

# Manifold Ranking with Sink Points for Update Summarization

Pan Du, Jiafeng Guo, Jin Zhang, Xueqi Cheng  
Institute of Computing Technology, CAS  
Beijing 100190, China  
{dupan,guojiafeng,zhangjin}@software.ict.ac.cn, cxq@ict.ac.cn

## ABSTRACT

Update summarization aims to create a summary over a topic-related multi-document dataset based on the assumption that the user has already read a set of earlier documents of the same topic. Beyond the problems (i.e., topic relevance, salience, and diversity in extracted information) tackled by topic-focused multi-document summarization, the update summarization must address the novelty problem as well. In this paper, we propose a novel extractive approach based on manifold ranking with sink points for update summarization. Specifically, our approach leverages a manifold ranking process over the sentence manifold to find topic relevant and salient sentences. More important, by introducing the sink points into sentence manifold, the ranking process can further capture the novelty and diversity based on the intrinsic sentence manifold. Therefore, we are able to address the four challenging problems above for update summarization in a unified way. Experiments on benchmarks of TAC are performed and the evaluation results show that our approach can achieve comparative performance to the existing best performing systems in TAC tasks.

## Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*abstracting methods*; I.2.7 [Artificial Intelligence]: Natural Language Processing—*text analysis*

## General Terms

Algorithms, Performance, Experimentation

## Keywords

Update Summarization, Multi-document Summarization, Manifold Ranking with Sink Points

## 1. INTRODUCTION

There has been increasing interests in text summarization with the massive explosion in the amount of data on

the Web. As the real time data becomes the trends, it is important that the summarization can take the temporal dimension into account so that the obsolete information which has been presented to users in the past can be removed in the summary. Update summarization aims to write a short summary over a set of topic-related multi-document dataset, under the assumption that the user has already read a given set of earlier documents of the same topic. Give this definition, we can find that there are four major problems that update summarization need to address: 1) *Topic Relevance*: The summary is based on a topic-related multi-document dataset, where a topic here represents user's information need. Therefore, the summary must stick to the topic users are interested in. 2) *Salience*: Not all the sentences in documents deliver information of equal importance about the topic. The summary has to neglect those trivial content because of the length limitation, and keep the salient information as much as possible. 3) *Diversity*: There should be less redundant information in the summary. In this way, we can fully leverage the limited summary space to cover as much information as possible about the topic. 4) *Novelty*: Given a specified topic and two chronologically ordered document datasets, the summary need to focus on the new information conveyed by the later dataset as compared with the earlier one under that topic.

Update summarization has attracted a lot of research interests recently. Most previous work can be categorized as extractive approach, which composes a summary by extracting sentences from target document dataset. Different ways were introduced in these approaches to address the novelty problem, including time decay factor [1, 5], sentence filtering [7], topic analysis [4] and negative influence diffusion [3]. However, most of these approaches [7, 3, 4, 5] did not address the four problems for update summarization simultaneously. In contrary, they need some additional steps to address diversity or novelty with some heuristics.

In this paper, we propose a novel extractive approach based on manifold ranking with sink points for update summarization. This approach can iteratively extract sentences to form an update summary by simultaneously addressing the four problems above in a unified way. Specifically, our approach leverages a manifold ranking process [8] over sentence manifold, which can help find topic relevant and salient sentences. More important, we introduce the sink points into the sentence manifold, which denote the sentences whose ranking scores are fixed to the minimum ranking score during the manifold ranking process. Therefore, the ranking scores of other sentences close to the sink points will be nat-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.  
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...\$10.00.

urally penalized during the ranking process based on the intrinsic sentence manifold. By turning both the sentences from the earlier document dataset and that already selected for summary into sink points, we can capture both novelty and diversity during the ranking process. In this way, we are able to address the four problems for update summarization simultaneously in a unified way. We conduct empirical experiments based on the benchmark datasets of TAC2008 and TAC2009. The ROUGE evaluation results show that our approach can achieve comparative performance to the existing best performing systems in TAC tasks and significantly outperform other baseline methods.

## 2. OUR APPROACH

Neither novelty nor diversity required by updated summarization can be addressed by the traditional manifold ranking process. Work [6] leveraged an additional step with a greedy algorithm to address the diversity problem. Unlike their approach, here we introduce the sink points into the sentence manifold to help capture both novelty and diversity during the ranking process. The sink points are sentences whose ranking scores are fixed to the minimum ranking score (i.e. zero in our case) on the manifold during the ranking process. Intuitively, we can imagine the sink points as the "black holes" on the manifold, where ranking scores spreading to them will be "absorbed" and no ranking scores would be able to "escape" from them. In this way, the ranking scores of other sentences close to the sink points (i.e. sentences sharing similar information with the sink points) will be naturally penalized during the ranking process based on the intrinsic sentence manifold.

As we know, novelty focuses on the new information conveyed by the later dataset as compared with the earlier one under the given topic. Thus, we need to penalize the sentences in the later dataset which convey similar information as the earlier dataset. While diversity focuses on the redundant information among the selected sentences in a summary. Thus we need to penalize the sentences conveying similar information as the sentences already in a summary. Therefore, by turning both the sentences from the earlier dataset and that already selected for summary into sink points, we can capture both novelty and diversity similarly based on the intrinsic sentence manifold during the ranking process. In this way, we address the four problems for update summarization in a unified way.

Formally, let  $\chi = \{x_0, x_1, \dots, x_s, x_{s+1}, \dots, x_n\} \subset R^m$  denote a set of data points over the sentence manifold (each point represents one sentence), where point  $x_0$  denotes the pseudo sentence of topic description,  $x_1, \dots, x_s$  denote sentences set as sink points, and  $x_{s+1}, \dots, x_n$  denote the rest free sentences to be ranked, named *free points*. Hereafter, sentence and point will not be discriminated unless otherwise specified. Let  $f : \chi \rightarrow R$  denote a ranking function which assigns a ranking value  $f_i$  to each point  $x_i$ . We can view  $f$  as a vector  $f = [f_0, f_1, \dots, f_s, f_{s+1}, \dots, f_n]^T$ . We also define a vector  $y = [y_0, y_1, \dots, y_s, y_{s+1}, \dots, y_n]^T$ , in which  $y_0 = 1$  since  $x_0$  is the topic sentence, and  $y_i = 0$  ( $1 \leq i \leq n$ ) for all the other sentences. The algorithm of update summarization based on manifold ranking with sink points is as follows:

1. Form the affinity matrix  $W$  for the sentence manifold, where  $W_{ij} = \text{sim}(x_i, x_j)$  if there is an edge linking  $x_i$

and  $x_j$ . Note that  $\text{sim}(x_i, x_j)$  is defined by the cosine similarity between sentences in our approach.

2. Construct the Matrix  $S = D^{-1/2}WD^{-1/2}$  in which  $D$  is a diagonal matrix with its  $(i, i)$ -element equal to the sum of the  $i$ -th row of  $W$ .
3. Iterate  $f(t+1) = \alpha SI_f f(t) + (1-\alpha)y$  until convergence, where  $0 \leq \alpha < 1$ , and  $I_f$  is an indicator matrix which is a diagonal matrix with its  $(i, i)$ -element equal to 0 if  $x_i$  is a sink point and 1 otherwise.
4. Let  $f_i^*$  denote the limit of the sequence  $\{f_i(t)\}$ . Each sentence  $x_i$  ( $s+1 \leq i \leq n$ ) gets its ranking scores  $f_i^*$ .
5. Select the sentence  $x_m$  with maximum score  $f_m^*$  as a summary sentence, set the corresponding diagonal element of  $I_f$  as 0.
6. If the summary length is not reached, go to step 3.

As we can see, the major difference between our algorithm (i.e. manifold ranking with sink points) and traditional manifold ranking algorithm is that we introduce the indicator matrix  $I_f$  into the iteration function in step 3. This indicator matrix  $I_f$  is used to fix the ranking scores of sink points at the minimum (i.e., zero in our case) during the iteration. As a result, the sink points will not spread any ranking score to their neighbors during the ranking process. It is important to know that, with the indicator matrix introduced, the new iteration algorithm can still achieve a global stable state.

**THEOREM 1.** *The sequence  $\{f(t)\}$  converges to*

$$f^* = (1 - \alpha)(I - \alpha SI_f)^{-1}y. \quad (1)$$

**PROOF.** According to iteration equation

$$f(t+1) = \alpha SI_f f(t) + (1 - \alpha)y.$$

used in the algorithm, we have

$$f(t) = (\alpha SI_f)^t f(0) + (1 - \alpha) \sum_{i=0}^{t-1} (\alpha SI_f)^i y.$$

Let  $\tilde{P} = D^{-1}WI_f$ ,  $\tilde{P}$  is the similarity transformation of  $SI_f$  as follows:

$$\begin{aligned} SI_f &= D^{1/2}D^{-1}WD^{-1/2}D^{1/2}I_fD^{-1/2} \\ &= D^{1/2}D^{-1}WI_fD^{-1/2} \\ &= D^{1/2}\tilde{P}D^{-1/2} \end{aligned}$$

hence  $\tilde{P}$  and  $SI_f$  have the same eigenvalues  $\lambda$ .

Note that  $|\tilde{P}_{ii}| = 0$ , according to Gershgorin circle theorem, we have

$$|\lambda| \leq \sum_{j \neq i} |\tilde{P}_{ij}| \leq 1$$

where  $|\lambda|$  represents the norm of any eigenvalue of  $\tilde{P}$ .

Since  $0 \leq \alpha < 1$ , and  $|\lambda| \leq 1$ ,

$$\begin{aligned} \lim_{t \rightarrow \infty} (\alpha SI_f)^t &= 0, \\ \lim_{t \rightarrow \infty} \sum_{i=0}^{t-1} (\alpha SI_f)^i &= (I - \alpha SI_f)^{-1}, \end{aligned}$$

Hence,

$$f^* = (1 - \alpha)(I - \alpha SI_f)^{-1}y.$$

□

We can use this closed form to compute the ranking scores of sentences directly. In large-scale real-world problems, however, an iteration algorithm is preferred.

### 3. EXPERIMENTS

#### 3.1 Data Sets

Update summarization has been one of the main tasks in TAC2008 and TAC2009 hold by NIST<sup>1</sup>. They have devoted a lot of manual labor to create the benchmark data for update summarization tasks. TAC2008 provided 48 topics and TAC2009 provided 44 topics. Each topic was composed of 20 relevant documents from the AQUAINT-2 collection of news articles, and the documents were divided into 2 Datasets: Document Set A and Document Set B. Each document set had 10 documents, and all the documents in set A chronologically preceded the documents in set B. In TAC task, a 100-word summary was required to be generated for each set of documents. The summary of Set B should be written under the assumption that the user has already read the content of set A and should inform the user of new information about the given topic.

#### 3.2 Evaluation Metric

ROUGE<sup>2</sup> has become the most frequently used toolkit for automatic summarization evaluation, as it produces the most reliable scores in correspondence with human evaluations. It measures summary quality by counting the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary and the ideal summaries created by humans. The n-gram recall measure, ROUGE-N, is computed as follows:

$$ROUGE - N = \frac{\sum_{S \in \{Refs\}} \sum_{gram_n \in S} Cnt_{match}(gram_n)}{\sum_{S \in \{Refs\}} \sum_{gram_n \in S} Cnt(gram_n)}.$$

where  $n$  stands for the length of the n-gram,  $Cnt_{match}(gram_n)$  is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries  $Refs$ , and  $Cnt(gram_n)$  is the number of n-gram in the reference summaries. In evaluation, we use the ROUGE-2 (bigram-based) and ROUGE-SU4 (an extended version of ROUGE-2) automatic metrics. The results were obtained with ROUGE version 1.5.5 with the settings used for TAC2008.

#### 3.3 Baseline Methods

For evaluation, we compared our approach with several baseline methods. One category of baseline methods consists of the top performing systems on update summarization tasks according to the ROUGE-2 metric on TAC2008 and TAC2009. Besides, we also leverage three other baseline systems, named Baseline-L, Baseline-U, and Baseline-MR, for comparison.

<sup>1</sup><http://www.nist.gov>

<sup>2</sup><http://haydn.isi.edu/ROUGE/>

Baseline-L and Baseline-U are two standard baseline methods provided by NIST on TAC. Baseline-L takes all the leading sentences (up to 100 words) in the most recent document. It provides a lower bound on what can be achieved with those extractive summarizer [2]. Baseline-U generates a summary consisting of sentences that have been manually selected from the dataset by a team of five human "sentence-extractors" from the University of Montreal. It provides an approximate upper bound on what can be achieved with a purely extractive summarizer. This baseline method is only available on TAC2009.

The Baseline-MR method could be considered as an extension of the method proposed in [6] on update summarization. It involves two major steps: (1) a traditional manifold ranking strategy is applied on sentence manifold constructed from document set B; (2) an additional greedy algorithm is then employed to penalize sentences based on the document set A and sentences already selected for summary. The major difference lies between Baseline-MR and our approach is that Baseline-MR employs an additional greedy algorithm to address the novelty and diversity, while our approach introduce sink points into manifold ranking for the same purpose.

Besides, we denote our approach based on manifold ranking with sink points as *MRSP*. For experiments, we set the only parameter  $\alpha$  of MRSP to 0.85, and the parameters of Baseline-MR are set as follows:  $\alpha = 0.9$ ,  $\omega = 1$ . Note here  $\alpha$  acts as a balance factor between the influence of the intrinsic manifold structure and the prior knowledge on each sentence in both methods, and  $\omega$  is the same penalty factor as used by Wan et al. [6]. We set the parameters to the specific values as the corresponding summarization approach can achieve its best performance. To represent document set A on our sentence manifold, we form a pseudo sentence for A by aggregating all the sentences. Note that there are other ways to represent the information in document set A on the sentence manifold, e.g., one may make a summary on A first and then turn the summary into a data point on the sentence manifold.

### 3.4 Evaluation Results

#### 3.4.1 Summarization Performance

The performance comparison based on update summarization tasks of TAC2008 and TAC2009 is shown in Table 1 and Table 2 respectively. Note that in Table 1, S14 represents the best performing system by ROUGE-2 on TAC2008. It is also an extractive summarization approach. In Table 2, S34 represents the best performing system by ROUGE-2 on TAC2009. However, since S34 is not a purely extractive summarization approach (with massive abstractive techniques), we also shown the best performing extractive summarization approach on TAC2009, dented as system S24, for better comparison.

From the results on TAC2008 shown in Table 1, we can see that our approach can achieve comparative performance to the best performing extractive approach S14 in terms of both ROUGE-2 and ROUGE-SU4, and significantly outperform the other two baseline methods (p-value<0.01). Similarly, from the results on TAC2009 shown in Table 2, we can observe that our approach can also obtain comparative performance to the best performing system S34, even though S34 uses massive abstractive techniques. Besides, our approach can significantly outperform the best performing extractive

**Table 1: Performance Comparison on TAC2008**

|             | ROUGE-2        | ROUGE-SU4      |
|-------------|----------------|----------------|
| <b>MRSP</b> | <b>0.10217</b> | <b>0.13778</b> |
| S14         | 0.101          | 0.137          |
| Baseline-MR | 0.09171        | 0.13124        |
| Baseline-L  | 0.060          | 0.094          |

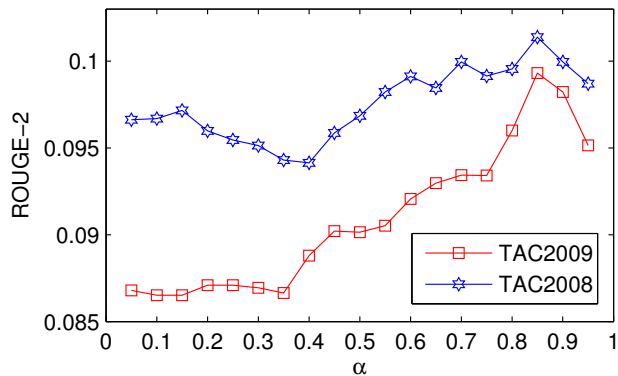
**Table 2: Performance Comparison on TAC2009**

|             | ROUGE-2        | ROUGE-SU4      |
|-------------|----------------|----------------|
| S34         | 0.10386        | 0.13851        |
| <b>MRSP</b> | <b>0.09932</b> | <b>0.13771</b> |
| Baseline-U  | 0.09820        | 0.13631        |
| S24         | 0.09615        | 0.13650        |
| Baseline-MR | 0.09002        | 0.12908        |
| Baseline-L  | 0.05142        | 0.09091        |

approach S24 on TAC2009 and all the other three baseline methods ( $p$ -value $<0.01$ ). It is interesting to notice that the Baseline-U method, the supposed upper bound baseline system of TAC2009 provided by NIST, was also beaten by our approach.

### 3.4.2 Parameter Tuning

There is only one parameter  $\alpha$  in our proposed model, which is in fact a balance factor between the influence of the intrinsic manifold structure and the prior knowledge on each sentence. Figure 1 shows the influence of parameter  $\alpha$  on the summarization performance.

**Figure 1: ROUGE-2 vs. Parameter  $\alpha$  on MRSP**

As we can see, the summarization approach performs not so well when  $\alpha$  is small, which may be due to the over emphasis of the prior knowledge. However, we can also notice the degradation of performance when  $\alpha$  approaches 1, which shows putting too much weight on the influence of structure may not work well either. As a result, our Algorithm achieves the best performance when  $\alpha = 0.85$  approximately on both benchmarks of TAC2008 and TAC2009.

## 4. CONCLUSION

In this paper, we propose a novel approach based on manifold ranking with sink points for update summarization. By introducing the sink points into manifold ranking process,

the multiple problems of update summarization including topic relevance, salience, novelty, and diversity, can be simultaneously addressed in a unified way. Experiments on benchmark of TAC2008 and TAC2009 show that the proposed approach can achieve comparative performance to the existing best performing systems in TAC tasks and significantly outperform other baseline methods.

For the future work, it is interesting to apply our proposed algorithm in Information Retrieval and Recommendation scenarios where a ranking that can simultaneously considers relevance, representative, novelty and diversity is also expected.

## Acknowledgments

This research work was supported by the State Key Program of National Natural Science Foundation of China (Grant No. 60933005) and the Program of National Natural Science Foundation of China (Grant No.60903139).

## 5. REFERENCES

- [1] F. Boudin, M. El-Bèze, and J.-M. Torres-Moreno. A scalable MMR approach to sentence scoring for multi-document update summarization. In *Coling 2008: Companion volume: Posters*, pages 23–26, Manchester, UK, August 2008. Coling 2008 Organizing Committee.
- [2] H. T. Dang and K. Owczarzak. Overview of the tac 2009 summarization track (draft). In *Proceedings of the Second Text Analysis Conference (TAC2009)*, 2009.
- [3] W. Li, F. Wei, Q. Lu, and Y. He. PNR2: Ranking sentences with positive and negative reinforcement for query-oriented update summarization. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 489–496, Manchester, UK, August 2008. Coling 2008 Organizing Committee.
- [4] J. Steinberger and K. Ježek. Update summarization based on novel topic distribution. In *Proceedings of the 9th ACM symposium on Document engineering*, pages 205–213, New York, NY, USA, 2009. ACM.
- [5] X. Wan. Timedextrank: adding the temporal dimension to multi-document summarization. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 867–868, New York, NY, USA, 2007. ACM.
- [6] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2903–2908, Hyderabad, India, January 6-12 2007.
- [7] J. Zhang, X. Cheng, H. Xu, X. Wang, and Y. Zeng. Ictcas’s ictgrasper at tac 2008: Summarizing dynamic information with signature terms based content filtering. In *Proceedings of the First Text Analysis Conference (TAC2008)*, 2008.
- [8] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.