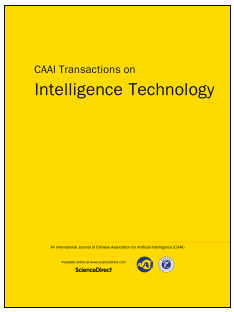

Density peaks clustering based integrate framework for multi-document summarization

Baoyan Wang, Jian Zhang, Yi Liu, Yuexian Zou

PII: S2468-2322(16)30056-7

DOI: 10.1016/j.trit.2016.12.005

Reference: TRIT 38

To appear in: *CAAI Transactions on Intelligence Technology*

Received Date: 14 October 2016

Accepted Date: 25 December 2016

Please cite this article as: B. Wang, J. Zhang, Y. Liu, Y. Zou, Density peaks clustering based integrate framework for multi-document summarization, *CAAI Transactions on Intelligence Technology* (2017), doi: 10.1016/j.trit.2016.12.005.

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
[c] PKU Shenzhen Institute
[d] PKU-HKUST Shenzhen-Hong Kong Institute
[e] School of Computer Science and Network Security Dongguan University of Technology

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].

***Index Terms*—**Multi-document summarization, Integrated Score Framework, Density Peaks Clustering, Sentences Rank

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 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