# Clustering Sentences with Density Peaks for Multi-document Summarization

**Yang Zhang**
Shenzhen Graduate School
Peking University, China
ecezhangy@sz.pku.edu.cn

**Yunqing Xia**
Dept. of Comp. Sci. and Tech.
Tsinghua University, China
yqxia@tsinghua.edu.cn

**Yi Liu**
IMSL, PKU-HKUST Shenzhen
Hong Kong Institution, China
yi.liu@imsl.org.cn

**Wenmin Wang**
Shenzhen Graduate School
Peking University, China
wangwm@ece.pku.edu.cn

## Abstract

Multi-document Summarization (MDS) is of great value to many real world applications. Many scoring models are proposed to select appropriate sentences from documents to form the summary, in which the clustering-based methods are popular. In this work, we propose a unified sentence scoring model which measures representativeness and diversity at the same time. Experimental results on DUC04 demonstrate that our MDS method outperforms the DUC04 best method and the existing clustering-based methods, and it yields close results compared to the state-of-the-art generic MDS methods. Advantages of the proposed MDS method are two-fold: (1) The density peaks clustering algorithm is firstly adopted, which is effective and fast. (2) No external resources such as Wordnet and Wikipedia or complex language parsing algorithms is used, making reproduction and deployment very easy in real environment.

## 1 Introduction

Document summarization is the process of generating a generic or topic-focused summary by reducing documents in size while retaining the main characteristics of original documents(Wang et al., 2011). The summary may be formed in a variety of different ways, which are generally categorized as abstractive and extractive(Shen et al., 2007). In this paper, we address the problem of generic multi-document summarization (MDS). An effective summarization method should properly consider the following three important issues: representativeness, diversity, conciseness.

Many scoring models are proposed to select appropriate sentences from documents to form the summary, in which the clustering-based methods are popular. Some researchers address the sentence scoring task in an *isolation* manner(Radev et al., 2004; Wang et al., 2008; Wan and Yang, 2008) (i.e., clustering and ranking are two independent steps). Others handle the sentence ranking task in a *mutuality* manner(Cai and Li, 2013; Cai et al., 2010; Wang et al., 2011) (i.e., clustering improves ranking and vice versa). Two drawbacks of the existing clustering-based methods are worth noting. First, extra algorithms are required to determine the number of clusters beforehand. Second, models are required to rank or score sentences within and across the clusters after clustering.

Our proposed MDS method is inspired by the recent work on density peaks clustering (DPC) algorithm published on *Science* (Rodriguez and Laio, 2014). The underlying assumption is that *cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities*. In this paper, we adapt the density peaks clustering algorithm(Rodriguez and Laio, 2014) to simultaneously cluster sentences and rank them in the *mutuality* manner. Thanks to the density peaks clustering algorithm, we *do not* need to set the number of clusters and *do not* need a post-processing module to reduce redundancy. From the view of summarization task, *DPC* is superior to other clustering methods because it can not only find the best cluster centers, but also do rank all data points, including



*Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1262–1267,
Denver, Colorado, May 31 – June 5, 2015. 

cluster centers, within and across clusters at the same time. Experimental results on the DUC2004 demonstrate that our method outperforms the best method in DUC04 and yields close results compared to the state-of-the-art unsupervised MDS methods.

The major contributions of this work are two-fold: Firstly, a unified sentence scoring model is proposed to consider representativeness, diversity and conciseness at the same time. Secondly, the density peaks clustering algorithm is first applied in the MDS task. We further revise the clustering algorithm to address the summary length constraint.

## 2 Related Work

A vast number of methods are reported in literatures on MDS. The MDS methods can be generally categorized into abstractive and extractive. The extractive MDS can be also categorised into supervised and unsupervised. Several supervised learning methods have been developed for training accurate model for extract-based summarization. The unsupervised methods, on the other hand, also contribute a lot to MDS. In this work, we put our contributions in context of the sentence ranking-based extractive MDS under the unsupervised framework.

Several clustering-based MDS methods have also been proposed. For example, ClusterHITS is proposed to incorporate the cluster-level information into the process of sentence ranking(Wan and Yang, 2008). RankClus is proposed to update sentence ranking and clustering interactively and iteratively with frequency relationships between two sentences, or sentences and terms (Cai et al., 2010). Some kinds of matrix factorization methods are also explored in MDS methods(Gong and Liu, 2001; Lee et al., 2009; Wang et al., 2008; Wang et al., 2011; Shen et al., 2011). For example, matrix factorization methods is adopted to generate sentence clusters, in which non-negative factorization is performed on the term-document matrix using the term-sentence matrix as the base so that the document-topic and sentence-topic matrices could be constructed(Wang et al., 2008).

We follow the idea of clustering-based sentence ranking. Different from the previous work, we attempt to design a unified sentence scoring model to rank sentences and reduce redundancy at the same time.

## 3 Method

In this work, the density peaks sentence clustering (DPSC) method is designed for multi-document summarization.

### 3.1 Density Peaks Sentence Clustering

The density peaks clustering (DPC) algorithm is achieved upon the object similarity matrix. Objects are finally assigned density values and mini-distance values. In this work, we consider sentences as objects and follow the framework to calculate representativeness score and diversity score of each sentence in a unified model.

To construct the sentence similarity matrix for the DPC algorithm, we first segment documents into sentences and remove the non-stop words in the sentences. We then represent the sentences using bag-of-words vector space model, thus the cosine equation is applicable to calculate sentence similarity. The terms can be weighted with different schemes such as $boolean$ (occurring or not), $tf$ (term frequency) and $tf * isf$ (term frequency inverse sentence frequency). We finally choose the boolean scheme in our experiments because it performs best in our empirical study.

### 3.2 Representativeness Scoring

For document summarization, we need a representative score to quantify the degree how much a sentence is important in the documents. Enlightened by the DPC algorithm, we assume that when a sentence has more similar sentences (i.e., higher density), it will be considered more important or more representative. Thus we define the following function to calculate the representativeness score $s^{\text{REP}}(i)$ for each sentence $s_i$:

$$s^{\text{REP}}(i) = \frac{1}{K} \sum_{j=1, j \neq i}^{K} \chi(sim_{ij} - \delta), \qquad (1)$$

$$\chi(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

where $sim_{ij}$ denotes the similarity value between the $i$-th and $j$-th sentence, and $K$ denotes the number of sentences in the datasets. $\delta$ denotes a prede-

fined density threshold. Note that we set the density threshold following (Rodriguez and Laio, 2014), which attempts to exclude the sentences holding lower similarity with the current sentence.

### 3.3 Diversity Scoring

Most of the previous work handles diversity via reduce redundancy in a post processing module after the sentences are ranked. In this work, we measure diversity in the ranking model.

Diversity score of a sentence is measured by computing the minimum distance between the sentence $s_i$ and any other sentences with higher density score.

In order to reflect the above observation, we define the following function to calculate the diversity score $s^{\text{DIV}}(i)$:

$$s^{\text{DIV}}(i) = 1 - \max_{j:s^{\text{REP}}(j)>s^{\text{REP}}(i)} sim_{ij}. \quad (3)$$

For the sentence with the highest density, we conventionally take

$$s^{\text{DIV}}(i) = 1 - \min_{j\neq i} sim_{ij}. \quad (4)$$

The proposed diversity score looks similar to the famous *Maximum Marginal Relevance* (MMR) (Carbonell and Goldstein, 1998), which is widely used in removing redundancy by using a greedy algorithm to remove sentences that are too similar to the already selected ones. The difference lies that MMR selects a sentence by comparing it to those selected sentences while we compare it to all the other sentences in the dataset, thus it can enhance the diversity globally.

### 3.4 Length Scoring

It is widely accepted that summarization task has an important constraint, i.e., summary length. In order to satisfy this constraint, the length of selected sentences should be as short as possible. Based on this analysis, we propose the length score, which has relationship with the *effective length* and *real length*. The *real length* is defined as the number of word occurrences that a sentence contains. We then define the *effective length* as how many unique non-stop terms a sentence contains. We finally define the following function to calculate the length score $s^{\text{LEN}}(i)$.

The motivation to propose the length score is, shorter sentences with better representativeness score and diversity score are more favorable for the final summaries. Furthermore, as we use the Boolean scheme to measure sentence similarity, we only count unique words as effective sentence length.

$$s^{\text{LEN}}(i) = \frac{el(s_i)}{\max_{j=1}^{K} el(s_j)} \times \log \frac{\left(\max_{j=1}^{K} rl(s_j)\right)}{rl(s_i)}, \quad (5)$$

where $el(s_i)$ returns the *effective length* of sentence $s_i$, and $rl(s_i)$ the *real length* of sentence $s_i$.

### 3.5 Unified Sentence Scoring

Now we integrate representativeness score, diversity score and length score in the following unified sentence scoring function:

$$s^{\text{DPSC}}(i) = s^{\text{REP}}(i) \times s^{\text{DIV}}(i) \times s^{\text{LEN}}(i). \quad (6)$$

The assumption is obviously that we need those sentences which are as representative, diversified as possible and contain unique terms as many as possible within a limited length.

In calculation, we simply apply logarithm since:

$$s^{\text{DPSC}}(i) \sim \log s^{\text{REP}}(i) + \log s^{\text{DIV}}(i) + \log s^{\text{LEN}}(i) \quad (7)$$

### 3.6 Summary Generation

As three scores above including the representativeness, diversity and length constraint are measured in a unified sentence scoring model, generating a summary with out method is basically achieved by selecting the higher ranking sentences. In other words, our summary contains more representative and diversified information in the limited length.

**Complexity Analysis**: Suppose $K$ is the total number of sentences in the document collection. The complexity in calculating the sentence similarity matrix is $O(K^2)$. As the complexity in the function of representativeness scoring, diversity scoring and length scoring are all $O(K)$, the total time complexity of our DPSC method is $O(K^2) + O(K) + O(K) \sim O(K^2)$.

## 4 Evaluation

Two experiments are reported in this paper: comparing the MDS methods and tuning the density threshold. For both experiments, we use the DUC2004(task 2)[1] dataset, which is annotated manually for generic MDS. We adopted ROUGE (Lin, 2004) version 1.5.5[2] and take F-measure of ROUGE-1, ROUGE-2 and ROUGE-SU as our evaluation metrics. In pre-processing, we use the Porter Stemmer[3] in sentence segmenting, stop-word removing and word stemming. Note that our MDS method is purely unsupervised, and uses no training or development data.

### 4.1 The MDS Methods

We selected three categories of baselines[4]:

(1) DUC04 MDS methods: *DUC04Best* (Conroy et al., 2004).

(2) Clustering-based MDS methods: *Centroid* (Radev et al., 2004), *ClusterHITS* (Wan and Yang, 2008), *SNMF* (Wang et al., 2008), *RTC* (Cai and Li, 2013), *FGB* (Wang et al., 2011), and *AASum* (Canhasi and Kononenko, 2013).

(3) Other state-of-the-art MDS methods: *LexRank* (graph-based method) (Erkan and Radev, 2004), *CSFO* (optimization-oriented method) (Lin and Bilmes, 2011) and *WCS* (aggregation-oriented method) (Wang and Li, 2012).

For our *DPSC* method, we adopt the following settings: (1) Density threshold is set 0.22 as it is empirically found as optimal in Section 4.2 in the DUC04 dataset. (2) Term weighting scheme is set *Boolean*. In our experiments, *Boolean* is found outperforming *tf* and *tfisf* in sentence representation, this is because term repetition happens less frequently in short text units like sentences than that in documents. Experimental results of the MDS methods are presented in Table 1. Note the ROUGE values of some MDS methods are not reported in the literatures and marked with $'-'$ in Table 1.

According to Table 1, *DPSC* outperforms *DUC04Best*, which ignores the cross-sentence information to solve the diversity problem. *DPSC* outperforms most clustering-based methods except for *AASum*, which performs slightly better than *DPSC* on ROUGE-1. *AASum* is a very complex MDS method which fully exploits the advantages of clustering and the flexibility of matrix factorization. A weakness of the approach is that the number of archetypes must be predefined, and a post-processing module is required to reduce redundancy (Canhasi and Kononenko, 2013).

Table 1: Experimental results of the MDS methods on DUC04.

| System | ROUGE-1 | ROUGE-2 | ROUGE-SU |
|---|---|---|---|
| DUC04Best | 0.38224 | 0.09216 | 0.13233 |
| Centroid | 0.36728 | 0.07379 | 0.12511 |
| ClusterHITS | 0.36463 | 0.07632 | – |
| SNMF | – | 0.08400 | 0.12660 |
| RTC | 0.37475 | 0.08973 | – |
| FGB | 0.38724 | 0.08115 | 0.12957 |
| AASum | **0.41150** | 0.09340 | 0.13760 |
| LexRank | 0.37842 | 0.08572 | 0.13097 |
| CSFO | 0.38900 | – | – |
| WCS | 0.39872 | **0.09611** | 0.13532 |
| DPSC | 0.39075 | 0.09376 | **0.14000** |

*DPSC* also outperforms *LexRank* and *CSFO*, and yields close results compared with *WCS*. According to Table 1, *DPSC* performs slightly worse than *WCS*. The marginal performance gain of *WCS* comes from the aggregation strategy, namely, multiple MDS systems are required. As a comparison, *DPSC* is a pure and simple MDS method, exhibiting much lower complexity.

*DPSC* method is also advantageous on usability, because it does not involve any external resources such as Wordnet and Wikipedia or very complex natural language processing algorithms such as sentence parsing. Moreover, DPSC is a very fast MDS method. Thus it can be easily reproduced and deployed in real environment.

### 4.2 Density Threshold

Following (Rodriguez and Laio, 2014), we design an experiment on DUC04 dataset to investigate how the density threshold influences quality of the summaries. We tune the density threshold by varying it from 0.10 to 0.40(see the X-axis in Figure 1).

Figure 1 shows that on the specific dataset (i.e., DUC04), *DPSC* reaches the best ROUGE score

[1] http://duc.nist.gov/duc2004/tasks.html

[2] Options used: -a -c 95 -b 665 -m -n 4 -w 1.2

[3] http://tartarus.org/martin/PorterStemmer/

[4] Interested readers can refer to details in the references.

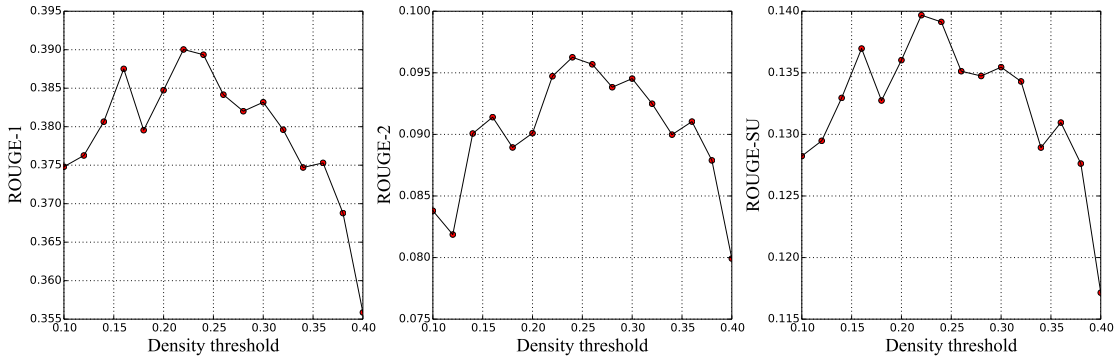


Figure 1: ROUGE curves of DPSC method varying the density threshold.

when the density threshold is set around 0.22 while starts to drop significantly after 0.30. This indicates that 0.22 is a good setting for the density threshold on DUC04.

## 5 Conclusion and Future Work

In this paper we report the density peaks sentence clustering (DPSC) method for multi-document summarization. Different from the prior work which deals with representativeness and redundancy independently, a unified sentence scoring model is designed in DPSC to combine the representativeness score, the diversity score and the length score of each sentence. Experimental results on DUC04 dataset show that DPSC outperforms the DUC04 best method and the existing clustering-based methods. Meanwhile, it yields close results when compared with the state-of-the-art generic MDS methods. It is thus verified that density peaks clustering algorithm is able to handle MDS effectively.

However, this work is still preliminary. We will study semantic text similarity to improve the sentence similarity matrix. We will then apply the proposed method in query-based multi-document summarization.

## Acknowledgement

This work is partially supported by Natural Science Foundation of China (61272233, 61373056, 61433018) and Shenzhen Peacock Scheme(183003656). We thank the reviewers for the insightful comments.

## References

Xiaoyan Cai and Wenjie Li. 2013. Ranking through clustering: An integrated approach to multi-document summarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(7):1424–1433.

Xiaoyan Cai, Wenjie Li, You Ouyang, and Hong Yan. 2010. Simultaneous ranking and clustering of sentences: a reinforcement approach to multi-document summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 134–142. Association for Computational Linguistics.

Ercan Canhasi and Igor Kononenko. 2013. Multi-document summarization via archetypal analysis of the content-graph joint model. *Knowledge and Information Systems*, pages 1–22.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM.

John M Conroy, Judith D Schlesinger, Jade Goldstein, and Dianne P Oleary. 2004. Left-brain/right-brain multi-document summarization. In *Proceedings of the Document Understanding Conference (DUC 2004)*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1):457–479.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM.

Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. 2009. Automatic generic document summarization

based on non-negative matrix factorization. *Information Processing & Management*, 45(1):20–34.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Alex Rodriguez and Alessandro Laio. 2014. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *IJCAI*, volume 7, pages 2862–2867.

Chao Shen, Tao Li, and Chris HQ Ding. 2011. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (plsa) with sentence bases. In *AAAI*.

Xiaojun Wan and Jianwu Yang. 2008. 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*, pages 299–306. ACM.

Dingding Wang and Tao Li. 2012. Weighted consensus multi-document summarization. *Information Processing & Management*, 48(3):513–523.

Dingding Wang, Tao Li, Shenghuo Zhu, and Chris Ding. 2008. 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*, pages 307–314. ACM.

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, and Yihong Gong. 2011. Integrating document clustering and multidocument summarization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(3):14.