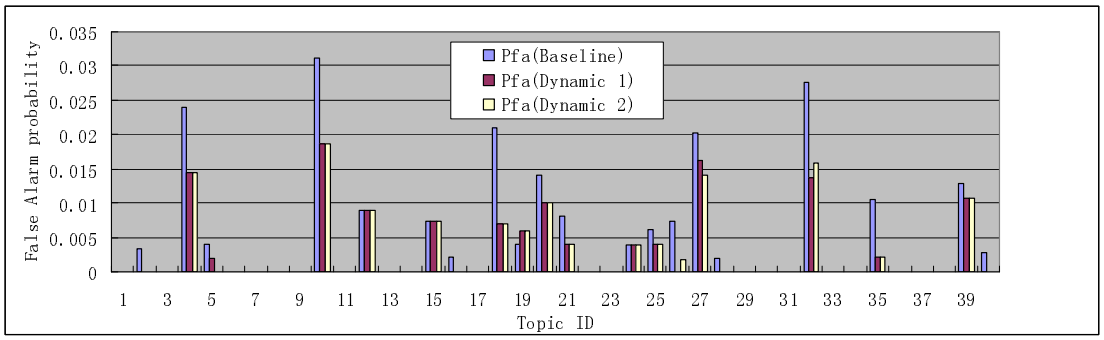

Figure 2: P_{fa} for individual topic

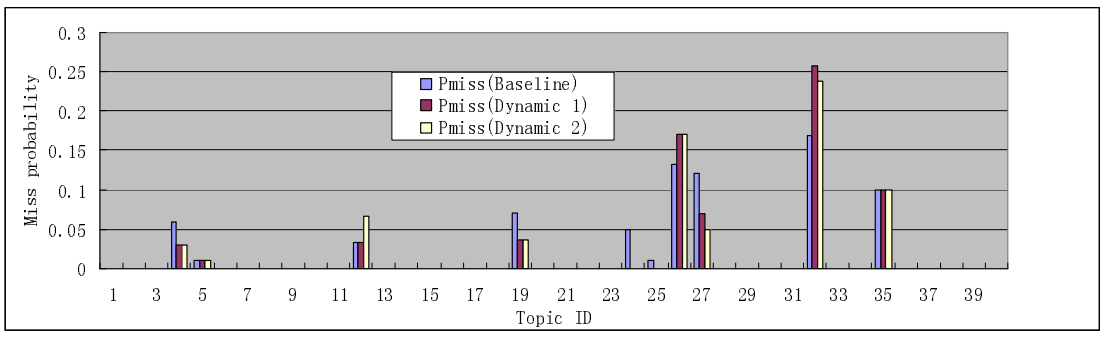

Figure 3: P_{miss} for individual topic

deficiency of vector model itself. Furthermore, we know that the extending story is chosen by cosine similarity, which results that the extending story and the extended story are usually topically linked from the same perspectives, seldom from different perspectives. Therefore the method of information extending may sometimes turn the above first problem worse and have no impact on the second problem. So mining more useful information or making more use of other useful resources to solve these problems will be the next work. In addition, how to represent this information with a proper model and seeking better or more proper representation models for TDT stories are also important issues. In a word, the method of information extending has been verified efficient in story link detection and can provide a reference to improve the performance of some other similar systems whose data must be processed serially, and it is also hopeful to combined with other improvement technologies.

7 Conclusion

Story link detection is a key technique in TDT research. Though many approaches have been tried, there are still some characters ignored. After analyzing the characters and deficiency in TDT stories and story link detection, this paper presents a method of dynamic information extending to improve the system performance by focus on two problems: information deficiency and topic evolution. The experiment results indicate that this method can effectively improve the performance on both miss and false alarm rates, especially the later one. However, we should realize that there are still some problems to solve in story link detection. How to compare general topically linked stories and how to compare stories describing a TDT topic from different angles will be very vital to improve system performance. The next work will focus on mining more and deeper useful information in TDT stories and exploiting more proper models to represent them.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (60403050), Program for New Century Excellent Talents in University (NCET-06-0926) and the National Grand Fundamental Research Program of China under Grant(2005CB321802).

References

James Allan, Victor Lavrenko, Daniella Malin, and Russell Swan. 2000. Detections, bounds, and timelines: Umass and tdt–3. In *Proceedings of Topic Detection and Tracking (TDT–3)*, pages 167–174.

J. Allan, A. Bolivar, M. Connell, S. Cronen-Townsend, A Feng, F. Feng, G. Kumaran, L. Larkey, V. Lavrenko, and H. Raghavan. 2003. Umass tdt 2003 research summary. In *proceedings of TDT workshop*.

James Allan, editor. 2002. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Norvell, Massachusetts.

Francine Chen, Ayman Farahat, and Thorsten Brants. 2004. Multiple similarity measures and source-pair information in story link detection. In *HLT-NAACL*, pages 313–320.

Margaret Connell, Ao Feng, Giridhar Kumaran, Hema Raghavan, Chirag Shah, and James Allan. 2004. Umass at tdt 2004. In *TDT2004 Workshop*.

Niek Hoogma. 2005. The modules and methods of topic detection and tracking. In *2nd Twente Student Conference on IT*.

Victor Lavrenko, James Allan, Edward DeGuzman, Daniel LaFlamme, Veera Pollard, and Stephen Thomas. 2002. Relevance models for topic detection and tracking. In *Proceedings of Human Language Technology Conference (HLT)*, pages 104–110.

LDC. 2003. Topic detection and tracking - phase 4. Technical report, Linguistic Data Consortium.

Ramesh Nallapati and James Allan. 2002. Capturing term dependencies using a language model based on sentence trees. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 383–390. ACM Press.

Ramesh Nallapati. 2003. Semantic language models for topic detection and tracking. In *HLT-NAACL*.

NIST. 2003. The 2003 topic detection and tracking task definition and evaluation plan. Technical report, National Institute of Standards and Technology(NIST).

Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 688–693. ACM Press.