IPSJ Transactions on Databases Vol. 1 No. 3 11–19 (Dec. 2008)

Regular Paper

Improving Document Representation for Story Link Detection by Modeling Term Topicality

Chirag Shah[†1,∗1] and Koji Eguchi[†1,†2]

Several information organization, access, and filtering systems can benefit from different kind of document representations than those used in traditional Information Retrieval (IR). Topic Detection and Tracking (TDT) is an example of such a domain. In this paper we demonstrate that traditional methods for term weighing do not capture topical information and this leads to inadequate representation of documents for TDT applications. We present various hypotheses regarding the factors that can help in improving the document representation for Story Link Detection (SLD)—a core task of TDT. These hypotheses are tested using various TDT collections. From our experiments and analysis we found that in order to obtain a *faithful* representation of documents in TDT domain, we not only need to capture a term's importance in traditional IR sense, but also evaluate its *topical* behavior. Along with defining this behavior, we propose a novel measure that captures a term's importance at the collection level as well as its discriminating power for topics. This new measure leads to a much better document representation as reflected by the significant improvements in the results.

1. Introduction

Document representation is one of the most common and critical stages of an information organization and access system. Several methods and models of document representation have been proposed based on the target application. Schemes such as vector space representations [19] and language models [18], which treat document as a bag of words, are quite common and successful in many Information Retrieval (IR) applications. They are general enough to be applicable to almost any IR-based application, however, certain domains or tasks require different approaches to document representations. In this paper we argue that Topic Detection and Tracking (TDT) [1] is one such domain that can benefit from a different kind of document representation than other tried and proposed models earlier. In order to support this argument, we analyze the peculiarities of TDT and propose a novel approach for document representation. In particular, we focus on term weighing and use Story Link Detection (SLD), a core task of TDT, as the target application.

To understand our focus and contribution in this paper, let us identify three major components in an IR implementation on the system side:

(1) The units of representation
(2) The method of weighing these units
(3) Matching criteria between a query and a document or between a document and a document

There have been a few studies focusing on the first of these components that demonstrated that using named entities as units of representation is a good idea for TDT applications [3),14),20]. Some studies investigated using noun phrases along with named entities [6]. A considerable amount of work has also been done on comparing two streams of text that includes cosine similarity [19], information-theoretic distance [16),23)], etc. As far as weighting the terms is concerned, there has not been much work that leverages the uniqueness of TDT. Most of the work so far has used the popular methods such as TF [8] and TFIDF [19]. In the work reported here we shall demonstrate how we can uniquely capture the topical nature of documents with the term weights in order to construct a better representation. Throughout our experiments we shall use the same units (all the words), and the same matching criteria (cosine similarity of two texts) to allow comparison of various term weighing methods.

We started this work by using some traditional IR methods for document representation for SLD task. As we moved from one method of term weighing to another, we realized their shortcomings. Our experiments and analysis showed that none of the traditional methods captured the *topical* information to weigh the terms, which we found was essential for TDT domain. This led us into investigating a better way to capture and incorporate such information. The main contributions of our work reported here are this realization and the methods that emerged from it.

†1 National Institute of Informatics, Tokyo
†2 Kobe University, Kobe, Japan
∗1 Presently with the University of North Carolina at Chapel Hill, USA

2. Background

In this section we review the TDT research and provide more specifics of SLD task. We also show the uniqueness of TDT tasks and contrast with traditional IR systems. With this explanation, we try to motivate the use of topical information to represent the news stories.

2.1 TDT

The Topic Detection and Tracking (TDT) research has provided a standard platform for addressing event-based organization of broadcast news and evaluating such systems [1]. The governing motivation behind such research was to provide a core technology for a system that would monitor broadcast news and alert an analyst to new and interesting events happening around the world [2]. The research program of TDT focuses on five tasks: story segmentation, first story detection, cluster detection, tracking, and story link detection. Each is viewed as a component technology whose solution will help address the broader problem of event-based news organization. The details of each of these tasks are beyond the scope of this paper. We, instead, want to focus on the general characteristics of TDT and specific details of the story link detection task.

To appreciate the unique nature of TDT, it is important to understand the notion of a topic. In TDT, a topic is defined to be a set of news stories that are strongly related by some seminal real-world event. For instance, when hurricane Katrina hit the coast, it became the seminal event that triggered a new topic. The stories that discussed the origin of this hurricane, the damage it did to the places, the rescue efforts, etc., were all parts of the original topic. Stories on another hurricane (occurring in the same region or time) could make up another topic. This shows an important contrast with typical IR. Along with hurricane Katrina, a query "hurricane" will bring up the documents about other hurricane related events. On the other hand, for the query "hurricane Katrina", some of the stories that followed the original event of hurricane Katrina may not be considered as "about" hurricane Katrina by the traditional IR measures and would be ranked very low in the retrieval set.

This contrast indicates that the notion of an event-based topic is narrower than a subject-based topic; it is built upon its triggering event. Hereafter, we focus on dealing with an event-based topic rather than a subject-based. A typical *topic* would have a start time and it would fade off from the news at some point of time. Since it is a specific event, it happened not only at some particular time, but in a specific location, and usually with an identifiable set of participants. In other words, a topic is well defined in scope and possibly in time. In this paper we would ignore the temporal nature of a topic and just focus on the scope.

2.2 SLD

The Story Link Detection (SLD) task evaluates a TDT system that detects if two stories are "linked" by the same event. For TDT, two stories are linked if they discuss the same event. Unlike other TDT tasks, link detection was not motivated by a hypothetical application, but rather the task of detecting when stories are linked is a "kernel" function from which the other TDT tasks can be built.

2.2.1 The Uniqueness of SLD

For this task, a set of pairs of documents are given to compare with each other and we need to declare if they are on the same *topic*, i.e., event-based topic as defined in section 2.1. It is important to note that although this might seem similar to finding the document similarity, there is a fine difference here. It is possible that two documents do not share many common terms but belong to the same topic. It is also possible that two documents may have several terms that match with one another, but they talk about different topics.

2.2.2 Evaluation

The performance of any TDT system is characterized in terms of the probabilities of missed detection (P_{Miss}) and false alarm errors (P_{Fa}) [9]. These error probabilities are linearly combined into a single detection cost, C_{Det}, by assigning costs to missed detection and false alarm errors and specifying an *a priori* probability of a target. The resulting formula is

$$C_{Det} = (C_{Miss} * P_{Miss} * P_{Target} + C_{Fa} * P_{Fa} * (1 - P_{Target})) \quad (1)$$

where P_{Miss} = (No. of missed detection)/(No. of targets), and P_{Fa} = (No. of false alarms)/(No. of non-targets), on the basis of human-judged gold standard data determining whether or not a document is on a target topic. C_{Miss} and C_{Fa} are the costs of a missed detection and a false alarm, respectively, and are pre-specified by TDT as $C_{Miss} = 1.0$ and $C_{Fa} = 0.1$. P_{Target} is the *a priori*

probability of finding a target. This detection cost C_{Det} is then normalized as given below.

$$(C_{Det})_{Norm} = \frac{C_{Det}}{\min\left(C_{Miss} * P_{Target}, C_{Fa} * (1 - P_{Target})\right)} \quad (2)$$

The cost function defined above is standard in TDT, and so we used this as an evaluation measure in this paper.

2.2.3 Recent Work on TDT with Emphasis on SLD

Although SLD is considered to be the core technology for almost all of the tasks in TDT, not much has been done specifically for SLD. The efforts have rather been focused on more practical applications such as new event detection and tracking. However, since different tasks in TDT are inter-related, we could use the experiences of one task for other tasks. Undoubtedly, the most popular approach for SLD, and TDT in general, has been the vector space model. This has included Carnegie Mellon University (CMU) [24], NorthEastern University in China (NEU) [8], University of Iowa (UIowa) [7], and University of Massachusetts at Amherst (UMass) [5]. Relevance models, another popular method for IR, have also been widely used for document representation and matching. This model has been heavily utilized by UMass. Relevance model [16] uses the new story as the query and all the past stories as the collection [15]. It then performs retrieval and expands the new story with top words. Finally, the similarity between the models of the new story and the training stories (from past stories) is computed using Kullback-Leibler (KL) divergence. A few of the approaches that are not very frequently used, but yet worth mentioning are voting scheme among documents similarity functions [24], and graph-based comparison [7].

What is common among the above methods is their neglect in considering the *topicality* of the terms while representing the documents. As we shall see in the following section, this will become the focus of our investigation.

3. Hypotheses and Proposed Methods

There are several hypotheses and conjectures in the field of TDT that have driven much of the research done in this field in the recent years. However, there is a lack of work that studies them systematically to understand their validity as well as effects. This is a key strength of our work; we studied and experimented with a more unified approach that takes one hypothesis at time and tests it using either the tools available or with new proposed methods.

In this section we present various hypotheses about the factors that can affect the document representation. To be specific, these hypotheses are relating to the term weighing schemes.

3.1 Hypothesis-1: Capturing a Term's Importance at the Document and/or Collection Level Provides a Faithful Representation

The majority of document representation models consider terms as the units of representation and incorporate their *importance* in a given document and/or collection by the means of some frequencies or probabilities.

Method-1. TFIDF on all the words—Baseline

TFIDF based representation of documents is widely used in document similarity [19], document classification [17], and document clustering [12] literature. It incorporates a term's importance at the document level by using its frequency in that document (TF), and its importance at the collection level by using its inverse document frequency (IDF). We adopted this approach as our baseline, which is a typical bag-of-words approach. TF values were found using the following equation.

$$TF(t,d) = \frac{freq(t,d)}{freq(t,d) + 0.5 + \frac{1.5 * DocLen(d)}{Avg\_DocLen}} \quad (3)$$

where $freq(t,d)$ is the raw frequency of term t in a given document d, $DocLen(d)$ is the length of the given document, and Avg_DocLen is the average length of the documents. IDF values were found using the following formulation.

$$IDF(t) = \log \frac{N+1}{N_t + 0.5} \quad (4)$$

where N is the total number of documents in the collection and N_t is the number of documents in which term t occurs.

We construct vectors for each document using $TF(t,d) \times IDF(t)$ and then find cosine between two vectors. This score becomes the similarity measurement for the given documents. If this score is above the threshold, then the given pair of stories are said to be on the same topic. Later we shall see how to derive this

threshold from the training data.

Method-2. Information Content (IC)

Another approach to finding a term's importance is by measuring the amount of information it carries. Information content (IC) of a term t is found using the following equation.

$$IC(t) = -\log_2 P(t|coll) \quad (5)$$

where $P(t|coll)$ is the probability of term t occurring in collection *coll* which is estimated as the following relative frequency.

$$P(t|coll) = \frac{\sum_{d \in coll} freq(t,d)}{\sum_{d \in coll} \sum_{t \in d} freq(t,d)} \quad (6)$$

Method-3. Pointwise KL (PKL) divergence scores

KL divergence or its variations are widely used in language modeling framework [16),23)]. In general, it is useful for finding how different two probability distributions are. We conjectured that the further a term's *behavior* in a document is different from the *behavior* of the world, the more useful it is for the representation. In other words, if a term is unique, it is more important than those terms that are not, such as stopwords. This intuition can be realized using pointwise KL divergence, as some researchers used in different contexts [13),22)]. The traditional KL divergence, modified for pointwise computation, results in the following formulation.

$$PKL(t,d) = P(t|d) \log \frac{P(t|d)}{P(t|coll)} \quad (7)$$

where $PKL(t,d)$ is the pointwise KL divergence between document d and collection *coll* with respect to term t, $P(t|d)$ is the probability of term t occurring in a given document d, and $P(t|coll)$ is the probability of t occurring in *coll*, which is calculated as given in equation (6). $P(t|d)$ is estimated as the following.

$$P(t|d) = \frac{freq(t,d)}{\sum_{t \in d} freq(t,d)} \quad (8)$$

We implemented these three weighing schemes using TDT2 collection as training and TDT3 collection for testing[⋆1]. This means that for method 1, IDF values for all the terms were found using TDT2 (training collection) and combined with TF values found using TDT3 (testing collection) to compute the TFIDF scores. For method 2, we found the information content of each term using TDT2 and used it to represent the terms of TDT3. For method 3, $P(t|coll)$ were estimated using TDT2 collection and $P(t|d)$ using TDT3. For each of these methods, once we computed the weights for different terms, we created vectors of these terms with their associated weights. Following a simple vector space model guidelines, the matching was performed using cosine function and the cost was calculated using Equation (2). The results of these systems are displayed later in Figure 1. It is important to note here that the lower the cost is, the better the system is.

3.2 Hypothesis-2: Capturing the Topical Nature of a Term Provides a Better Representation for TDT-like Applications

We realized that in none of the methods tried before, the information about topics is explicitly captured. It seemed interesting and useful to us to see what would happen if we incorporated such information while weighing the terms. However, it was not clear to us how exactly we could go about defining and capturing *topicality*. Following are two possible methods of doing so. They are based on some intuition, but by no means the best ways to capture *topicality* of terms.

Method-4. Topicality scores

We started with a very simple approach of defining and capturing *topicality*. Our formulation was basically derived from the following intuition.

- In the documents on the same topic, a term that occurs frequently is more useful than the terms that are not very frequent.
- A term that is more frequent in a topic and less frequent in other topics is more useful.

It can be seen that the desired characteristics of a topical terms are very similar to the formulation of TFIDF—the first point is similar to finding TF and the

⋆1 This is based on the assumption that we cannot know how the whole collection is, in advance, in the context of TDT.

second point is similar to finding IDF. However, there are some basic differences here, which are mainly in point two. In case of normal IDF, we are interested in calculating how frequently the term occurs in the entire collection, whereas in the formulation that we gave here, we are interested in finding how frequent the term is in the documents of different topics. This is translated in the following formula.

$$\begin{aligned}\text{Topicality}(t) &= (\text{Prob. of } t \text{ occurring in a topic}) \\ &\quad \times (\text{Prob. of } t \text{ not occurring in any other topic}) \\ &= \max_{i \in \{1,\cdots,K\}} \left[P(t|k_i) \times \left(1 - P(t|k_{-i})\right)\right] \end{aligned} \tag{9}$$

where $k_{-i} = \cup_{j \in \{1,\cdots,K\}, j \neq i} k_j$, K is the total number of topics, and k_i is i^{th} topic. Given topics assigned to documents, we can estimate $P(t|k_i)$ as the following.

$$P(t|k_i) = \frac{\sum_{d \in k_i} freq(t,d)}{\sum_{d \in k_i} \sum_{t \in d} freq(t,d)} \tag{10}$$

In this paper, we used the number of topics and the topic labels that are given along with TDT collections, for training.

Method-5. Topical Information Content

Earlier we proposed to use information content of a term as a mean to find its weight in document representation. This information content was measured with respect to the collection. We now change it slightly to evaluate it with respect to the topics. This new measure is defined below, to which we would refer as the Topical Information Content (TIC).

$$TIC(t) = \max_{i \in \{1,\cdots,K\}} (-\log P(t|k_i)) \tag{11}$$

where $P(t|k_i)$ is the probability of term t given topic k_i. Once again, we used TDT2 collection for training and TDT3 for testing. This means that for method 4, the *Topicality* was computed using TDT2 collection and used for representing TDT3 terms. Similarly, for method 5, the information content with respect to the topics was found on TDT2 and used for TDT3.

From the results, we found that using only *topical* information resulted in better performance over the baseline, but did not give any advantage over the other methods of term weighing that did not use *topical* information. However, if the *importance* of a term in the collection and that in a topic are two relatively orthogonal factors, then we might be able to combine them in a way that would outperform a method that uses only one of them. Thus, we came up with the following hypothesis.

3.3 Hypothesis-3: Combining a Term's Importance at Collection Level and Its Topicality Provides a Better Representation than that of Either of them Used Separately.

Method-6. Topical KL (TKL) divergence scores

From our experiments so far, we came to a realization that Pointwise KL (PKL) is a good way of capturing a term's importance. We now approach the calculation for PKL with topicality in the formulation. Earlier we defined the PKL divergence with respect to the document and the collection. Now we will measure it with respect to the topic and the collection. This new formulation is given below.

$$TKL(t) = \max_{i \in \{1,\cdots,K\}} \left[P(t|k_i) \log \frac{P(t|k_i)}{P(t|coll)}\right] \tag{12}$$

where k_i is a topic and *coll* is the collection. Once again, we find the scores for each topic for a given term and take the maximum. While representing a document by a vector, each term is weighted by its corresponding topical KL divergence score.

We found both $P(t|k)$ and $P(t|coll)$ using our training collection TDT2 and used these values to compute a term's weight on TDT3 collection. The result obtained by this system along with all the previous systems is displayed in **Fig. 1**. As we can see, our proposed approach obtains the least cost. The actual values of these costs are given in **Table 1**.

It is important to note here that Method-6 (Topical KL) is the combination of the ideas behind method-3 (PKL) and method-4 (Topicality). We selected this combination because method-4 (Topicality) alone worked better than method-5 (Topical IC) alone and method-3 (PKL) worked better than method-1 (TFIDF) or method-2 (IC), as shown in Table 1.

To measure how significant these changes were, we used standard two-tailed

paired t-test. According to this test, all of our Methods 2-6 were tested significantly different from the baseline, Method-1. Since all of our systems that we built to test these hypotheses produced better results than the baseline, we can claim that our proposed methods were significantly better than the traditional TFIDF method. However, one can argue that such a test that is parametric in nature and assumes a certain kind of underlying distribution, may not be appropriate for usage in our case. We, therefore used McNemar's statistical significance test [10], which is non-parametric. McNemar's test uses only the number of discordant pairs, that is, the pairs for which we have different decisions in given two techniques. Let us define the pairs that transferred from YES to NO to be R and the pairs that transferred from NO to YES to be S. We then calculated the Chi-Square value using the following equation.

$$\chi^2 = \frac{(|R-S|-1)^2}{R+S} \tag{13}$$

Using this Chi-Square value and associated p-value that is given by this test, we can claim that if there were really no association between the given techniques, there is probability p of chance that the observed odds ratio would be so far from 1.0 (no association). From these tests, we found the improvements of our methods statistically significant than that of the baseline system.

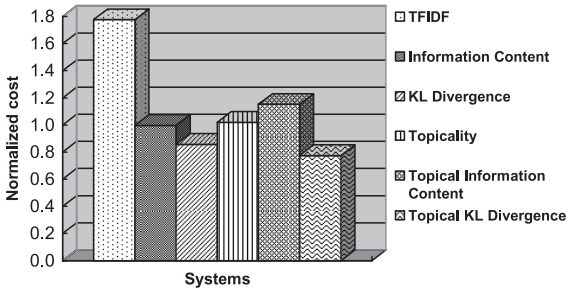

Fig. 1 Normalized detection cost for various systems.

4. Additional Experiments

In order to support our hypothesis-3 with additional empirical results, we carried out experiments using TDT4 collection, too. This collection is significantly different than TDT2 and TDT3 collections. A summary of the size of the collections and the number of document-pairs for which the gold standard judgments are available, is given in **Table 2**. The results of our baseline, plain topicality, and topical KL systems with TDT4 collection are given in **Tables 3** and **4**. As we can see, using merely topicality scores does worse than the baseline, but our proposed system of topical KL scores still outperforms the baseline. Since the testing collection differs quite a bit from the training collections in these experiments, the improvements achieved by our proposed method are not as dramatic as the results reported in Table 1, but they are still significant. **Table 5** shows how the improvements given by our topical KL scores are significant comparing with the baseline method using TFIDF scores. The results indicate that the improvements of our method can still be said to be statistically significant than that of the baseline method.

We found these results interesting and tried to do further analysis. It turns out

Table 2 Collection statistics.

Coll.	Documents	Topics	Document-pairs
TDT2	83979	100	10080
TDT3	67111	60	11736
TDT4	98245	80	20000

Table 1 Cost for various systems. Training on TDT2, testing on TDT3.

System	Normalized detection cost	Improvement over the baseline
TFIDF scores (baseline)	1.7758	–
Information content	0.9971	43.85%
PKL divergence	0.8570	51.74%
Topicality scores	1.0197	42.58%
Topical information content	1.1526	35.09%
Topical KL scores	0.7749	56.36%

Table 3 Training on TDT2, testing on TDT4.

System	Normalized Cost	Improvement over the baseline
TFIDF scores (baseline)	1.1104	–
Topicality scores	1.1392	−2.59%
Topical KL scores	1.0383	6.49%

Table 4 Training on TDT3, testing on TDT4.

System	Normalized Cost	Improvement over the baseline
TFIDF scores (baseline)	0.9254	–
Topicality scores	0.9915	−7.14%
Topical KL scores	0.8983	2.93%

Table 5 Results of significance tests.

Training	Testing	Two tailed paired t-test p-value	McNemar's test p-value
TDT2	TDT4	0.0000	0.0000
TDT3	TDT4	0.0000	2.46e-05

Table 6 Overlap of unique named entities between training and testing collections.

	Coll.	No. of named entities	Coll.	No. of named entities
Training	TDT2	71536	TDT3	140513
Testing	TDT3	68977	TDT4	348456
Common		22006		40922

that although the vocabulary of TDT4 is not very different than those of TDT2 and TDT3, the nature of the topics is quite different. One way of understanding how different the document topics between two collections were, is to see how many named entities they have in common. For this objective, we picked up a couple of extreme cases: TDT2-TDT3 pair as the best case and TDT3-TDT4 pair as the worst case, and compared them from the standpoint of how much the named entities in each pair are overlapped, as shown in **Table 6**.

These statistics show that for our first set of experiments that were trained on TDT2 collection and tested on TDT3 collection, we had a high overlap of named entities (about one third of the named entities of the testing data were present in the training data). However, when we used TDT3 collection for training and TDT4 collection for testing for our second set of the experiments, the testing data had just little over one tenth of the named entities from the training data.

This is an indication that there must have been a higher similarity of topics between TDT2 and TDT3 than that of TDT3 and TDT4. Since our *Topical scores* method takes into account only the topical information, it did not do as well on TDT3-TDT4 setting as it did on TDT2-TDT3 pair. Thus, using only topical information from TDT2 or TDT3 on TDT4 collection hurts the performance. On the other hand, our proposed method that uniquely combines the topical as well as the overall collection information still does the best. This demonstrates the robustness of our proposed approach.

5. Conclusion

In this paper we presented a novel approach for term weighing that incorporated a term's importance at the collection level at the same time capturing its topicality. With our analysis and experiments, we showed that traditional IR techniques of term weighing do not consider the topical information, which is essential for TDT tasks. We selected SLD as our target application, which is a core task of TDT. Through a set of hypotheses testing, experiments, and analysis, we realized that while traditional IR methods of term weighing do not capture topical information explicitly, the information that the topics provide is still very useful. We then proposed a unique way of combining topical information with the information that a traditional IR term weighing scheme provides. This method consistently outperformed the baseline across different TDT collections that we used. Another advantage of this model is that since it is based on well-studied information theoretic and probabilistic frameworks, it becomes easier and more effective to analyze and understand it.

The core contribution of our work can be summarized as the following.

(1) Understanding that the document representation scheme that works for most of the information processing applications may be fit for all the situations.

(2) Realization that document representation in TDT domain can benefit from capturing information that incorporates topical behavior of terms and/or documents.

(3) Finding that if we have a good way of combining a term's importance in a collection as well as its topical nature, we can perform better than using only either of them.

The experiments reported here are done for SLD task only. However, given that SLD is at the core of TDT and the generalizable nature of our findings, we conjecture that our proposed methods, which gave significant improvements, should help in other TDT tasks as well.

The state-of-the-art systems for SLD use external sources to do better deductions [5),24)]. The systems that do not use such extra knowledge are shown to give cost 1.0 or higher depending on the algorithm used for matching and retrieval. It is important to note that our contribution is not in surpassing any state-of-the-art system, since we are not using any external knowledge, but in proposing an effective mechanism of representing news stories with minimal information available. It is our conjecture that given additional information that high performance systems use, we would also be able to do much better.

There are several directions that can be explored based on our findings reported in this paper. For instance, as noted earlier, in TDT a topic is defined in scope and time. Our work focused on a topic's scope only. A nice extension of our model would be incorporating temporal aspect of topic as well. We have provided a theoretically verifiable and objective way of computing topicality of terms, which was shown to be reasonably robust. One of the ways in which our work could be extended is by providing a correlation analysis of collection size and the robustness of the term weighing functions. In the same spirit, the presented framework can be used for further exploring various behavioral attributes of a term that relate to its topicality. For instance, one could study the correlations among part of speech of a term (noun, verb, adjective, etc.), its information content, and its topicality. Findings from such a study could be used for term selection. Another line of our future direction is to incorporate "unsupervised" topic modeling [4),11),21)], which can be cast to our term weighting method, while this paper assumed training data are labeled with topics.

References

1) Allan, J.(ed.): *Topic Detection and Tracking*, Kluwer Academic Publishers (2002).
2) Allan, J., Carbonell, J., Doddington, G.R., Yamron, J. and Yang, Y.: Topic Detection and Tracking Pilot Study Final Report, Technical report, NIST (1998).
3) Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lvrenko, V., Hoberman, R. and Caputo, D.: Topic-based Novelty Detection, Technical report, Center for Language and Speech Processing, John Hopkins University (1999).
4) Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
5) Connell, M., Feng, A., Kumaran, G., Raghavan, H., Shah, C. and Allan, J.: UMass at TDT 2004, Technical report, University of Massachusetts, Amherst (2004).
6) Eichmann, D.: Experiments with Tracking/Detection/etc. using Entities and Noun Phrases, Technical report, University of Iowa (2001).
7) Eichmann, D.: Link Detection, Technical report, University of Iowa (2004).
8) Fiscus, J. and Wheatley, B.: Overview of the TDT 2004 Evaluation and Results, Technical report, NIST (2004).
9) Fiscus, J.G. and Doddington, G.R.: Topic Detection and Tracking Evaluation Overview, *Topic Detection and Tracking*, Allan, J. (ed.), Kluwer Academic Publishers, chapter 2, pp.17–32 (2002).
10) Gillick, L. and Cox, S.: Some Statistical Issues in the Comparison of Speech Recognition Algorithms, *Proc. 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP-89*), Glasgow, UK, pp.532–535 (1989).
11) Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proc. 22nd Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, California, USA, pp.50–57 (1999).
12) Kanungo, T., Mount, D.M., Netanyahu, N., Piatko, C., Silverman, R. and Wu, A.Y.: An Efficient k-means Clustering Algorithm: Analysis and Implementation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.24, pp.881–892 (2002).
13) Kelly, D., Diaz, F., Belkin, N.J. and Allan, J.: A User-Centered Approach to Evaluating Topic Models, *Proc. 26th European Conference on Information Retrieval* (*ECIR 2004*), Sunderland, UK, pp.27–41 (2004).
14) Kumaran, G. and Allan, J.: Text Classification and Named Entities for New Event Detection, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, pp.297–304 (2004).
15) Lavrenko, V., Allan, DeGuzman, J.E., LaFlamme, D., Pollard, V. and Thomas, S.: Relevance Models for Topic Detection and Tracking, *Proc. 2nd Human Language Technology Conference*, San Diego, California, USA, pp.104–110 (2002).
16) Lavrenko, V. and Croft, W.B.: Relevance-Based Language Models, *Proc. 24th Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, Louisiana, USA, pp.120–127 (2001).
17) Lewis, D.D.: Evaluating and Optmizing Autonomous Text Classification Systems, *Proc. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, pp.246–254 (1995).

18) Ponte, J.M. and Croft, W.B.: A Language Modeling Approach to Information Retrieval, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp.275–281 (1998).
19) Salton, G.(ed.): *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley (1989).
20) Shah, C., Croft, W.B. and Jensen, D.: Representing Documents with Named Entities for Story Link Detection (SLD), *Proc. 15th ACM Conference on Information and Knowledge Management*, Arlington, Virginia, USA (2006).
21) Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, Vol.101, No.476, pp.1566–1581 (2006).
22) Tomokiyo, T. and Hurst, M.: A Language Model Approach to Keyphrase Extraction, *Proc. ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp.33–40 (2003).
23) Zhai, C. and Lafferty, J.: Model-based Feedback in the Language Modeling Approach to Information Retrieval, *Proc. 10th ACM Conference on Information and Knowledge Management*, Atlanta, Georgia, USA, pp.403–410 (2001).
24) Zhang, Y. and Callan, J.: CMU DIR Supervised Tracking Report, Technical report, Carnegie Mellon University (2004).

(Received June 4, 2008)
(Accepted October 8, 2008)

(Editor in Charge: *Makoto Iwayama*)

Chirag Shah is a doctoral student at University of North Carolina (UNC) at Chapel Hill, USA working with Gary Marchionini and Diane Kelly. He obtained his MTech in Computer Science from Indian Institute of Technology (IIT) Madras, India and MS in Computer Science from University of Massachusetts (UMass) at Amherst, USA where he worked with Bruce Croft, James Allen, and David Jensen. The work reported in this paper was done during his internship with Koji Eguchi at National Institute of Informatics (NII) in Tokyo, Japan. His current research interests are collaborative IR, interactive IR systems, user feedback, and Web crawling and archiving.

Koji Eguchi is an Associate Professor in the Department of Computer Science and Systems Engineering, Kobe University, Japan, and a Visiting Associate Professor at the National Institute of Informatics (NII), Japan. His research interests include information retrieval, web computing and data mining.