PageRank with Text Similarity and Video Near-Duplicate Constraints for News Story Re-ranking

Xiaomeng Wu[1], Ichiro Ide[2], and Shin'ichi Satoh[1]

[1] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku Tokyo 101-8430, Japan
{wxmeng,satoh}@nii.ac.jp
[2] Graduate School of I.S., Nagoya University
Furo-cho, Chikusa-ku Nagoya 464-8630, Japan
ide@is.nagoya-u.ac.jp

Abstract. Pseudo-relevance feedback is a popular and widely accepted query reformulation strategy for document retrieval and re-ranking. However, problems arise in this task when assumed-to-be relevant documents are actually irrelevant which causes a drift in the focus of the reformulated query. This paper focuses on news story retrieval and re-ranking, and offers a new perspective through the exploration of the pair-wise constraints derived from video near-duplicates for constraint-driven re-ranking. We propose a novel application of PageRank, which is a pseudo-relevance feedback algorithm, and use the constraints built on top of text to improve the relevance quality. Real-time experiments were conducted using a large-scale broadcast video database that contains more than 34,000 news stories.

Keywords: PageRank, Video Near Duplicate, News Story Re-ranking, Video Data Mining.

1 Introduction

News videos are broadcast everyday across different sources and times. To make full use of the overwhelming volume of news videos available today, it is necessary to track the development of news stories from different sources, mine their dependencies, and organize them in a semantic way. News story retrieval is a fundamental step for news topic tracking, threading, video summarization, and browsing from among these research efforts. News story retrieval aims at searching for evolving and historical news stories according to the topics, such as *the Trial of Saddam Hussein* and *2006 North Korean nuclear test.*

1.1 Background

News story retrieval is normally studied under the Query by Example theme using the textual features as the underlying cues [3,4,11]. Relevance feedback is a

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 533–544, 2010.

popular and widely accepted query reformulation strategy for this task. Some researchers have attempted to automate the manual part of relevance feedback, which is also known as pseudo-relevance feedback [1,5,10]. Pseudo-relevance feedback is obtained by assuming that the top k documents in the resulting set containing n results (usually where $k \ll n$) are relevant, and has the advantage in that assessors are not required. However, in pseudo-relevance feedback, problems arise when assumed-to-be relevant documents are actually irrelevant, which causes a drift in the focus of the reformulated query [10]. How to reduce the inappropriate feedback taken from irrelevant documents or how to guarantee the relevance quality is the main focused issue for pseudo-relevance feedback studies.

To tackle this issue, we offer a new perspective that explores the pairwise constraints derived from video near duplicates for news story retrieval and constraint-driven re-ranking. The main points of discussion include: (1) a novel scheme for pseudo-relevance feedback on the basis of near duplicates built on top of text, (2) a novel application of PageRank as a pseudo-relevance feedback algorithm used for constraint-driven re-ranking, and (3) real-time experiments conducted on a large-scale broadcast video database containing more than 34,000 news stories.

1.2 Framework Overview

Our system works on a large-scale broadcast video database. The system uses a news story as the search query and outputs stories depicting the query topic from within the database. A news story is formally defined as a semantic segment within a news video, which contains a report depicting a specific topic or incident. A story is described as a group of shots. Each shot is described by using a set of representative keyframes and closed-captions. Figure 1 depicts our proposed news story retrieval and re-ranking system.

Initially, candidate news stories that are similar to the query are searched using a topic-tracking method based only on the textual information. The reason

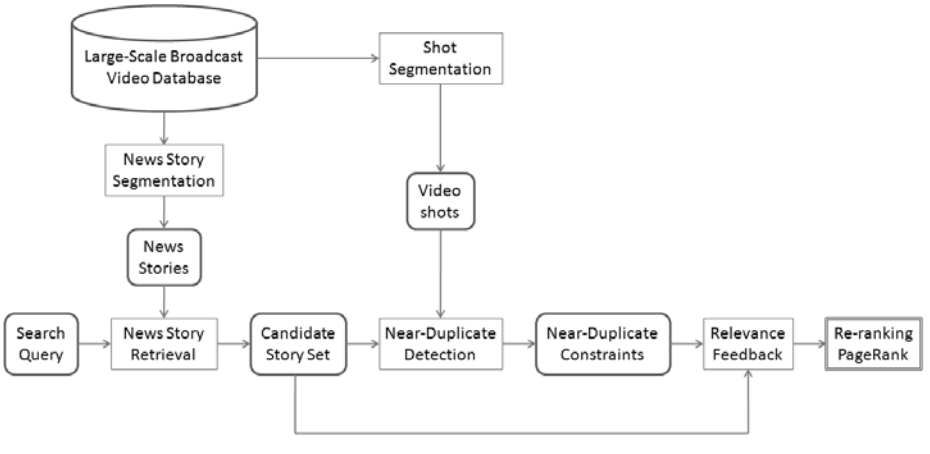

Fig. 1. Proposed news story retrieval and re-ranking system

that we use a text-based method first before using the near-duplicate detection is because the latter requires visual information processing and is computationally far more expensive than textual information processing. On the other hand, the coverage of near duplicates is normally insufficient for conducting a complete and thorough topic tracking compared that of the textual information. After text-based topic tracking, near duplicates are detected from the set of candidate news stories and used to group together the stories that share these near duplicates. We believe that video near-duplicate constraints are highly-useful for guaranteeing a higher relevance quality than the textual constraints. PageRank is then applied as a pseudo-relevance feedback algorithm on the basis of these pairwise constraints, and used to re-rank news stories depicting the same topic.

2 News Story Retrieval Based on Textual Information

News story retrieval is normally studied under the Query by Example theme with the textual features used as the underlying cues [3,4,11]. In news videos, the focal point or content of evolving stories depicting the same topic normally varies slowly with time. The news story retrieval method used in this work should be robust enough for this focal-point variation. In this paper, [3] is used for this purpose, which uses the semantic and chronological relation to track the chain of related news stories in the same topic over time.

A story boundary is first detected from a closed-caption text broadcasted simultaneously with the video. The resemblances between all story combinations are evaluated by adopting a cosine similarity between two keyword frequency vectors generated from two news stories. When the resemblance exceeds a threshold, the stories are considered related and linked. Tracking is achieved by considering the children stories related to the search query as new queries to search for new children stories. This procedure forms a simple story link tree starting from the story of interest, i.e. the search query. Children stories are defined as news stories related to a parent, under the condition that the time stamps of the children stories always chronologically succeed their parent. The link tree can also be considered a set of candidate news stories that is similar to the search query, which is further used for near-duplicate detection.

3 Video Near-Duplicate Detection

In addition to the textual features, in broadcast videos, there are a number of video near duplicates, which appear at different times and dates and across various broadcast sources. Near duplicates, by definition, are sets of shots composed of the same video material used several times from different sources or material involving the same event or the same scene, as shown in Fig. 2. These near duplicates basically form pairwise equivalent constraints that are useful for bridging evolving news stories across time and sources.

After text-based news story retrieval, near duplicates are detected only from the set of candidate news stories. This can not only dramatically reduce the

Fig. 2. Near duplicates across different stories of two topics. The label under each image is the program name and the airdate. Above: *Trial of Saddam Hussein*. Below: *2006 North Korean nuclear test*.

computation burden due to visual information processing, but also reduce the probability of potential errors caused by near-duplicate detection. We used an interest-point-based algorithm with a local description for the near-duplicate detection. This algorithm was proposed by Ngo et al. [9] and proved to be robust to variations of translation and scaling introduced due to video editing and different camerawork.

They tested their algorithm using a keyframe database instead of a video archive [9]. We extract multiple keyframes from each video shot to extend it to near-duplicate shot detection. The shot length is equally divided, and the frames at the points of division are selected as the keyframe. This is to tolerate the variation introduced by the camera and object motion. In equation terms, given the shot length L, the $(i \times L/(N+1))^{\text{th}}$ frames are extracted as the keyframe, where $i = 1 \cdots N$. N indicates the number of keyframes and is empirically set to three in this paper. To tolerate the significant impact of video captions, we propose cropping the keyframe beforehand so that only the central part is used for the near-duplicate detection. On the other hand, we also manually excluded anchorperson shots that are not related to the topic while highly possible to be detected as near duplicates. Since anchorperson-shot detection has been extensively studied and many good algorithms were already proposed, this process can be automated if needed.

4 PageRank with Near-Duplicate Constraints

4.1 PageRank

Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. PageRank [2] is a variant of the Eigenvector centrality measure. PageRank (**PR**) is iteratively defined as

$$\mathbf{PR} = d \times \mathbf{S}^* \times \mathbf{PR} + (1 - d) \times \mathbf{p} \quad (1)$$

$\mathbf{S}^*$ is the column normalized adjacency matrix of $\mathbf{S}$, where $\mathbf{S}_{i,j}$ measures the weight between node i and j. d is a damping factor to be empirically determined, and $\mathbf{p} = [\frac{1}{n}]_{n\times 1}$ is a uniform damping vector, where n is the total number of nodes in the network. Repeatedly multiplying $\mathbf{PR}$ by $\mathbf{S}^*$ yields the dominant eigenvector of the matrix $\mathbf{S}^*$. Although $\mathbf{PR}$ has a fixed-point solution, in practice, it can often be estimated more efficiently using iterative approaches. PageRank converges only when matrix $\mathbf{S}^*$ is aperiodic and irreducible. The former is generally true for most applications and the latter usually requires a strongly connected network, a property guaranteed in practice by introducing a damping factor d. It is generally assumed that the damping factor is set around 0.85.

The simplicity and effectiveness of PageRank for text mining were demonstrated through document summarization studies [12,14], which suggested combining PageRank with text similarity. However, except the study [7], little attention has been paid to applying PageRank to the document retrieval task. This is because PageRank usually requires a strongly connected network where the group of good nodes should comprise a majority of all the nodes in this network. This fundamental assumption is feasible for document summarization because people normally seek to summarize a set of documents, all of which are strongly related to a certain topic. However, for document retrieval, the set of documents searched for by the query normally contains errors that are irrelevant to the topic of interest. The potential existence of these assumed-to-be good nodes that are actually bad nodes will have a large impact on the majority distribution of the network, and thus, they cause a drift in the focus of PageRank. This issue is similar to the problem existing in pseudo-relevance feedback studies that were introduced in Sec. 1.1.

4.2 Near-Duplicate Constraints

A variety of algorithms have been proposed for applying near-duplicate constraints to news video retrieval tasks. Zhai et al. [6] linked news stories by combining image matching and textual correlation. Hsu et al. [8] tracked four topics with near duplicates and semantic concepts, and found that near duplicates significantly improve the tracking performance. These two works use video near duplicate and textual information as two independent modalities. Since each modality is individually processed and fusion is based only on the score functions of their processing results, the potential inter-modal relationships between the two modalities have not been thoroughly explored and thus are wasted. Apart from these multimodality fusion studies, Wu et al. [13] presented a system built on near-duplicate constraints, which are applied on top of text to improve the story clustering. This work depends on manual near-duplicate labeling, which is impossible to handle within large-scale databases.

In this paper, we apply PageRank as a pseudo-relevance feedback algorithm and integrate video near-duplicate constraints to guarantee the relevance quality. Given a story used as a search query, candidate stories similar to the query are

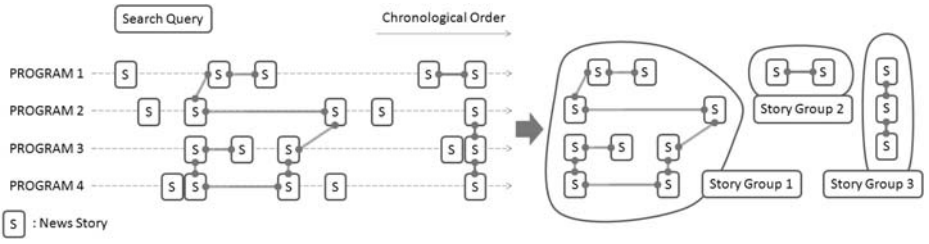

Fig. 3. Group stories sharing near duplicates

searched across various news programs (Fig. 3). Two stories are linked together if they share at least one pair of near duplicates. This kind of two stories can be regarded as having a must-link constraint, which indicates that they discuss the same topic. Stories are then clustered into groups based on these links. We make the following two assumptions.

- **Assumption 1**
  Most stories in the same story group depict the same topic.
- **Assumption 2**
  The largest story group depicts the same topic as the query.

Assumption 1 is feasible in most cases because near duplicates are detected only from the set of candidate stories that are similar to the query, so that the probability of potential errors caused by near-duplicate detection is small. **Assumption 2** is also feasible because most near duplicates are shared between stories depicting the same topic as the query. From another point of view, the noise or outlier topics are normally different from each other so that fewer near duplicates are shared between them.

The experimental results from news story grouping based on ten search queries are listed in Table 1 (Sec. 5.1 for more information on these queries). From #Story (#TP), we can see that most news stories clustered in the largest group depict the same topic as the search query (except for T6). In other words, both **Assumption 1** and **Assumption 2** described above were feasible in our experiment. For T6, the news topic is on a child abduction-murder that occurred in Hiroshima, and the query story was broadcasted on 2005/12/01. The next day, another child abduction-murder was reported near Tokyo. In most news programs, these two news topics were continuously broadcasted. Due to the high similarity between them, the story segmentation method [3] used in this paper failed to segment them from each other, so that the stories of the query topic also contained shots of the noise topic. As a result, stories depicting these two topics were clustered in the same story group based on the near-duplicate constraints.

Given both **Assumption 1** and **Assumption 2** are feasible, it is reasonable to assume that stories in the largest story group would satisfy the PageRank requirement in a strongly connected network where the group of good nodes comprises a majority among all the nodes in this network (Sec. 4.1). Based on

Table 1. Experimental results of news story grouping. #Candidate: number of stories in the candidate story set. #TP: number of relevant stories depicting the same topic as the search query. #Group: number of clustered story groups based on near-duplicate constraints. #Story: number of stories in the largest story group.

Topic Number	#Candidate (#TP)	#Group	#Story (#TP)
T1	106 (43)	11	9 (9)
T2	370 (174)	3	72 (72)
T3	164 (115)	3	75 (71)
T4	35 (13)	3	10 (10)
T5	93 (64)	1	47 (47)
T6	148 (25)	7	28 (4)
T7	119 (60)	7	26 (22)
T8	35 (34)	3	13 (13)
T9	48 (44)	2	30 (30)
T10	65 (54)	3	39 (38)

Fig. 4. Examples of near duplicates depicting *Trial of Saddam Hussein*. #story: number of stories sharing the corresponding near duplicate.

these assumptions, the largest story group is chosen as the relevance feedback. For example, story group 1 shown in Fig. 3 is the largest story group, so it will be used as the relevance feedback and further integrated with PageRank. Examples of near duplicates within the largest story group depicting *the Trial of Saddam Hussein* are shown in Fig. 4.

4.3 Applying PageRank

To define the text similarity, we use term frequency-inverse document frequency (*tf-idf*) weighting, which is one of the best-known schemes for text mining, to represent each story. To do so, a semantic analysis is first applied to the compound nouns extracted from each story to generate a keyword vector for four semantic classes, *general*, *personal*, *locational/organizational*, and *temporal*. From our experimental results, we found that the keywords from the *temporal* class are normally not helpful for identifying news stories depicting the same topic. Therefore, only the compound nouns of the other three classes are used as the keywords.

For each story, a keyword vector $\boldsymbol{V}$ can be created as follows. The keyword similarity between two stories is thus defined by the cosine similarity shown in Eq. 3.

$$\boldsymbol{V} = (tfidf(t_1), tfidf(t_2), \cdots, tfidf(t_N)) \tag{2}$$

$$\cos(\boldsymbol{V}_i, \boldsymbol{V}_j) = \frac{\boldsymbol{V}_i \cdot \boldsymbol{V}_j}{\|\boldsymbol{V}_i\|\|\boldsymbol{V}_j\|} \tag{3}$$

The set of stories may be represented by a cosine similarity matrix $\mathbf{S}$, where each entry $\mathbf{S}_{i,j} = \cos(\boldsymbol{V}_i, \boldsymbol{V}_j)$ in the matrix is the similarity between the corresponding story pair (s_i, s_j). Different from document summarization studies where sentences are regarded as nodes, we regard one document or one story as one node. Thus, we have the overall relevancy ($\mathbf{PR}$) of each story given its similarity to other stories, iteratively defined by using PageRank (Eq. 1).

This is equivalent to using all the stories searched for by the query as relevance feedback, which is highly sensitive to the potential existence of irrelevant stories. To guarantee the relevance quality, we put restrictions on the adjacency matrix $\mathbf{S}$ based on the video near-duplicate constraints. In Sec. 4.2, stories are clustered into groups based on the near-duplicate constraints. Given the largest story group being denoted by $\mathbb{S}$, each entry in the adjacency matrix $\mathbf{S}$ is defined by using Eq. 4. In other words, we only regard stories in $\mathbb{S}$ as high-quality nodes, and for each node s_j in the whole network, we only use the connection between s_j and these high-quality nodes $\{s_i : s_i \in \mathbb{S}\}$ as a vote of support to iteratively define the overall relevancy of s_j.

$$\mathbf{S}_{i,j} = \begin{cases} \cos(\boldsymbol{V}_i, \boldsymbol{V}_j) & (s_i \in \mathbb{S}) \\ 0 & (s_i \notin \mathbb{S}) \end{cases} \tag{4}$$

5 Experiments

5.1 Database

We tested our system using a large-scale broadcast video database comprised of actually broadcasted videos from 2005/10/19 to 2007/01/19. These videos were broadcasted from six different news programs produced by three different Japanese TV stations. Closed-captions were segmented into stories using the algorithm developed by Ide et al. [3], and the videos were segmented into shots by comparing the RGB histograms between adjacent frames. The stories and shots were used as the basic units of analysis. The keywords were derived from a list of compound nouns extracted from the closed-captions [3], while the keyframes were derived by equally dividing the shot and selecting the points of division. The set of near-duplicate pairs was detected using the algorithm developed by Ngo et al. [9]. The database was comprised of 34,279 news stories (compared to around 800 news stories used by Wu et al. [13]).

Ten search queries were selected for experimentation, as listed in Table 2, including five Japanese and five foreign news stories. The design of these queries is based on the biggest topics of the important domestic and international news stories from 2005/10/19 to 2007/01/19. The candidate stories that were similar to these queries were searched for across the six news programs, and the near duplicates were detected from the set of candidate news stories. The duration

Table 2. Ten search queries selected for experimentation

Topic Number	Topic	Duration	Domestic / Foreign
T1	*Trial of Saddam Hussein (1)*	15 months	Foreign
T2	*Architectural forgery in Japan*	2 months	Domestic
T3	*Fraud allegations of Livedoor*	2 months	Domestic
T4	*Trial of Saddam Hussein (2)*	2 months	Foreign
T5	*7 July 2005 London bombings*	1 month	Foreign
T6	*Murder of Airi Kinoshita*	1 month	Domestic
T7	*Murder of Yuki Yoshida*	1 month	Domestic
T8	*Murder of Goken Yaneyama*	1 month	Domestic
T9	*2006 North Korean missile test*	1 month	Foreign
T10	*2006 North Korean nuclear test*	1 month	Foreign

within which the search was conducted varied from 1 to 15 months. Our experiments on news story re-ranking were conducted based on our proposed algorithm discussed in Sec. 4.

5.2 News Story Re-ranking

The stories searched for using the ten queries were ranked based on PageRank with video near-duplicate constraints. To evaluate the performance, we compared our algorithm with the four baseline algorithms listed as follows. Note that **BL2**, **BL3**, and **BL4** can also be considered traditional algorithms under the pseudo-relevance feedback theme.

- **Baseline 1 (BL1)**
  The cosine similarity between each news story and the original query is evaluated and used for the story ranking.

- **Baseline 2 (BL2)**
  After ranking the stories using **BL1**, the top-k ($k = 10, 20, 30$) stories were chosen as the relevance feedback. The Rocchio algorithm [1], which is a classic algorithm for extracting information from relevance feedback, is used to re-rank the stories. The Rocchio algorithm formula is defined by using Eq. 5, where $\boldsymbol{V}_m$ is the expanded query, $\boldsymbol{V}_0$ is the original query, $\mathbb{D}_{rel}$ and $\mathbb{D}_{irr}$ are the sets of relevant and irrelevant stories, respectively, and $\alpha = 1$, $\beta = 1$, and γ are weights. In this baseline, we allow only positive feedback, which is equivalent to setting $\gamma = 0$. Note that $|\mathbb{D}_{rel}| = k$. The cosine similarity between each news story and the expanded query is then evaluated and used for story re-ranking.

$$\boldsymbol{V}_m = \alpha \times \boldsymbol{V}_0 + \beta \times \frac{\sum_{\boldsymbol{V}_i \in \mathbb{D}_{rel}} \boldsymbol{V}_i}{|\mathbb{D}_{rel}|} - \gamma \times \frac{\sum_{\boldsymbol{V}_i \in \mathbb{D}_{irr}} \boldsymbol{V}_i}{|\mathbb{D}_{irr}|} \tag{5}$$

- **Baseline 3 (BL3)**
  Apply PageRank to all stories searched for by the query.

– **Baseline 4 (BL4)**
After ranking the stories using **BL1**, the top-k ($k = 10, 20, 30$) stories are chosen as the relevance feedback. PageRank is used for the story re-ranking. Each entry in the adjacency matrix $\mathbf{S}$ is defined by using Eq. 6, where $\mathbb{D}_{rel}$ is the set of top-k stories.

$$\mathbf{S}_{i,j} = \begin{cases} \cos(\boldsymbol{V}_i, \boldsymbol{V}_j) & (s_i \in \mathbb{D}_{rel}) \\ 0 & (s_i \notin \mathbb{D}_{rel}) \end{cases} \tag{6}$$

5.3 Experimental Results and Discussions

An evaluation using the average precision ($AveP$) was performed using Eq. 7. In Eq. 7, r denotes the rank, N the number of stories searched, $rel()$ a binary function on the relevance of a given rank, and $P()$ the precision at a given cut-off rank. N_{rel} denotes the number of relevant stories with $N_{rel} \leq N$. Table 3 lists the results, where PRND denotes our proposed algorithm using PageRank with Near-Duplicate constraints and MAP is the Mean Average Precision.

$$AveP = \frac{\sum_{r=1}^{N}(P(r) \times rel(r))}{N_{rel}} \tag{7}$$

Table 3. Experimental results of story re-ranking ($AveP$: %)

	BL1	BL2_10	BL2_20	BL2_30	BL3	BL4_10	BL4_20	BL4_30	**PRND**
T1	82.6	69.92	67.58	80.37	65.99	69.73	69.22	82.02	**96.38**
T2	87.92	87.47	91.35	93	67.61	92.06	92.41	94.35	**97.45**
T3	88.4	96.92	97.05	97	97.33	97.89	97.54	97.78	**98.36**
T4	81.68	94.85	90.79	90.99	85.93	95.3	80.35	62.91	**100**
T5	93.46	97.37	97.67	97.63	96.08	97.11	98.12	97.95	**99.42**
T7	85.36	96.86	98.12	98.15	93.91	98.21	99.09	99.3	**99.42**
T8	99.46	99.83	**100**	**100**	**100**	**100**	**100**	**100**	**100**
T9	98.28	98.62	98.74	98.64	98.62	**98.97**	98.85	98.74	98.87
T10	94.92	97.09	97.22	97.13	95.13	97.27	**97.58**	96.95	97.34
MAP	89.87	92.55	93.63	95.01	85.81	94.22	94.11	95.2	**98.25**

From Table 3, we can see that our proposed re-ranking algorithm outperformed the baselines for most topics, and the MAP of PRND is higher than all baselines. The main reason for this is that our re-ranking algorithm based on PageRank improves the informativeness and representativeness of the original query (**BL1**), and near-duplicate constraints guarantee a higher relevance quality than the textual constraints used in traditional pseudo-relevance feedback algorithms (**BL2**, **BL3**, and **BL4**).

In particular, for T1, T2, and T4, we can see they had a higher level of improvement than the other topics. By observing the results, we found that ranking based on the original query tends to associate higher rank to stories closer to the query in terms of their airdates. Therefore, stories temporally closer

to the query (normally within one week in our experiments) tend to comprise the majority from among the top-k stories chosen as the relevance feedback in **BL2** and **BL4**. This is because the focal point of the evolving stories depicting the same topic normally varies over time, so the stories temporally distant from the query tend to have less similarity to the query. For long topics like T1, T2, and T4, stories that are relevant but temporally distant from the query tend to acquire less contribution or a lower vote of support from the top-k relevant stories. On the other hand, evolving stories always repeatedly use the same representative shots, even if their airdates are distant from each other. Figure 2 shows two examples, from which we can see that the detected near-duplicates guaranteed a larger coverage of evolving stories in terms of airdate. This leads to increased coverage of textual information for the relevance feedback when applying PageRank based on these near-duplicate constraints. This can explain the greater improvement of T1, T2, and T4 compared to the other topics. This is also considered one of the contributions of our proposed re-ranking algorithm. For another long topic T3, because the focal-point variation is small in this case, the problem described above was not reflected in this experiment.

6 Conclusion

This paper focuses on news story retrieval and re-ranking, and offers a new perspective by exploring the pairwise constraints derived from video near duplicates for constraint-driven re-ranking. Compared to some other similar works, we use the constraints built on top of text to improve the relevance quality. As a future work, an experiment is under development for evaluating the actual time necessary to process the dataset during each phase of our proposed system. Also, we are developing an experiment for integrating the near-duplicate constraints into pseudo-relevance feedback, and comparing it to PageRank. Another future work is quantitatively evaluating the performance of near-duplicate detection algorithm, and checking up on whether the false alarms will have a large impact on the proposed system.

References

1. Rocchio, J.: Relevance Feedback in Information Retrieval. The SMART Retrieval System (1971)
2. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks 30, 107–117 (1998)
3. Ide, I., Mo, H., Katayama, N., Satoh, S.: Topic Threading for Structuring a Large-Scale News Video Archive. In: Enser, P.G.B., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 123–131. Springer, Heidelberg (2004)
4. Mo, H., Yamagishi, F., Ide, I., Katayama, N., Satoh, S., Sakauchi, M.: Key Image Extraction from a News Video Archive for Visualizing Its Semantic Structure. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 650–657. Springer, Heidelberg (2004)

5. Qin, Z., Liu, L., Zhang, S.: Mining Term Association Rules for Heuristic Query Construction. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 145–154. Springer, Heidelberg (2004)
6. Zhai, Y., Shah, M.: Tracking news stories across different sources. In: ACM Multimedia, pp. 2–10 (2005)
7. Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., Ma, W.-Y.: Improving web search results using affinity graph. In: SIGIR, pp. 504–511 (2005)
8. Hsu, W.H., Chang, S.-F.: Topic Tracking Across Broadcast News Videos with Visual Duplicates and Semantic Concepts. In: ICIP, pp. 141–144 (2006)
9. Ngo, C.-W., Zhao, W., Jiang, Y.-G.: Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation. In: ACM Multimedia, pp. 845–854 (2006)
10. Song, M., Song, I.-Y., Hu, X., Allen, R.B.: Integration of association rules and ontologies for semantic query expansion. Data Knowl. Eng. 63, 63–75 (2007)
11. Lin, F., Liang, C.-H.: Storyline-based summarization for news topic retrospection. Decis. Support Syst. 45, 473–490 (2008)
12. Wan, X., Yang, J.: Multi-document summarization using cluster-based link analysis. In: SIGIR, pp. 299–306 (2008)
13. Wu, X., Ngo, C.-W., Hauptmann, A.G.: Multimodal News Story Clustering With Pairwise Visual Near-Duplicate Constraint. IEEE Transactions on Multimedia 10, 188–199 (2008)
14. Otterbacher, J., Erkan, G., Radev, D.R.: Biased LexRank: Passage retrieval using random walks with question-based priors. Inf. Process. Manage. 45, 42–54 (2009)