Decayed DivRank for Guided Summarization

Pan Du, Jipeng Yuan, Xianghui Lin, Jin Zhang, Jiafeng Guo, Xueqi Cheng
Institute of Computing Technology
Chinese Academy of Sciences, Beijing 100190, China
{dupan,yuanjipeng,xianghuilin,zhangjin,guojiafeng}@software.ict.ac.cn
cxq@ict.ac.cn

Abstract

Guided summarization is essentially an aspect-based multi-document summarization, where aspects can be taken as specified queries in summarization. We proposed a novel ranking algorithm, *Decayed DivRank (DDRank)* for guided summarization tasks of TAC2011. DDRank can address relevance, importance, diversity, and novelty simultaneously through a decayed vertex-reinforced random walk process in sentence ranking. With sentence ranked by DDRank, top-ranked sentences are supposed to be able to present aspect-related, important, diverse, and novel information contained in documents. Hence, we can select the top-ranked sentences as summary candidates directly. Aspects are approximately obtained by pLSA, and are adopted to generate sentence prior for DDRank in our approach.

1 Introduction

Guided summarization[1] has been one the main tasks of Text Analysis Conference for two years. This task aims to write a 100-word summary of a set of 10 newswire articles for a given topic. The topic falls into a predefined category. Given a list of aspects for each category, the summary must cover all these aspects if the information can be found in the documents. Besides, guided summarization also demands an update summary, similar to the update summarization[2] in TAC2009. Update summarization aims at generating summaries assuming the user has read some articles before. Specifically, given the topic, the task is to write two summaries, one for document set A and the other for document set B. The summary for document set A is a guided summary. The update summary for document set B is also guided one but should be written under the assumption that the user of the summary has already read the documents in document set A. Each summary should be well-organized, in English, using complete sentences.

Guided Summarization caters for two emerging demands of information processing. One is the aspect-specific requirement, the other is time-dependent requirement. A user expects the summary to contain information specific to the particular aspects of the event. Meanwhile, since new information is created as the events develop, a user may want the summary to contain fresh information, to save time. However, much of current work has focused on the specified static document collection without attempting to capture the changes over time or trying to give the aspect-based information. The classic problem of summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs [14], which has been studied in many variations and has been addressed through a lot of summarization techniques [9, 8, 5, 22, 19, 4]. However, the demands of novel and aspect-specific information have not been fully recognized yet.

The goal of guided summarization task is to address these two new demands of summarization simultaneously. By providing concise, aspect-specific summaries of the

[1] http://www.nist.gov/tac/2011/Summarization/
[2] http://www.nist.gov/tac/2009/Summarization/

periodical dynamic information devoted to a common topic, guided summary can save the users from browsing the web content with much redundancy. We can formulate the guided summarization task as aspect-based update summarization, which can be valuable for periodically monitoring the important changes of specific aspect from the documents varying over a given time period.

Guided summarization provides clearer requirements of automatic summary when faced with specific categories of documents. The difficulty lies in mining those specific aspects. Discovering the changes in the event is also a challenge. There are five categories in total, each category with a separate list of aspects. The categories and the corresponding aspects are listed as follows:

Accidents and Natural Disasters: what happened; date; location; reasons for accident/disaster; casualties; damages; rescue efforts/countermeasures

Attacks: what happened; date; location; casualties; damages; perpetrators; rescue efforts/countermeasures

Health and Safety: what is the issue; who is affected; how they are affected; why it happens; countermeasures

Endangered Resources: description of resource; importance of resource; threats to resource; countermeasures

Trials and Investigations: who is under investigation, who is investigating/suing; why (general); specific charges; sentence/consequences; how do they plead/react to charges

Summaries are supposed to find all the aspects corresponding to the category. Besides, an update summary is also required for each document collection B. Update summarization is essentially a temporal extension of topic-focused multi-document summarization. As defined in [1], the temporal summarization is to summarize from web documents over a given time interval. The temporal summarization focuses on the identification of changes between web documents. Both the requirements of novel information and aspect-specific information of guided summarization are seldom well addressed in current state of the art.

In this paper, we introduce Decayed DivRank [7], DDRank to cope with guided summarization. DDRank is a unified approach which aims to address topic-relevance, importance, information diversity and novelty simultaneously in ranking. By modeling aspects as subtopics, we adopt aspects-guided DDrank to rank sentences for guided summarization. Evaluation results on datasets of TAC2011indicate the effectiveness of our approach.

In Section 2, we give an overview of the related works. The proposed DDRank are demonstrated in section 3. The experiments and evaluation followed in section 4. Finally, we conclude this paper with a summary and discussion of results in TAC2011, and look ahead to future work.

2 Related Work

2.1 DDRank

DDRank is an extended version of DivRank [15]. DivRank is essentially a query-independent ranking model, which addresses importance and diversity simultaneously by leveraging a vertex-reinforced random walk process. The reinforcement mechanism makes the vertices which have been heavily visited be more heavily visited in the future. Therefore the differences on visiting times between vertices are enlarged during the random walk process. In other words, the vertex-reinforcement in random walk introduces a "rich-get-richer" phenomenon, which means the prestige vertices will take advantage of its neighbors in ranking. In this way, all the prestige vertices will be ranked much higher than their neighbors. Diversity among the top-ranked vertices is then naturally achieved in the form of diversity among different prestige vertices. On the other hand, the criteria of relevance, which rely on the network structure, can not be addressed by DivRank anymore. As the transition probability of vertex-reinforced random walk is changed during the ranking process, the corresponding network structure changes as well.

DDRank extends DivRank by introducing a decayed factor on reinforcement, so as to embrace the capability of relevance in ranking. Decayed factor is designed according to the distance between query vertex and vertices to be ranked. The reinforcement on vertices near query is weaker than those far away from query. Hence, the local structure is partially reserved to achieve relevance during the vertex-reinforced random walk. In this way, relevance, salience and diversity can be addressed in a unified ranking process.

2.2 Update Summarization

Update summarization is a temporal extension of topic-focused multi-document summarization [21, 6, 10], by focusing on summarizing up-to-date information contained in the new document set given a past document set. A major approach for update summarization is extractive summarization [13, 9, 16]. In the extractive approach, update summarization is reduced to a sentence ranking problem, which composes a summary by extracting the most representative sentences from target document set. There are four goals a ranking algorithm for update summarization aims to achieve:

- *Topic Relevance*: The summary is based on a topic-related multi-document set, where a topic represents user's information need (either a short query or narrative). Therefore, the summary must stick to the topic users are interested in.

- *Importance*: Not all the sentences in the documents deliver information of equal importance about the topic. The summary has to neglect trivial content and include important information instead.

- *Diversity*: There should be less redundant information in the summary, so that the limited summary space can cover as much information as possible about the topic.

- *Novelty*: Given a specified topic and two chronologically ordered document sets, the summary needs to focus on the new information conveyed by the later dataset as compared with the earlier one under that topic.

Technically, *novelty* can be considered as a special kind of diversity since it focuses on the difference between sentences of newcoming documents and those of earlier documents, while *diversity* focuses on the difference between sentences selected already and those to be selected next.

Update summarization is most commonly used in a dynamic web environment. Allan et al. [1] generated temporal summaries over news stories on a certain event, which could be considered as an early form of update summarization. Recently, Boudin et al. [2] described a scalable sentence scoring method, SMMR derived from MMR [3], where candidate sentences were selected according to a combined criterion of query relevance and dissimilarity with previously read sentences. However, neither MMR nor SMMR took the influence of importance into consideration. Wan et al. [20] presented the TimedTextRank algorithm, a PageRank variation with a time factor, to select new and important sentences for update summarization. They achieved diversity through an additional penalty step based on cosine similarity measurement in a heuristic way. Li et al. [11] presented a positive and negative reinforcement ranking strategy PNR^2 to capture novelty for update summarization. They also penalized redundancy similarly as [20] to encourage diversity. It's hard to address the four goals of update summarization in a unified way.

3 Decayed DivRank

In this paper, we propose a novel approach DDRank [7] to address diversity as well as relevance and importance in ranking in a unified way.

Before describing the DDRank approach, we first introduce the original DivRank. The iteration process of DivRank is described as equation 1:

$$f_{t+1}^T = \alpha f_t^T (P_0 N_t) D_t^{-1} + (1-\alpha) r^T, \; 0 \le \alpha \le 1 \qquad (1)$$

where $P_0 = \beta P + (1-\beta) I$, $0 \le \beta \le 1$ and $D_t(i,i) = \sum_{j=1}^{n} P_0(i,j) N_t(j,j)$, f^T is the ranking score vector and r^T is the prior vector about relevance. P is the primitive transition matrix acquired from the adjacent relationship of a weighted network. P_0 is the new transition matrix on which the vertex-reinforced random walk depend. I is an identity matrix to forge self-links. The self-links in P_0 help to prevent the vertices from losing the profit already acquired during the reinforcement. N_t is a diagonal matrix with each diagonal element recording the visiting times of corresponding object. It acts as the reinforcing factor during the random walk process. Matrix D_t is to re-normalize

PN_t into a transition matrix P_t, and to make sure the process will eventually converge.

We now describe our DDRank model, a query dependent ranking model where relevance, prestige and diversity are addressed simultaneously, To capture the relevance, we try to preserve the local structure around the query, and improve the competitiveness of these relevant objects during the DivRank process. For this purpose, we modify the DivRank algorithm by re-weighting the reinforcement on each object according to the relevance between the corresponding object and the query. In this way, we can achieve two goals: 1) If the object is more relevant to the query, it will be more competitive. 2) The competition between objects near the query is weaker than that away from it, which makes tradeoff between relevance and diversity. Therefore, we can balance relevance, prestige and diversity during the vertex-reinforced random walk.

Formally, the DDRank model can be described as follows:

$$f_{t+1}^T = \alpha f_t^T (P_0 N_t^{1-r}) D_t^{-1} + (1-\alpha) r^T,\ 0 \le \alpha \le 1 \qquad (2)$$

where $P_0 = \beta P + (1-\beta) I,\ 0 \le \beta \le 1$ and $D_t(i,i) = \sum_{j=1}^{n} P_0(i,j) N_t(j,j)^{1-r}$. If the network is ergodic, after a sufficiently large t, the reinforced random walk defined by equation 2 also converges to a stationary distribution π. Then this distribution is used to rank the vertices in the information network by DDRank.

From equation 2 we can find that the ratio of N_t between a couple of neighbors (denoted as $\tilde{a} \approx \left[\frac{N_t(1)}{N_t(2)}\right]^{1-r} = a^{1-r}$) is suppressed according to their relevance r because the first order derivative form of $\tilde{a}$ on r is $\tilde{a}'_r = -ln(a)a^{1-r}$. We have $\tilde{a}'_r > 0$ if $a < 0$, and $\tilde{a}'_r < 0$ if $a > 0$. Similarly, the first order derivative form of reinforcement (denoted as $\tilde{e} = N_t^{(1-r)}$) on r is $\tilde{e}'_r = -ln(N_t) N_t^{1-r}$. Noting that $N_t \le 1$, we have $\tilde{e}'_r \ge 0$. This means the more relevant objects are attached more competitiveness than the less relevant ones in DDRank. Relevance r is approximated by the similarity between the query and others in our experiment.

4 Experiments

4.1 Data Set

The test dataset of TAC2011 is composed of 44 topics, divided into five categories: Accidents and Natural Disasters, Attacks, Health and Safety, Endangered Resources, Investigations and Trials. Each topic has a topic ID, category, title, and 20 relevant documents which have been divided into 2 sets: Document Set A and Document Set B. Each document set has 10 documents, and all the documents in Set A chronologically precede the documents in Set B. Unlike in previous years, there is no topic narrative, because the category and its aspects already define what information the reader is looking for. The summary for Document Set A should be a straightforward query-focused summary. The update summary for Document Set B is also query-focused but should be written under the assumption that the user of the summary has already read the documents in Document Set A. The documents for summarization come from the AQUAINT and AQUAINT-2 collections of news articles. The AQUAINT corpus of English News Text consists of documents taken from the New York Times, the Associated Press, and the Xinhua News Agency newswires (LDC catalog number LDC2002T31). The collection spans the years 1999-2000 (1996-2000 for Xinhua documents). The AQUAINT-2 collection spans the time period of October 2004 - March 2006; articles are in English and come from a variety of sources including Agence France Presse, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times-Washington Post News Service, New York Times, and the Associated Press.

4.2 Evaluation Metric

ROUGE[12], Recall Oriented Understudy for Gisting Evaluation, is a metric adopted by TAC for automatic summarization evaluation. There are several variants that can be used in practice with provided toolkits. ROUGE-N measures summary quality by counting overlapping units of n-gram between the candidate summary (peer) and the reference summaries (model). The evaluation metrics we adopted in our training process are ROUGE-1, ROUGE-2 and ROUGE-SU4 respectively. ROUGE-N is computed as follows:

$$ROUGE-N = \frac{\sum\limits_{S \in ReferenceSummries} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in ReferenceSummries} \sum\limits_{gram_n \in S} Count(gram_n)} \quad (3)$$

Where n stands for the length of n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. ROUGE-SU4 is a skip-bigram co-occurrence measure with addition of unigrams as counting unit.

The ROUGE toolkit reports scores for 1-, 2-, 3-, and 4-gram. We show three of the ROUGE metrics in the experimental results, at a confidence level of 95%: ROUGE-1, ROUGE-2, ROUGE-SU4.

Pyramid[17, 18] is a manual metric used for summary evaluation in TAC. Its kernel concept is Summary Content Units, referred as SCUs, which are semantically motivated, sub-sentential units that are variable in length but no bigger than a sentential clause. SCUs emerge from annotation of a collection of human summaries for the same input. They are identified by noting information that is repeated across summaries, whether the repetition is as small as a modifier of a noun phrase or as large as a clause. The weight an SCU obtains is directly proportional to the number of reference summaries that support that piece of information. The evaluation method that is based on overlapping SCUs in human and automatic summaries is described in the Pyramid method.

4.3 Experimental Results

Our proposed approaches for guided summarization, which is an extension of update summarization, performs well in Text Analysis Conference of 2011. The guided summarization task in TAC 2011 requires the generation of 100-word summaries for 44 topics. Each topic has a topic category and 20 relevant documents which have been divided into 2 sets: Document Set A and Document Set B. Each document set has 10 documents, and all the documents in Set A chronologically precede the documents in Set B. The generated summaries are evaluated by the National Institute of Standards and Technology (NIST[3]). All summaries were truncated to 100 words before being evaluated by manual and automatic metrics. The evaluation results of our DDRank-based system of Run 12 and 26 are demonstrated in Table 1 and Table 2 respectively.

Table 1: Evaluation Results of DDRank (Run 12) in TAC11.

Metric	Score	Rank
Pyramid - A	0.420	19
Pyramid - B	0.351	14
BE - A	0.06829	21
BE - B	0.05717	4
ROUGE-2 - A	0.10917	17
ROUGE-2 - B	0.07992	15
ROUGE-SU4 - A	0.14541	12
ROUGE-SU4 - B	0.12062	10

Table 2: Evaluation Results of DDRank (Run 26) in TAC11.

Metric	Score	Rank
Pyramid - A	0.435	14
Pyramid - B	0.335	9
BE - A	0.07099	13
BE - B	0.05717	3
ROUGE-2 - A	0.11324	11
ROUGE-2 - B	0.07992	14
ROUGE-SU4 - A	0.14901	9
ROUGE-SU4 - B	0.12062	9

Both of runs 12 and 26 are DDRank-based methods. The query of run 12 is merely the topic title given by NIST, while that of 26 is acquired by pLSA, a simple topic model. Each sentence is assigned a score measuring the probability that it belongs to certain aspects acquired by pLSA. The max score of each sentence with respect to aspects is adopted as its prior for DDRank. Comparing

[3] http://www.nist.gov/

to other participants, the performance of our system 12 and 26 perform well on BE metric for set B, which means that it captured the update nature of set B successfully. However, its performance on sets A is not good enough, which might be caused by the unique setting of parameter α. Performance of run 26 is somehow better than that of 12, as can be seen in Table 1 and 2, which means mining specific aspects does benefit for guided summarization. However, both run 12 and 26 failed to achieve an exciting performance this year. This is probably because that the extracted information on aspects are not qualified enough to generate a good guided summary.

5 Conclusions and Future Work

In this paper, we proposed a novel approach for guided summarization task of TAC 2011. Our approach is based on DDRank, which can address relevance, prestige, diversity, and novelty simultaneously in ranking. According to the evaluation results, it is helpful to explicitly mine the aspects of each category for guided summarization. However, a more effective aspects mining approach is required to enhance the summarization performance. We will consider a refined aspect model to improve the performance of DDRank for guided summarization in our future work.

6 Acknowledgments

This research work was funded by the National High-tech R&D Program of China under Grant No. 2010AA012502, and the National Natural Science Foundation of China under Grant No. 61003166, Grant No. 60903139 and Grant No. 60933005.

References

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18, New York, NY, USA, 2001. ACM.

[2] 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.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM.

[4] X. Cheng, P. Du, J. Guo, X. Zhu, and Y. Chen. Ranking on data manifold with sink points. *Knowledge and Data Engineering, IEEE Transactions on*, PP(99):1, 2011.

[5] J. M. Conroy and D. P. O'leary. Text summarization via hidden markov models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407, New York, NY, USA, 2001. ACM.

[6] J. M. Conroy and J. D. Schlesinger. Classy query-based multidocument summarization. In *In Proceedings of DUC'2005*, 2005.

[7] P. Du, J. Guo, and X.-Q. Cheng. Decayed divrank: capturing relevance, diversity and prestige in information networks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 1239–1240, New York, NY, USA, 2011. ACM.

[8] P. Du, J. Guo, J. Zhang, and X. Cheng. Manifold ranking with sink points for update summarization. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1757–1760, New York, NY, USA, 2010. ACM.

[9] G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, 2004.

[10] E. Hovy, C. yew Lin, L. Zhou, and J. Fukumoto. Automated summarization evaluation with basic elements. In *In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC)*, 2006.

[11] 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.

[12] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In S. S. Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[13] C.-Y. Lin and E. Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[14] I. Mani. Recent developments in text summarization. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 529–531, New York, NY, USA, 2001. ACM.

[15] Q. Mei, J. Guo, and D. Radev. Divrank: the interplay of prestige and diversity in information networks. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1009–1018, New York, NY, USA, 2010. ACM.

[16] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In D. Lin and D. Wu, editors, *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[17] A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4, 2007.

[18] A. Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152, 2004.

[19] A. Nenkova, L. Vanderwende, and K. McKeown. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 573–580, New York, NY, USA, 2006. ACM.

[20] X. Wan. Timedtextrank: 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.

[21] F. Wei, W. Li, Q. Lu, and Y. He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 283–290, New York, NY, USA, 2008. ACM.

[22] H. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 113–120, New York, NY, USA, 2002. ACM.