CAAI Transactions on Intelligence Technology 2 (2017) 26–30

Contents lists available at ScienceDirect

CAAI Transactions on Intelligence Technology

journal homepage: http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Original article

Density peaks clustering based integrate framework for multi-document summarization

Baoyan Wang [a], Jian Zhang [b, e, *], Yi Liu [c, d], Yuexian Zou [a]

[a] ADSPLAB, School of ECE, Peking University, Shenzhen, 518055, China
[b] Shenzhen Raisound Technologies, Co., Ltd, China
[c] PKU Shenzhen Institute, China
[d] PKU-HKUST Shenzhen-Hong Kong Institute, China
[e] School of Computer Science and Network Security Dongguan University of Technology, China

ARTICLE INFO

Article history:
Received 14 October 2016
Accepted 25 December 2016
Available online 20 February 2017

Keywords:
Multi-document summarization
Integrated score framework
Density peaks clustering
Sentences rank

ABSTRACT

We present a novel unsupervised integrated score framework to generate generic extractive multi-document summaries by ranking sentences based on dynamic programming (DP) strategy. Considering that cluster-based methods proposed by other researchers tend to ignore informativeness of words when they generate summaries, our proposed framework takes relevance, diversity, informativeness and length constraint of sentences into consideration comprehensively. We apply Density Peaks Clustering (DPC) to get relevance scores and diversity scores of sentences simultaneously. Our framework produces the best performance on DUC2004, 0.396 of ROUGE-1 score, 0.094 of ROUGE-2 score and 0.143 of ROUGE-SU4 which outperforms a series of popular baselines, such as DUC Best, FGB [7], and BSTM [10].

1. Introduction

With the explosively growing of information overload over the Internet, consumers are flooded with all kinds of electronic documents i.e. news, emails, tweets, blog. Now more than ever, there are urgent demands for multi-document summarization (MDS), which aims at generating a concise and informative version for the large collection of documents and then helps consumers grasp the comprehensive information of the original documents quickly. Most existing studies are extractive methods, which focus on extracting salient sentences directly from given materials without any modification and simply combining them together to form a summary for multi-document set. In this article, we study on the generic extractive summarization from multiple documents. Nowadays, an effective summarization method always properly considers four important issues [1,2]:

- Relevance: a good summary should be interrelated to primary themes of the given multi-documents as possible.
- Diversity: a good summary should be less redundant.
- Informativeness: the sentences of a good summary should conclude information as much as possible.
- Length Constraint: the summary should be extracted under the limitation of the length.

The extractive summarization methods can fall into two categories: supervised methods that rely on provided document-summary pairs, and unsupervised ones based upon properties derived from document clusters. The supervised methods consider the multi-document summarization as a classification/regression problem [3]. For those methods, a huge amount of annotated data is required, which are costly and time-consuming. For another thing, unsupervised approaches are very enticing and tend to score sentences based on semantic grouping extracted from the original documents. Researchers often select some linguistic features and statistic features to estimate importance of original sentences and then rank sentences.

Inspired by the success of cluster-based methods, especially density peaks clustering (DPC) algorithm on bioinformatics, bibliometric, and pattern recognition [4], in this article we propose a novel method to extract sentences with higher relevance, more informativeness and a better diversity under the limitation of length for sentences ranking based on Density Peaks Clustering (DPC). First, thanks to the DPC, it is not necessary to provide the

* Corresponding author. Shenzhen Raisound Technologies, Co., Ltd, China.
E-mail address: 13925876721@163.com (J. Zhang).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.12.005

established number of clusters in advance and do the post-processing operation to remove redundancy. Second, we attempt to put forward an integrated score framework to rank sentences and employ the dynamic programming solution to select salient sentences.

This article is organized as follows: Section 2 describes related research work about our motivation in detail. Section 3 presents our proposed Multi-Document Summarization framework and the summary generation process based on dynamic programming technology. Section 4 and Section 5 give the evaluation of the algorithm on the benchmark data set DUC2004 for the task of multi-document summarization. We then conclude at the end of this article and give some directions for future research.

2. Related work

Various extractive multi-document summarization methods have been proposed. For supervised methods, different models have been trained for the task, such as hidden Markov model, conditional random field and REGSUM [5]. Sparse coding [2] was introduced into document summarization due to its useful in image processing. Those supervised methods are based on algorithms that a large amount of labeled data is needed for precondition. The annotated data is chiefly available for documents, which are mostly relevant to the trained summarization model. Therefore, it's not necessary for the trained model to generate a satisfactory summary when documents are not parallel to the trained model. Furthermore, when consumers transform the aim of summarization or the characteristics of documents, the training data should be reconstructed and the model should be retrained necessarily.

There are also numerous methods for unsupervised extracted-based summarization presented in the literature. Most of them tend to involve calculating salient scores for sentences of the original documents, ranking sentences according to the saliency score, and utilizing the top sentences with the highest scores to generate the final summary. Since clustering algorithm is the most essential unsupervised partitioning method, it is more appropriate to apply clustering algorithm for multi-document summarization. The cluster based methods tend to group sentences and then rank sentences by their saliency scores. Many methods use other algorithms combined with clustering to rank sentences. Wan et al. [6] clustered sentences first, consulted the HITS algorithm to regard clusters as hubs and sentences as authorities and then ranked and selected salient sentences by the final gained authority scores. Wang et al. [7]translated the cluster-based summarization issue to minimizing the Kullback-Leibler divergence between the original documents and model reconstructed terms. Cai et al. [8] ranked and clustered sentences simultaneously and enhanced each other mutually. Other typical existing methods include graph-based ranking, LSA based methods, NMF based methods, submodular functions based methods, LDA based methods. Wang et al. [9]used the symmetric non-negative matrix factorization (SNMF) to softly cluster sentences of documents into groups and selected salient sentences from each cluster to generate the summary. Wang et al. [10] used generative model and provided an efficient way to model the Bayesian probability distributions of selecting salience sentences given themes. Wang et al. [11]combined different summarization results from single summarization systems. Besides, some papers considered reducing the redundancy in summary, i.e. MMR [12]. To eliminate redundancy among sentences, some systems selected the most important sentences first and calculated the similarity between previously selected ones and next candidate sentence, and add it to the summary only if it included sufficient new information.

We follow the idea of cluster-based method in this article. Different from previous work, we attempt to propose an integrated weighted score framework that can order sentences by evaluating salient scores and remove redundancy of summary. We also use the dynamic programming solution for optimal salient sentences selection.

3. Proposed method

In this section, we discuss the outline of our proposed method as illustrated in Fig. 1. We show a novel way of handling the multi-document summarization task by using DPC algorithm. All documents are first represented by a set of the sentences as raw input of the framework. After the corpus is preprocessed, DPC is employed to get relevance scores and diversity scores of sentences simultaneously. Meanwhile, the number of effective words will be applied to obtain informativeness scores of sentences. What's more, a length constraint is used to ensure the extracted sentenced have a proper length. In the end, we attempt to use an integrated scoring framework to rank sentences and generate the summary based on the dynamic programming algorithm. The DPC based summarization method mainly includes the following steps:

3.1. Pre-processing

Before using our method to deal with the text data, a pre-processing module is indispensable. After the given corpus of English documents, $C\ corpus = \{d\ 1,d2, \ldots,d\ i, \ldots,d\ cor\}$, which d i denotes the i-th document in *C corpus* and those documents are same or similar topics, splitting apart into individual sentences, $S = \{s\ 1,s\ 2, \ldots s\ i, \ldots,s\ sen\}$ where s i means the i-th sentence in *C corpus*, we utilize an undefined forward stop words list to remove all stop words and Porter's stemming algorithm to perform stem of remaining words.

3.2. Sentence estimation factors

3.2.1. Relevance score

In this section, we show a relevance score to measure the extent how much a sentence is relevant to residual sentences in the

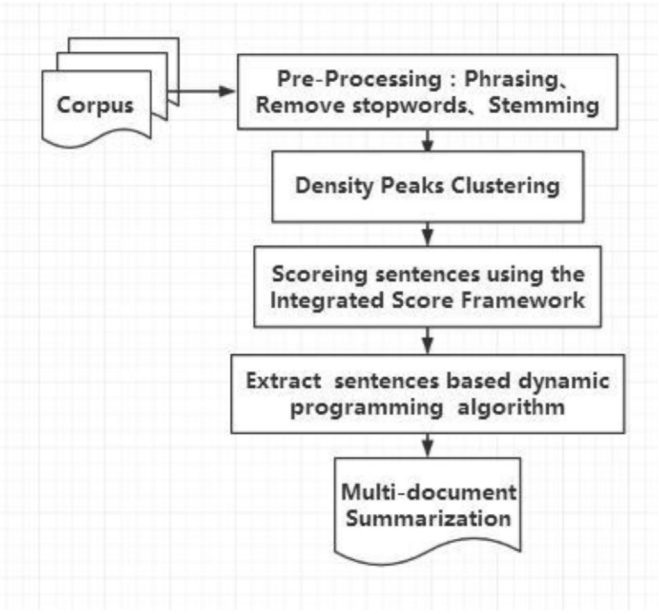

Fig. 1. The outline of our proposed framework.

documents. One of the underlying assumptions of DPC is that cluster centers are characterized by a higher density than their neighbors. Inspired by the assumption, we assume that a sentence will be deemed to be higher relevance and more representational when it possesses higher density meaning owning more similar sentences. As the input of the DPC algorithm is similarity matrix among sentences, the sentences are represented by bag-of-words vector space mode primarily, and then cosine similarity formula is applied to calculate the similarity among sentences. The reason why terms are weighted with Binary schemes, which Term weighting Wij is set 1 if term tj appears at least once in the sentence, is that the frequency of term repetition tend to be less in sentences than that in documents. Thus we define the function to compute the Relevance Scoring SCrele(i) for each sentence si as following:

$$Sim_{ij} = \frac{\sum_t W_{ti} * W_{tj}}{\sum_t W_{ti} * \sum_t W_{tj}}, W_{ij} = \begin{cases} 1 & t_j \in s_i \\ 0 & else \end{cases} \tag{1}$$

$$SC_R(i) = \sum_{j=1}^{K} f(Sim_{ij} - \omega) \ \ , f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & else \end{cases} \tag{2}$$

where *Sim ij* represents the cosine similarity numerical value between the *i*-th and *j*-th sentence, *K* denotes the total number of sentences in the documents and *T* denotes the total number of terms in the documents. ω represents the predefined value of density threshold. $SCR\,(i)$ should be normalized in order to adapt to the comprehensive scoring model.

$$SC_{rele}(i) = SC_R^*(i) \Big/ \max_j SC_R^*(i) \tag{3}$$

In this section, the density threshold ω is determined following the study [4] to exclude the sentences, which hold lower similarity values with the others.

3.2.2. Diversity score

In this section, diversity scoring is presented to argue a good summary should not include analogical sentences. A document set usually contains one core topic and some subtopics. In addition to the most evident topic, it's also necessary to get the sub-topics most evident topic so as to better understand the whole corpus. In other words, sentences of the summary should be less overlap mutually so as to eliminate redundancy. Maximal Marginal relevance (MMR), one of the typical methods reducing redundancy, uses a greedy approach to sentence selection through combing criterion of query relevance and novelty of information. Another hypothesis of DPC is that cluster centers also are characterized by a relatively large distance from points with higher densities, which ensure the similar sentences get larger difference scores. Therefore, by comparing with all the other sentences of the corpus, the sentence with a higher score could be extracted, which also can guarantee the diversity globally. The diversity score *SCdiv*(*i*) is defined as the following function.

$$SC_{div}(i) = 1 - \max_{SC_R^*(j) > SC_R^*(i)} Sim_{ij} \tag{4}$$

Note that diversity score of the sentence with the highest density is assigned 1 conventionally.

3.2.3. Informativeness score

Relevance score and diversity score measure the relationship between the sentences. In this section, Informative content words are employed to calculate the internal informativeness of sentences. Informative content words are the non-stop words and parts of speech are nouns, verbs and adjectives.

$$SC_{Inf}^*(i) = \sum_{j=1}^{T} W_{ij} \tag{5}$$

It's also necessary to normalize the informativeness scoring as follows:

$$SC_{infor}(i) = SC_{Inf}^*(i) \Big/ \max_j SC_{Inf}^*(j) \tag{6}$$

3.2.4. Length constraint

The longer sentence is, the more informativeness it owns, which causes the longish sentences tend to be extracted. The total number of words in the summary usually is limited. The longer sentences are, the fewer ones are selected. Therefore, it is requisite to provide a length constraint. Length of sentences li range in a large scope. On this occasion, we should lead in a smoothing method to handle the problem. Taking logarithm is a widely used smoothing approach. Thus the length constraint is defined as follows in (7).

$$SC_{len}^* = \log\left(\max_j L_j / L_i \right) \tag{7}$$

It needs to be normalized as the previous operations:

$$SC_{len}(i) = SC_{len}^*(i) \Big/ \max_j SC_{len}^*(j) \tag{8}$$

3.3. Integrated score framework

The ultimate goal of our method is to select those sentences with higher relevance, more informativeness and better diversity under the limitation of length. We define a function comprehensively considering the above purposes as follows:

$$SC^*(i) = SC_{rele}(i)^a * SC_{div}(i)^b * SC_{infor}(i)^c * SC_{len}(i) \tag{9}$$

In order to calculate concisely and conveniently, the scoring framework then is changed to:

$$SC(i) = \alpha \log SC_{rele}(i) + \beta \log SC_{div}(i) + \gamma \log SC_{infor}(i) + \log SC_{len}(i) \tag{10}$$

Note that in order to determine how to tune the parametersα, β, and γ of the integrated score framework, we carry out a set of experiments on development dataset. The value of α, β, and γ was tuned by varying from 0 to 1.5, and chose the values, with which the method performs best.

3.4. Summary generation process

The summary generation is regarded as the 0–1 knapsack problem:

$$\arg \max \sum (SC(i) * x_i) \tag{11}$$

Subject to $\sum_i l_i x_i \leq L, x_i = \{0, 1\}$

The 0–1 knapsack problem is NP-hard. To alleviate this problem we utilize the dynamic programming solution to select sentences

until the expected length of summaries is satisfied, shown as follows.

$$
\begin{aligned}
&1.\ S[i][0] = 0 \ \ \forall i \in [1, K] \\
&2.\ \textit{for}\ i : 1 \ldots K \\
&3.\qquad \textit{for}\ l : 1 \ldots L \\
&4.\qquad\qquad S^{*} = S[i-1][l] \\
&5.\qquad\qquad S^{**} = S[i-1][l-l_i] + SC^{*}[i] \\
&6.\qquad\qquad S[i][l] = \max\{S^{*}, S^{**}\} \\
&7.\ \textit{return}\ \ \arg\max S[i][l]
\end{aligned}
$$

where S[i][l] stands for a high score of summary, that can only contain sentences in the set { *s 1,s 2, ...s i* } under the limit of the exact length l.

4. Experimental setup

4.1. Datasets and evaluation metrics

We evaluate our approach on the open benchmark data sets DUC2004 and DUC2007 from Document Understanding Conference (DUC) for summarization task. Table 1 gives a brief description of the datasets. There are four human-generated summaries, of which every sentence is either selected in its entirety or not at all, are provided as the ground truth of the evaluation for each document set.

In this section, DUC2007 is used as our development set to investigate howα, β, and γ relate to integrated score framework. ROUGE version 1.5.5 toolkit [13], widely used in the research of automatic documents summarization, is applied to evaluate the performance of our summarization method in experiments. Among the evaluation methods implemented in Rouge, Rouge-1 focuses on the occurrence of the same words between generated summary and reference summary, while Rouge-2 and Rouge-SU4 concerns more over the readability of the generated summary. We report the mean value over all topics of the recall scores of these three metrics in the experiment.

4.2. Baselines

We study with the following methods for generic summarization as the baseline methods to compare with our proposed method, which of them are widely applied in research or recently released in literature.

1: DUC best: The best participating system in DUC2004;
2: Cluster-based methods: KM [10], FGB [7], ClusterHITS [6], NMF [14], RTC [8];
3: Other state-of-the-art MDS methods: Centroid [15], LexPageRank [16], BST M [10], WCS [11].

5. Experimental results

We evaluate our method on the DUC 2004 data with $\alpha = 0.77$, $\beta = 0.63$, $\gamma = 0.92$ which was our best performance in the experiments on the development data DUC 2007. The results of these experiments are listed in Table 2. Fig. 2 visually illustrates the comparison between our method with the baselines so as to better demonstrate the results. We subtract the KM score from the scores of residual methods and then plus the number 0.01 in the figure, thus the distinction among those methods can be observed more distinctly. We show ROUGE-1, ROUGE-2 and ROUGE-SU Recall-measures in Table 2.

From Table 2 and Fig. 2, we can have the following observed results: our result is on the verge of the human-annotated result and our method clearly outperforms the DUC04 best team work. It is obvious that our method outperforms most rivals significantly on the ROUGE-1 metric and the ROUGE-SU metric. In comparison with the WCS, the result of our method is slightly worse. It may be due to the aggregation strategy used by WCS. The WCS aggregates various summarization systems to produce better summary results. Compared with other cluster-based methods, ours consider the informativeness of sentences and do not need to set the clusters' number. By removing one from the four scores of the integrated score framework, the results show that effectiveness of the method is reduced. In other words, the four scores of the integrated score framework have a promoting effect for the summarization task. In a word, it is effective for our proposed method to handle MDS task.

Table 1
Description of the dataset.

	DUC 2004	DUC 2007
Number of document sets	50	45
Number of news articles	10	20
Length Limit of summary	665 bytes	250 words
Data source	TDT	AQUAINT

Table 2
Overall performance comparison on DUC2004 dataset using ROUGE evaluation tool. Remark: "-" indicates that the corresponding method does not authoritatively release the results.

DUC best	0.38224(5)	0.09216(3)	0.13233(3)
Centroid	0.36728(9)	0.07379(8)	0.12511(8)
LexPageRank	0.37842(6)	0.08572(6)	0.13097(5)
NMF	0.36747(8)	0.07261(10)	0.12918(7)
FGB	0.38724(4)	0.08115(7)	0.13096(6)
KM	0.34872(11)	0.06937(9)	0.12115(9)
ClusterHITS	0.36463(10)	0.07632(8)	–
RTC	0.37475(7)	0.08973(5)	–
WCS	0.39872(1)	0.09611(1)	0.13532(2)
BSTM	0.39065(3)	0.09010(4)	0.13218(4)
OURS	0.39677(2)	0.09432(2)	0.14356(1)
OURS-*SCrele*	0.28956	0.04655	0.07665
OURS-*SCdiv*	0.36409	0.07927	0.12517
OURS-*SCinfor*	0.37416	0.08194	0.12974
OURS-*SClen*	0.38640	0.08936	0.13688

6. Conclusion

In this paper, we proposed a novel unsupervised method to

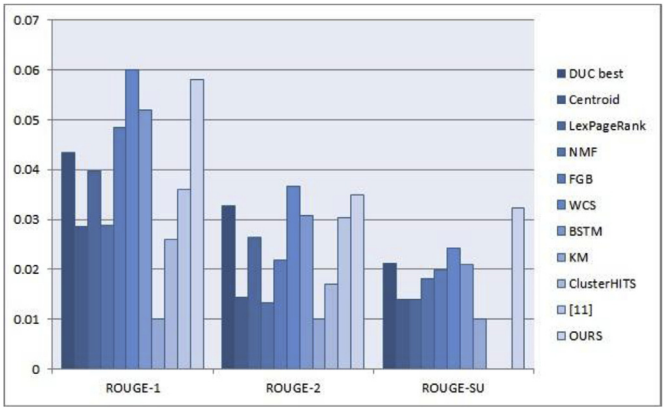

Fig. 2. Comparison of the methods in terms of ROUGE-1, ROUGE-2, and ROUGESU Recall-measures.

handle the task of multi-document summarization. For ranking sentences, we proposed an integrated score framework. Informative content words are used to get the informativeness, while DPC was employed to measure the relevance and diversity of sentences at the same time. We combined those scores with a length constraint and selected sentences based dynamic programming at last. Extensive experiments on standard datasets show that our method is quite effective for multi-document summarization.

In the future, we will introduce external resources such as Wordnet and Wikipedia to calculate the sentence semantic similarity, which can solve the problems of the synonym and the multivocal word. We will then apply our proposed method in topic-focused and updated summarization, to which the tasks of summarization have turned.

Acknowledgments

This work is partially supported by NSFC (No: 61271309, No. 61300197), Shenzhen Science & Research projects (No: CXZZ20140509093608290).

References

[1] T. Ma, X. Wan, Multi-document summarization using minimum distortion, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 354–363.
[2] H. Liu, H. Yu, Z.-H. Deng, Multi-document summarization based on two- level sparse representation model, in: AAAI, 2015, pp. 196–202.
[3] Z. Cao, F. Wei, L. Dong, S. Li, M. Zhou, Ranking with recursive neural networks and its application to multi-document summarization, in: AAAI, 2015, pp. 2153–2159.
[4] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science (6191) (2014) 1492–1496.
[5] K. Hong, A. Nenkova, Improving the estimation of word importance for news multi-document summarization, in: EACL, 2014, pp. 712–721.
[6] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2008, pp. 299–306.
[7] D. Wang, S. Zhu, T. Li, Y. Chi, Y. Gong, Integrating document clustering and multi-document summarization, ACM Trans. Knowl. Discov. Data (TKDD) 5 (3) (2011) 14.
[8] X. Cai, W. Li, Ranking through clustering: an integrated approach to multi-document summarization, IEEE Trans. Audio, Speech, Lang. Process. 21 (7) (2013) 1424–1433.
[9] D. Wang, T. Li, S. Zhu, C. Ding, Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2008, pp. 307–314.
[10] D. Wang, S. Zhu, T. Li, Y. Gong, Multi-document summarization using sentence-based topic models, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACL, 2009, pp. 297–300.
[11] D. Wang, T. Li, Weighted consensus multi-document summarization, Inf. Process. Manag. 48 (3) (2012) 513–523.
[12] J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, Multi-document summarization by sentence extraction, in: Proceedings of the 2000 NAACL- ANLP Workshop on Automatic Summarization-Volume 4, ACL, 2000, pp. 40–48.
[13] P. Over, J. Yen, Introduction to duc-2001: an intrinsic evaluation of generic news text summarization systems, in: Proceedings of DUC 2004 Document Understanding Workshop, Boston, 2004.
[14] D. Wang, T. Li, C. Ding, Weighted feature subset non-negative matrix factorization and its applications to document understanding, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 541–550.
[15] D.R. Radev, H. Jing, M. Sty's, D. Tam, Centroid-based summarization of multiple documents, Inf. Process. Manag. 40 (6) (2004) 919–938.
[16] Q. Mei, J. Guo, D. Radev, Divrank: the interplay of prestige and diversity in information networks, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 1009–1018.