Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Unsupervised document summarization from data reconstruction perspective

Zhanying He [a], Chun Chen [a,*], Jiajun Bu [a], Can Wang [a], Lijun Zhang [a], Deng Cai [b], Xiaofei He [b]

[a] Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University, Hangzhou 310027, China
[b] State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 2 April 2014
Received in revised form
2 July 2014
Accepted 20 July 2014
Communicated by M. Wang
Available online 28 September 2014

Keywords:
Document summarization
Data reconstruction
Manifold adaptive kernel

ABSTRACT

Due to its wide applications in information retrieval, document summarization is attracting increasing attention in natural language processing. A large body of recent literature has implemented document summarization by extracting sentences that cover the main topics of a document with a minimum redundancy. In this paper, we take a different perspective from data reconstruction and propose a novel unsupervised framework named *Document Summarization based on Data Reconstruction* (DSDR). Specifically, our approach generates a summary which consist of those sentences that can best reconstruct the original document. To model the relationship among sentences, we firstly introduce the linear reconstruction which approximates the document by linear combinations of the selected sentences. We then extend it into the non-negative reconstruction which allows only additive, not subtractive, linear combinations. In order to handle the nonlinear cases and respect the geometrical structure of sentence space, we also extend the linear reconstruction in the manifold adaptive kernel space which incorporates the manifold structure by using graph Laplacian. Extensive experiments on summarization benchmark data sets demonstrate that our proposed framework outperform state of the art.

1. Introduction

With the explosion of the textual information on the World Wide Web, people are overwhelmed by innumerable accessible documents. This means that we are in great need for technologies like document summarization that can better help users digest the information on the Web. Summarization techniques address this problem by condensing the document into a short piece of text covering the main topics. For example, search engines can provide users with snippets as the previews of the document contents, and help them to find the desired document. News sites usually describe hot news topics in concise headlines to facilitate browsing all news. Both the snippets and headlines are specific forms of document summary in real applications. Especially in the micro-blogging services, such as Twitter, Weibo and Tumblr, a hot topic can yield millions of short massages including enormous noises and redundancies. The possible solution is to summarize the massive tweets into a set of short text pieces covering the main topics [1].

Document summarization can be categorized as abstractive summaries or extractive summaries. Given a document, the abstractive summary is generated from complex natural language processing like information fusion, sentence compression and reformulation. Obviously, it is a difficult task for computer to automatically generate a satisfactory summary by abstraction. So the common practice is to perform extractive summarization in which a subset of existing sentences is used to form a final summary. Most of the existing generic summarization approaches use a ranking model to select sentences from a candidate set [2–4]. But these methods suffer from the redundancy problem in that top ranked sentences usually share much information in common. Although there are some methods [5–7] trying to reduce the redundancy, selecting sentences which have both good information coverage and minimum redundancy is a non-trivial task.

The motivation of our work is that the traditional methods usually solve the document summarization as a natural language problem rather than a data reconstruction problem although the second has been explored greatly in the literature of machine learning such as dimension reduction and feature selection. So in this paper, we propose a novel unsupervised summarization framework from the perspective of data reconstruction. As far as we know, our work is the first to treat the document summarization as a data reconstruction problem. We argue that a good summary should consist of those sentences that can best reconstruct the original document. Therefore, the reconstruction error becomes a natural criterion for measu ring the quality of summary. The new framework, namely *Document*

[*] Corresponding author.
E-mail addresses: hezhanying@zju.edu.cn (Z. He), chenc@zju.edu.cn (C. Chen), bjj@zju.edu.cn (J. Bu), wcan@zju.edu.cn (C. Wang), zljzju@zju.edu.cn (L. Zhang), dengcai@cad.zju.edu.cn (D. Cai), xiaofeihe@cad.zju.edu.cn (X. He).

http://dx.doi.org/10.1016/j.neucom.2014.07.046
0925-2312/© 2015 Published by Elsevier B.V.

Summarization based on Data Reconstruction (DSDR), finds the summary sentences by minimizing the reconstruction error. DSDR learns a reconstruction function for each candidate sentence of an input document and then formulates an objective function minimizing the error to obtain an optimal summary. The geometric interpretation is that DSDR tends to select sentences that span the intrinsic subspace of candidate sentence space, so that it is able to cover the core information of the document.

We firstly introduce the linear reconstruction to model the relationship between the document and the summary. The linear reconstruction aims to approximate the document by linear combinations of the selected summary sentences. Further, inspired by previous studies which indicate the existence of psychological and physiological evidence for parts-based representation in the human brain [8–10], we assume that document summary should consist of the parts of sentences, and introduce the non-negative constraints into the DSDR framework. With the non-negative constraints, our method leads to parts-based representation so that no redundant information needs to be subtracted from the combination. Still another issue to be addressed in document summary is the nonlinearity of the sentence space, as recent research [11] shows that the raw sentences are supposed to be highly nonlinear in distribution. The linear functions therefore lead to suboptimal fit in that neither the linear reconstruction nor the non-negative linear reconstruction respect the nonlinear manifold structure of sentence space. So we propose a novel nonlinear reconstruction which is performed in the manifold adaptive kernel space by using graph Laplacian [11–13]. By extracting sentences which can reconstruct the document in the kernel space, we are able to produce a better summary than the classical methods.

It is worthwhile to highlight the following three contributions of our proposed DSDR framework in this paper:

- We propose a novel unsupervised summarization framework from the perspective of data reconstruction which as we known is the first work to treat the document summarization from such a perspective.
- We firstly introduce the linear reconstruction and a greedy optimization method to solve the problem efficiently and effectively. Further, we propose the non-negative reconstruction and the corresponding iterative method to get a global optimum. To handle the nonlinearity, we finally propose the nonlinear reconstruction based on the manifold adaptive kernel.
- The proposed framework should not be restricted to the three types of reconstruction mentioned in this paper. Actually it is suitable for any other data reconstruction types. Since DSDR is unsupervised and language independent, it can be extended to summarize non-English document easily and even multi-language document.

This work is an extended and improved follow-up to our earlier work [14]. In comparison, we add a substantially theoretical analysis about extending DSDR in the manifold adaptive kernel space. For both linear reconstruction and non-negative linear reconstruction, the details of the mathematical translations are introduced additionally. We also extend the experiments here, such as implementing DSDR in the manifold adaptive kernel space and comparing it with existing approaches.

Our paper is organized as follows. We briefly review the related work in Section 2. In Section 3, we introduce the details of the *Document Summarization based on Data Reconstruction* (DSDR) including the optimization algorithms. Finally, we experimentally demonstrate the effectiveness of our proposed approaches in Section 4 and conclude in Section 5.

2. Related work

Recently, lots of extractive document summarization methods have been studied. Most of them involve assigning salient scores to sentences or paragraphs of the original document and composing the result summary of the top units with the highest scores. The computation rules of salient scores can be categorized into three groups [15]: feature based measurements, lexical chain based measurements and graph based measurements [4]. Salient scores in feature based measurements are usually related with various features such as term frequency, position, length, and topic presentation. The first method proposed in [16] ranks the sentences which are represented by the weighted term frequency vectors according to the relevance scores to the whole document. In the second type of measurements, a lexical chain is defined by a coherent sequence of related nouns, verbs and others. Sentence scores are then computed according to the lexical chain. In [17], the semantic relations of terms in the same semantic role are discovered by using the WordNet [18]. The relations are finally used in pairwise semantic similarity calculations which serve for the construction of their semantic similarity matrix. A tree pattern expression for extracting information from syntactically parsed text is proposed in [19]. In the graph based measurements, the sentence scores propagate around the graph on the basic idea that the score of one sentence affects scores of its neighbor sentences in the graph. Algorithms like PageRank [2] and HITS [3] are used in the sentence score propagation based on the graph constructed through the semantic affinity among sentences. In [4], it is also shown that this kind of measurements can improve single-document summarization by integrating multiple documents of the same topic.

Almost all the mentioned document summarization methods based on sentence scores have to incorporate with the adjustment of term weights which is one of the most important factors that influence summarization performance [20]. The adjustment process is used to eliminate the redundant information while it is not necessary when methods without saliency scores are applied in summarization. For extracting sentences, the methods without saliency scores include classification-based methods [21,22], clustering-based methods [23], as well as model-based methods [5–7]. Inspired by the latent semantic indexing (LSA), Ref. [16] applies the singular value decomposition (SVD) to select highly ranked sentences for generic document summarization. Besides, to improve summarization performance, there are some other studies like clustering sentences into topic themes, improving the topic representation and also time series text. Ref. [17] uses symmetric non-negative matrix factorization (SNMF) to cluster sentences into groups and selects sentences from each group for summarization. And [24] analyzes five different topic representations and proposes a novel topic representation based on topic themes. In [24], authors propose a novel symbolic representation of time series for text processing.

However, all the above summarization methods aim to obtain the summary which covers the core information, but few conduct the extractive task from the data reconstruction perspective. We believe that a good generic summary should contain those sentences that can best reconstruct the document. So how to best reconstruct the original document by the selected sentences is the main focus of the proposed DSDR in this study.

Notation: Small letters (*e.g.* x) denote scalars. Lowercase bold letters (*e.g.* $\mathbf{x}$) denote column vectors and $\|\cdot\|$ denotes the vector l_2-norm. Uppercase letters (*e.g.* X) denote matrices or graphs. The matrix trace is denoted by $\mathrm{Tr}(\cdot)$ and the Forbenius norm of a matrix is denoted by $\|\cdot\|_F$. Script uppercase letters (*e.g.* $\mathcal{X}$) denote ordinary sets and $|\mathcal{X}|$ is the size of the set. Blackboard bold capital letters (*e.g.* $\mathbb{R}$) denote number sets.

3. The proposed framework

Suppose we have a document and its summary as shown in Fig. 1. It can be found that a good summary should match the following two conditions. First, the selected sentences are able to cover most information of all sentences so that they can represent the original document. And we call the process of covering as "reconstruction". Second, the reconstruction of these sentences should be concise so that the summary will keep minimum redundancy. So we believe that a good summary should contain those sentences that can be used to reconstruct the document as well as possible, namely minimizing the reconstruction error.

In the following, we describe the details of our proposed framework *Document Summarization based on Data Reconstruction* (DSDR) which minimizes the reconstruction error for summarization. The algorithm procedure of DSDR is as follows:

- After stemming and stop-word elimination, we decompose the document into individual sentences and create a weighted term-frequency vector for every sentence. All the sentences form the *candidate set*.
- For the document (or, a set of documents), DSDR aims to find an optimal set of representative sentences to approximate the entire document (or, the set of documents), by minimizing the reconstruction error.

We denote the candidate sentence set as $V=[\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n]^T$ where $\mathbf{v}_i\in\mathbb{R}^d$ is a weighted term-frequency vector for sentence i. Here notice that we use V to represent both the matrix and the candidate set $\{\mathbf{v}_i\}$. Suppose there are totally d terms and n sentences in the document, we will have a matrix V in the size of $n\times d$. We denote the summary sentence set as $X=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_m]^T$ with $m<n$ and $X\subset V$.

Given a sentence $\mathbf{v}_i\in V$, DSDR attempts to represent it with a reconstruction function $f_i(X)$ given the selected sentence set X. Denoting the parameters of f_i as $\mathbf{a}_i$, we obtain the reconstruction error of $\mathbf{v}_i$ as

$$L(\mathbf{v}_i,f_i(X;\mathbf{a}_i))=\|\mathbf{v}_i-f_i(X;\mathbf{a}_i)\|^2,$$

where $\|\cdot\|$ is the L_2-norm.

By minimizing the sum of reconstruction errors over all the sentences in the document, DSDR picks the optimal set of representative sentences. The objective function of DSDR can be formally defined as

$$\min_{X,\mathbf{a}_i}\sum_{i=1}^{n}\|\mathbf{v}_i-f_i(X;\mathbf{a}_i)\|^2.$$

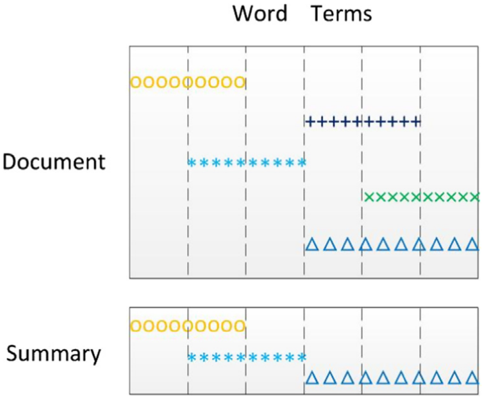


Fig. 1. Shown in this diagram, the columns separated by vertical dashes denote word terms, and the five lines denote sentences which covering different word terms. For example, the sentence denoted by circles covers the first and second word terms. Suppose we have a document with five sentences. Given the overlap between the triangle sentence and the other two, it is obvious that three sentences denoted by circles, stars and triangles will make a good summary. Because from these three sentences, we can obtain most of the information in a brief way.

The result summary must cover the main content so that it can reconstruct the original document and it must keep less redundancy so that it can minimize the reconstruction error.

3.1. Linear reconstruction

To model the relationship between sentences, we firstly define the reconstruction functions $f_i(X)$ as a linear function

$$f_i(X;\mathbf{a}_i)=\sum_{j=1}^{m}\mathbf{x}_j a_{ij},\quad X=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_m]^T. \tag{1}$$

Namely a candidate sentence $\mathbf{v}_i$ can be approximately represented as

$$\mathbf{v}_i\approx\sum_{j=1}^{m}\mathbf{x}_j a_{ij},\quad 1\le i\le n.$$

Now, the reconstruction error of the document can be obtained as

$$\sum_{i=1}^{n}\|\mathbf{v}_i-X^T\mathbf{a}_i\|^2$$

The solution from minimizing the above equation often exhibits high variance and results in high generalization error especially when the dimension of sentence vectors is smaller than the number of sentences. The variance can be reduced by shrinking the coefficients $\mathbf{a}_i$, if we impose a penalty on its size. Inspired by ridge regression [25], we penalize the coefficients of linear reconstruction error in DSDR as follows:

$$\begin{aligned}\min_{X,A}\quad & J=\sum_{i=1}^{n}\|\mathbf{v}_i-X^T\mathbf{a}_i\|^2+\lambda\|\mathbf{a}_i\|^2\\ \text{s.t.}\quad & X\subset V,\quad |X|=m\\ & A=[\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n]^T\in\mathbb{R}^{n\times m}.\end{aligned} \tag{2}$$

The set $\{\mathbf{x}_i\}$ includes the selected representative sentences from the original candidate sentence set V and will be used as the document summary finally. λ is the regularization parameter controlling the amount of shrinkage. In limited case, if $m=n$, the solution of (2) should be $X=V$ which reasonably means that the optimal summary with length of n is the document itself.

The optimization problem in (2) faces two combinatorial challenges:

- Evaluating the best reconstruction error of one candidate sentence $\mathbf{v}_i$, we would find the optimal X with the size of m out of exponentially many options.
- The optimal set for $\mathbf{v}_i$ is usually not optimal for $\mathbf{v}_j$. So to reconstruct all the candidate sentences, we would have to search over an exponential number of possible sets to determine the unique optimal X.

Actually, a similar problem that selects $m<n$ basic vectors from n candidates to approximate a single vector in the least squares criterion has been proved to be NP hard [26].

Inspired by the previous work in [27], the optimization problem in (2) is equivalent to the following problem:

$$\begin{aligned}\min_{X}\quad & J=\mathrm{Tr}[V(X^TX+\lambda I)^{-1}V^T]\\ \text{s.t.}\quad & X\subset V,\quad |X|=m\end{aligned} \tag{3}$$

where V is the candidate sentence set, X is the selected sentence set, I is the identity matrix, and $\mathrm{Tr}[\cdot]$ is the matrix trace calculation.

Proof. By fixing X and setting the derivative of (2) with respect to A to be zero

$$\frac{\partial J}{\partial A} = -2VX^T + 2AXX^T + 2\lambda A = 0,$$

we can obtain the optimal A^* as an expression of X

$$A^* = VX^T(XX^T + \lambda I)^{-1}.$$

Submitting this optimal A^* into the objective function (2), we can get

$$\begin{aligned}
&\mathrm{Tr}[(V-AX)(V-AX)^T] + \lambda\,\mathrm{Tr}(AA^T)\\
&\quad = \mathrm{Tr}(VV^T) - \mathrm{Tr}[VX^T(XX^T+\lambda I)^{-1}XV^T]\\
&\quad = \lambda\,\mathrm{Tr}[V(X^TX+\lambda I)^{-1}V^T]
\end{aligned}$$

where the Woodbury matrix identity [28]

$$(H+UCD)^{-1} = H^{-1} - H^{-1}U(C^{-1}+DH^{-1}U)^{-1}DH^{-1},$$

is applied in the second step with $H=\lambda I$, $U=X$, $C=I$ and $D=X^T$

$$\begin{aligned}
&\mathrm{Tr}[VX^T(XX^T+\lambda I)^{-1}XV^T]\\
&\quad = \frac{1}{\lambda}\mathrm{Tr}\left\{VX^T[I - X(X^TX+\lambda I)^{-1}X^T]XV^T\right\}\\
&\quad = \mathrm{Tr}\{V[I-\lambda(X^TX+\lambda I)^{-1}]V^T\}. \qquad \square
\end{aligned}$$

For the optimization problem of (3), we use a greedy algorithm to find the approximate solution. Given the previously selected sentence set X_1, DSDR selects the next new sentence $\mathbf{x}_i \in V$ as follows:

$$\begin{aligned}
&\min_{\mathbf{x}_i} \quad J(\mathbf{x}_i) = \mathrm{Tr}[V(X^TX+\lambda I)^{-1}V^T]\\
&\text{s.t.} \quad X = X_1 \cup \mathbf{x}_i, \quad \mathbf{x}_i \in V.
\end{aligned} \tag{4}$$

Denoting $P = X_1^TX_1 + \lambda I$, (4) can be rewritten as

$$\begin{aligned}
J(\mathbf{x}_i) &= \mathrm{Tr}[V(X^TX+\lambda I)^{-1}V^T]\\
&= \mathrm{Tr}[V(P+\mathbf{x}_i\mathbf{x}_i^T)^{-1}V^T]\\
&= \mathrm{Tr}\left[VP^{-1}V^T - \frac{VP^{-1}\mathbf{x}_i\mathbf{x}_i^TP^{-1}V^T}{1+\mathbf{x}_i^TP^{-1}\mathbf{x}_i}\right],
\end{aligned}$$

where the Woodbury matrix identity [28] is applied in the second step.

Since the candidate sentence set V and the selected sentence set X_1 are both fixed, $\mathrm{Tr}[VP^{-1}V^T]$ is a constant, so the objective function is the same as maximizing the second part in the trace

$$\max_{\mathbf{x}_i}\mathrm{Tr}\left[\frac{VP^{-1}\mathbf{x}_i\mathbf{x}_i^TP^{-1}V^T}{1+\mathbf{x}_i^TP^{-1}\mathbf{x}_i}\right] = \frac{\|VP^{-1}\mathbf{x}_i\|^2}{1+\mathbf{x}_i^TP^{-1}\mathbf{x}_i}.$$

To simplify the computation, we introduce a matrix $B = VP^{-1}V^T$. Then the index of the new sentence $\mathbf{x}_i$ can be obtained by

$$i = \arg\max_i \frac{\|B_{*i}\|^2}{1+B_{ii}},$$

where i is the index of the new sentence $\mathbf{x}_i$ in the candidate sentence set V, B_{*i} and B_{ii} are the ith column and diagonal entry of matrix B, respectively.

Once we find the new sentence $\mathbf{x}_i$, we add it into X_1 and update the matrix B as follows:

$$\begin{aligned}
B^t &= VP_t^{-1}V^T\\
&= V(P_{t-1}+\mathbf{x}_i\mathbf{x}_i^T)^{-1}V^T\\
&= B^{t-1} - \frac{B_{*i}^{t-1}[B_{*i}^{t-1}]^T}{1+B_{ii}^{t-1}}.
\end{aligned} \tag{5}$$

where the matrix B^{t-1} denotes the matrix B at the step $t-1$.

Algorithm 1. DSDR with linear reconstruction.

```
Input:
    • The candidate data set: $V = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]^T$
    • The number of sentences to be selected: $m$
    • The trade off parameter: $\lambda$
Output:
    • The set of $m$ summary sentences: $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^T \subseteq V$
1:  initialize $X \leftarrow \varnothing$;
2:  $B^0 \leftarrow VV^T/\lambda$;
3:  for $t=1$ to $m$ do
4:      for $i=1$ to $n$ and $\mathbf{x}_i \notin X$ do
5:          $score(\mathbf{x}_i) \leftarrow \|B_{*i}^{t-1}\|^2/(1+B_{ii}^{t-1})$
6:      end for
7:      $\mathbf{x}_i \leftarrow \arg\max_{\mathbf{x}_i} score(\mathbf{x}_i)$
8:      $X \leftarrow X \cup \mathbf{x}_i$
9:      $B^t \leftarrow B^{t-1} - B_{*i}^{t-1}[B_{*i}^{t-1}]^T/(1+B_{ii}^{t-1})$
10: end for
11: return $X$;
```

Initially the previously selected sentence set X_1 is empty. So the matrix P is initialized as

$$P_0 = \lambda I.$$

Then the initialization of the matrix B can be written as

$$B^0 = VP_0^{-1}V^T = \frac{1}{\lambda}VV^T.$$

We describe our sequential method for linear reconstruction in Algorithm 1. Given a document with n sentences and each sentence $\mathbf{v}_i \in \mathbb{R}^d$, Algorithm 1 generates a summary with m sentences with the complexity as follows:

- $O(n^2d)$ to calculate the initialization B^0 according to Step (2).
- $O(n^2m)$ for the Steps (3)–(10) where $O(n)$ to calculate $score(\mathbf{x}_i)$ in Step (5) and $O(n^2)$ to update the matrix B in Step (9).

The overall cost for Algorithm 1 is $O(n^2(d+m))$.

3.2. Non-negative linear reconstruction

The linear reconstruction optimization problem (2) in the previous section might come up with a_{ij}'s with negative values, which means that redundant information needs to be removed from the summary sentence set X. As shown in Fig. 2, suppose we have three summary sentences and one original sentence, we can obtain the combination parameters on the right. We can find that the negative value of a_{ij} meaning that there exists redundant information to be subtracted from the reconstruction. To compose a final summary, the two sentences denoted by stars and pluses are enough. This indicates that better optimization can be achieved by adding nonnegative constraint for a_{ij}.

Non-negative constraints on data representation has received considerable attention due to its psychological and physiological interpretation of naturally occurring data whose representation may be parts-based in the human brain [8–10]. Our non-negative linear reconstruction method leads to parts-based reconstruction because it allows only additive, not subtractive, combinations of the sentences.

In order to solve the optimization problem, we convert the objective function into a relaxed formula

$$\begin{aligned}
&\min_{\mathbf{a}_i} \quad J = \sum_{i=1}^{n}\|\mathbf{v}_i - V^T\mathbf{a}_i\|^2 + 2\sum_{j=1}^{n}\sqrt{\gamma\sum_{i=1}^{n}a_{ij}^2}\\
&\text{s.t.} \quad a_{ij} \ge 0 \quad \text{and} \quad \mathbf{a}_i \in \mathbb{R}^n.
\end{aligned}$$

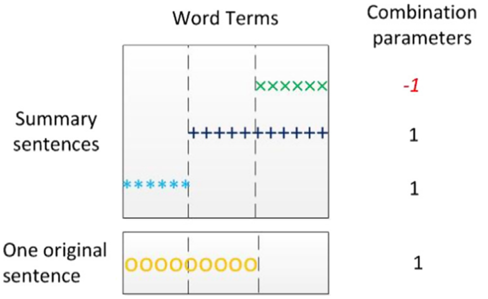


Fig. 2. Similar as Fig. 1, the columns denote word terms and the lines denote sentences. Suppose we have three summary sentences and one original sentence from the document. Because of the overlap between the sentences, it is obvious to obtain the linear combination parameters on the right. Actually, we can see that the two sentences denoted by stars and pluses are enough to compose a final summary. Therefore, the sentences with negative combination parameters can be moved from the summary.

Notice that the vector of $\mathbf{a}_i$ was a vector with size m, and now its size becomes n. We can find that $X^T\mathbf{a}_i$ and $V^T\mathbf{a}_i$ are the same if some elements of $\mathbf{a}_i$ are zeros. In order to guarantee that some elements of $\mathbf{a}_i$ will be zeros, we use the group sparse regularization. This regularization will result that some elements of $\mathbf{a}_1$, $\mathbf{a}_2$ and all other $\mathbf{a}_j$ to be zeros. So that the relaxed formula will approximate the original objective function.

One challenge introduced by this group sparse regularization is non-differentiability. As many other group sparse problems [29,30], we use a relaxation inequality

$$2\sum_{j=1}^{n}\sqrt{\gamma\sum_{i=1}^{n}a_{ij}^2}\le\sum_{j=1}^{n}\frac{\sum_{i=1}^{n}a_{ij}^2}{\beta_j}+\gamma\|\boldsymbol{\beta}\|_1,\quad \beta_j\ge 0.$$

For the sake of efficient optimization, following [31], we formulate the objective function of non-negative DSDR as follows:

$$\min_{\mathbf{a}_i,\boldsymbol{\beta}}\quad J=\sum_{i=1}^{n}\left\{\|\mathbf{v}_i-V^T\mathbf{a}_i\|^2+\sum_{j=1}^{n}\frac{a_{ij}^2}{\beta_j}\right\}+\gamma\|\boldsymbol{\beta}\|_1$$
$$\text{s.t.}\quad \beta_j\ge 0,\quad a_{ij}\ge 0\quad\text{and}\quad \mathbf{a}_i\in\mathbb{R}^n, \tag{6}$$

where $\boldsymbol{\beta}=[\beta_1,\dots,\beta_n]^T$ is an auxiliary variable to control the candidate sentences selection. Similar to LASSO [29], the L_1 norm of $\boldsymbol{\beta}$ will enforce some elements to be zeros. If $\beta_j=0$, then all $a_{1j},\dots,a_{nj}$ must be 0 which means the j-th candidate sentence is not selected. The new formulation in (6) is a convex problem and can guarantee a global optimal solution.

By fixing $\mathbf{a}_i$'s and setting the derivative of J with respect to $\boldsymbol{\beta}$ to be zero

$$\frac{\partial J}{\partial\beta_j}=\gamma-\sum_{i=1}^{n}\frac{a_{ij}^2}{\beta_j^2}=0,$$

we can obtain the minimum solution of $\boldsymbol{\beta}$

$$\beta_j=\sqrt{\frac{\sum_{i=1}^{n}a_{ij}^2}{\gamma}} \tag{7}$$

since $\beta_j\ge 0$ as stated in (6); once the $\boldsymbol{\beta}$ is obtained, the minimization under the non-negative constraints can be solved using the Lagrange method. Let α_{ij} be the Lagrange multiplier for constraint $a_{ij}\ge 0$ and $A=[a_{ij}]$, the Lagrange L is

$$L=J+\mathrm{Tr}[\alpha A^T],\quad \alpha=[\alpha_{ij}].$$

The derivative of L with respect to A is

$$\frac{\partial L}{\partial A}=-2VV^T+2AVV^T+2A\ \mathrm{diag}(\boldsymbol{\beta})^{-1}+\alpha,$$

where $\mathrm{diag}(\boldsymbol{\beta})$ is a matrix with diagonal entries of $\beta_1,\dots,\beta_n$. Setting the above derivative to be zero, α can be represented as

$$\alpha=2VV^T+2AVV^T-2A\ \mathrm{diag}(\boldsymbol{\beta})^{-1}.$$

Using the Kuhn–Tucker condition $\alpha_{ij}a_{ij}=0$, we get

$$(VV^T)_{ij}a_{ij}-(AVV^T)_{ij}a_{ij}-(A\ \mathrm{diag}(\boldsymbol{\beta})^{-1})_{ij}a_{ij}=0.$$

This leads to the following updating formula:

$$a_{ij}\leftarrow\frac{a_{ij}(VV^T)_{ij}}{[AVV^T+A\ \mathrm{diag}(\boldsymbol{\beta})^{-1}]_{ij}}. \tag{8}$$

The formulations in (7) and (8) are iteratively performed until convergence. For the convergence of this updating formula, we have the following Theorem 1.

Theorem 1. *Under the iterative updating rule as (8), the objective function J is non-increasing with fixed* $\boldsymbol{\beta}$*, and that the convergence of the iteration is guaranteed.*

Proof. To prove Theorem 1, we introduce an auxiliary function as

$$G(\mathbf{u},\mathbf{a}_i)=\sum_{j=1}^{n}\left\{\frac{(C\mathbf{a}_i)_j}{a_{ij}}u_j^2-2(VV^T)_{ij}u_j\right\},$$

where $C=VV^T+\mathrm{diag}(\boldsymbol{\beta})^{-1}$, and $\mathbf{u}=[u_1,\dots,u_n]^T$ is a positive vector. $G(\mathbf{u},\mathbf{a}_i)$ can also be identified as the sum of singular-variable functions

$$G(\mathbf{u},\mathbf{a}_i)=\sum_{j=1}^{n}G_j(u_j).$$

Let $F(\mathbf{a}_i)=\mathbf{a}_i^TC\mathbf{a}_i-2(VV^T)_{i*}\mathbf{a}_i$, Sha et al. [32] have proved that if a_{ij} updates as:

$$a_{ij}\leftarrow\arg\min_{u_j}G_j(u_j),$$

$G(\mathbf{u},\mathbf{a}_i)$ converges monotonically to the global minimum of $F(\mathbf{a}_i)$.

Algorithm 2. DSDR with non-negative linear reconstruction.

```
Input:
    • The candidate sentence set: $V=[\mathbf{v}_1,\mathbf{v}_2,\dots,\mathbf{v}_n]^T$
    • The trade off parameter: $\gamma>0$
Output:
    • The set of the summary sentences: $X\subseteq V$
Procedure:
    initialize $a_{ij}$, $\beta_j$;
    initialize $X\leftarrow\varnothing$;
    repeat
        $\beta_j=\sqrt{\frac{\sum_{i=1}^{n}\mathbf{a}_{ij}^2}{\gamma}}$;
        repeat
            $a_{ij}\leftarrow\frac{a_{ij}(VV^T)_{ij}}{[AVV^T+A\ \mathrm{diag}(\boldsymbol{\beta})^{-1}]_{ij}}$;
        until converge;
    until converge;
    $X\leftarrow\{\mathbf{v}_j|\mathbf{v}_j\subset V,\beta_j\ne 0\}$;
    return $X$;
```

Taking the derivation of $G_j(u_j)$ with respect to u_j and setting it to be zero, we obtain the updating formulation as

$$a_{ij}\leftarrow\frac{a_{ij}(VV^T)_{ij}}{[AVV^T+A\ \mathrm{diag}(\boldsymbol{\beta})^{-1}]_{ij}}, \tag{9}$$

which agrees with (8).

We can rewrite the objective function J as

$$J=\sum_{i=1}^{n} F(\mathbf{a}_i)+\mathrm{Tr}[VV^T]+\gamma\|\boldsymbol{\beta}\|_1. \tag{10}$$

Fixing $\boldsymbol{\beta}$, we can obtain the minimizer of J by minimizing each $F(\mathbf{a}_i)$ separately. Since the objective function J is the sum of all the individual terms $F(\mathbf{a}_i)$ plus a term independent of $\mathbf{a}_i$, we have shown that J is non-increasing with fixed $\boldsymbol{\beta}$ under the updating rule as (8). □

Algorithm 2 describes the DSDR with non-negative linear reconstruction. Suppose the maximum number of iterations for Step (4) and Step (6) are t_1 and t_2 respectively, the total computational cost for Algorithm 2 is $O(t_1(n+t_2(n^3)))$.

3.3. DSDR in manifold adaptive kernel space

Since the original sentence space is believed to be an nonlinear sub-manifold embedded in the ambient space [33], the linear reconstruction will lead to a sub-optimal fit. We thus need to extend our algorithm to consider the nonlinear geometric structure in the sentence space. The kernel trick is a usual way for discovering the nonlinear structure in the data by mapping the original nonlinear observations into a higher dimensional inner product space [34]. It can represent an implicit mapping of the sentences in a higher dimensional space. The most commonly used kernels include Gaussian kernel and polynomial kernel. Let $\mathcal{V}\in\mathbb{R}^d$ denote the candidate sentence space and $\mathcal{H}$ be the reproducing kernel Hibert space (RKHS) [35]. The feature mapping function $\varphi:\mathcal{V}\rightarrow\mathcal{H}$ is implicitly induced by a kernel function $\mathcal{K}:\mathcal{V}\times\mathcal{V}\rightarrow\mathbb{R}$ that defines the similarity between sentences in the original space. It can be shown that if there is a kernel function $\mathcal{K}(\cdot,\cdot)$ then the function $\varphi(\cdot)$ and the feature space $\mathcal{H}$ exist [36], and furthermore the kernel function is as follows:

$$\mathcal{K}(\mathbf{v}_i,\mathbf{v}_j)=\langle\varphi(\mathbf{v}_i),\varphi(\mathbf{v}_j)\rangle,\quad \mathbf{v}_i,\mathbf{v}_j\in\mathbb{R}^d. \tag{11}$$

With the kernel method, we can deal with the problem by a kernel function instead of the feature mapping $\varphi(\cdot)$. In the transformed feature space $\mathcal{H}$, we denote the candidate sentences by $\varphi(V)=[\varphi(\mathbf{v}_1),\dots,\varphi(\mathbf{v}_n)]^T$ and the summary sentences by $\varphi(X)=[\varphi(\mathbf{x}_1),\dots,\varphi(\mathbf{x}_m)]^T$. Since the choice of $\mathcal{K}$ is flexible, we use Gaussian kernel in our experimental setting. Namely the kernel function is formed as

$$\mathcal{K}(\mathbf{v}_i,\mathbf{v}_j)=e^{-\|\mathbf{v}_i-\mathbf{v}_j\|^2/2\sigma^2},$$

where σ is the parameter for Gaussian kernel.

However, such nonlinear structure captured by the data independent kernels may not be consistent with the intrinsic manifold structure, such as geodesic distance, curvature, and homology [11]. Vikas et al. [37] construct a family of data-dependent norms and propose the manifold adaptive kernel. Let $\mathcal{O}$ be a linear space with a positive semi-definite inner product (quadratic form) and let $S:\mathcal{H}\rightarrow\mathcal{O}$ be a bounded linear operator. We define $\tilde{\mathcal{H}}$ to be the space of functions from $\mathcal{H}$ with the modified inner product

$$\langle f,g\rangle_{\tilde{\mathcal{H}}}=\langle f,g\rangle_{\mathcal{H}}+\langle S(f),S(g)\rangle_{\mathcal{O}}.$$

Vikas et al. have shown that $\tilde{\mathcal{H}}$ is still a RKHS [37]. Given the sentences $\mathbf{v}_1,\dots,\mathbf{v}_n$, let $S:\mathcal{H}\rightarrow\mathbb{R}^n$ be the evaluation map

$$S(f)=(f(\mathbf{v}_1),\dots,f(\mathbf{v}_n))^T.$$

Denote $\mathbf{f}=(f(\mathbf{v}_1),\dots,f(\mathbf{v}_n))^T$ and $\mathbf{g}=(g(\mathbf{v}_1),\dots,g(\mathbf{v}_n))^T$. Notice that $\mathbf{f},\mathbf{g}\in\mathcal{O}$, thus we have

$$\langle S(f),S(g)\rangle_{\mathcal{O}}=\langle\mathbf{f},\mathbf{g}\rangle=\mathbf{f}^TM\mathbf{g}$$

where M is a positive semi-definite matrix. We define the ith column of the kernel matrix as

$$K_{*i}=(\mathcal{K}(\mathbf{v}_i,\mathbf{v}_1),\dots,\mathcal{K}(\mathbf{v}_i,\mathbf{v}_n))^T.$$

It can be shown that the reproducing kernel in $\tilde{\mathcal{H}}$ is

$$\tilde{\mathcal{K}}(\mathbf{v}_i,\mathbf{v}_j)=\mathcal{K}(\mathbf{v}_i,\mathbf{v}_j)-\lambda K_{*i}^T(I+MK)^{-1}MK_{*j}$$

where I is an identity matrix, K is the kernel matrix in $\mathcal{H}$, and $\lambda\geq 0$ is a constant controlling the smoothness of the functions. The choice of M is the key issue which makes the adaptive kernel be data-dependent. The graph Laplacian matrix is usually proposed since it models the underlying geometrical structure of the data.

In order to model the manifold structure, we construct a nearest neighbor graph G. For each sentence $\mathbf{v}_i$, we find its k nearest neighbors denoted by $N(\mathbf{v}_i)$ and put an edge between $\mathbf{v}_i$ and its neighbors. There are many choices for the weight matrix on the graph. A simple one is as follows [37,11]:

$$W_{ij}=\begin{cases}1 & \text{if } \mathbf{v}_i\in N(\mathbf{v}_j) \text{ or } \mathbf{v}_j\in N(\mathbf{v}_i);\\ 0 & \text{otherwise.}\end{cases}$$

The graph Laplacian [12,13,38] is defined as $L=D-W$ where D is a diagonal degree matrix given by $D_{ii}=\sum_j W_{ij}$. By setting $M=L$, we eventually get the following manifold adaptive kernel [37,11,39]:

$$\tilde{\mathcal{K}}(\mathbf{v}_i,\mathbf{v}_j)=\mathcal{K}(\mathbf{v}_i,\mathbf{v}_j)-\lambda K_{*i}^T(I+LK)^{-1}LK_{*j}.$$

In RKHS, the objective function of linear reconstruction can be rewritten as follows:

$$\min_{X,A}\ \sum_{i=1}^{n}\|\varphi(\mathbf{v}_i)-\varphi(X)^T\mathbf{a}_i\|^2+\lambda\|\mathbf{a}_i\|^2$$
$$\text{s.t.}\quad \varphi(X)\subset\varphi(V),\quad A=[\mathbf{a}_1,\mathbf{a}_2,\dots,\mathbf{a}_n]^T.$$

Given the previously selected sentence set $\varphi(X_1)$, the sequential optimization seeks the next new sentence $\varphi(\mathbf{x}_i)\subset\varphi(V)$ by

$$\min_{\varphi(\mathbf{x}_i)}\ J=\{\varphi(V)[\varphi(X)^T\varphi(X)+\lambda I]^{-1}\varphi(V)^T\}$$
$$\text{s.t.}\quad \varphi(X)=\varphi(X_1)\cup\varphi(\mathbf{x}_i),\quad \varphi(\mathbf{x}_i)\in\varphi(V).$$

Defining $P=\varphi(X)^T\varphi(X)+\lambda I$ and initializing $P_0=\lambda I$, we again introduce the matrix $B=\varphi(V)P^{-1}\varphi(V)^T$. Then the initialization of B is

$$B^0=\varphi(V)P_0^{-1}\varphi(V)^T=\tilde{K}/\lambda,\quad \tilde{K}_{ij}=\tilde{\mathcal{K}}(\mathbf{v}_i,\mathbf{v}_j).$$

And iteratively update B as follows:

$$B^t=B^{t-1}-\frac{B_{*i}^{t-1}[B_{*i}^{t-1}]^T}{1+B_{ii}^{t-1}},$$

where t denotes the tth iteration, B_{*i} and B_{ii} are the ith column and diagonal entries of matrix B, respectively.

At each step t, the index of the new sentence $\varphi(\mathbf{x}_i)$ can be selected by

$$i=\arg\max_i\frac{\|B_{*i}\|^2}{1+B_{ii}}.$$

Setting $t\leq m$, we can get the summary sentences iteratively. The sequential algorithm in kernel space is similar to that mentioned before, except that the input should include the definition of the kernel function $\mathcal{K}$ which is directly related to the initialization of the matrix B.

4. Experiments

In this study, we use the standard summarization benchmark data sets DUC 2006 and DUC 2007 for the evaluation. DUC 2006 and DUC 2007 contain 50 and 45 document sets respectively, with

25 news articles in each set. The sentences in each article have been separated by NIST.[1] And every sentence is either used in its entirety or not at all for constructing a summary. The length of a result summary is limited by 250 tokens (whitespace delimited).

4.1. Evaluation metric

We use the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit [40] which has been widely adopted by DUC for automatic summarization evaluation. ROUGE measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the peer summary (produced by algorithms) and the model summary (produced by humans). We choose two automatic evaluation methods ROUGE-N and ROUGE-L in our experiment. Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries and ROUGE-L uses the longest common subsequence (LCS) metric. ROUGE-N is computed as follows:

$$ROUGE-N = \frac{\sum_{S \in Ref} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in Ref} \sum_{gram_n \in S} Count(gram_n)}$$

where n stands for the length of the n-gram, Ref is the set of reference summaries. $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, and $Count(gram_n)$ is the number of n-grams in the reference summaries. Among these different scores, the unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most [41]. ROUGE toolkit reports separate scores for 1, 2, 3 and 4-gram, and also for the longest common subsequence (ROUGE-L). Given a reference summary of u sentences containing a total of m words and a result summary C of v sentences containing a total of n words, the F-measure of ROUGE-L score can be computed as follows:

$$ROUGE-L_R = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{m}$$

$$ROUGE-L_P = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, C)}{n}$$

$$ROUGE-L_F = \frac{2 \times ROUGE-L_R \times ROUGE-L_P}{ROUGE-L_R + ROUGE-L_P}$$

where r_i is a reference summary sentence and $LCS_{\cup}(r_i, C)$ is the LCS score of the union longest common subsequence between reference sentence r_i and candidate summary C. For example, if $r_i = w_1 w_2 w_3 w_4 w_5$, and C contains two sentences: $c_1 = w_1 w_2 w_6 w_7 w_8$ and $c_2 = w_1 w_3 w_8 w_9 w_5$, then the longest common subsequence of r_i and c_1 is "$w_1 w_2$" and the longest common subsequence of r_i and c_2 is "$w_1 w_3 w_5$". The union longest common subsequence of r_i, c_1, and c_2 is "$w_1 w_2 w_3 w_5$" and $LCS_{\cup}(r_i, C) = 4/5$. More information can be referred to the toolkit package [40].

4.2. Compared methods

To the best of our knowledge, our work is the first approach which treats the document summarization as a sentence reconstruction problem. It is important to note that our algorithm is unsupervised. Thus we do not compare with those supervised summarization systems [42–45]. Because they need to train the summarization model using human labeled training data. We compare our DSDR with several unsupervised state-of-the-art summarization approaches described briefly as follows:

- *Random*: Selects sentences randomly for each document set. We implement the random function offered by MATLAB.
- *Lead* [46]: For each document set, orders the documents chronologically and takes the leading sentences one by one.
- *LSA* [16]: Applies the singular value decomposition (SVD) on the terms by sentences matrix to select highest ranked sentences.
- *ClusterHITS* [47]: Considers the topic clusters as hubs and the sentences as authorities, then ranks the sentences with the authorities scores. Finally, the highest ranked sentences are chosen to constitute the summary. The number of clusters is the same as the number of document sets, namely 50 and 45 for DUC 2006 and DUC 2007, respectively.
- *SNMF* [17]: Uses symmetric non-negative matrix factorization (SNMF) to cluster sentences into groups and select sentences from each group for summarization. The group number is the same as the number of document sets.

For convenience, we denote DSDR with the linear reconstruction, DSDR with the non-negative reconstruction and DSDR in the manifold adaptive kernel space by "DSDR-lin", "DSDR-non" and "DSDR-adap" respectively.

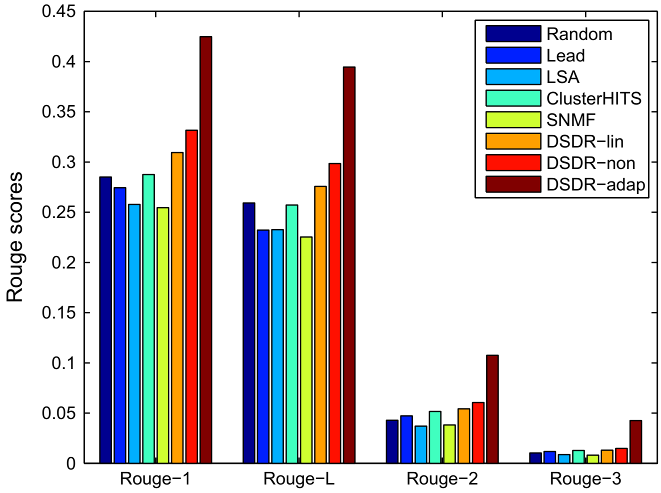


Fig. 3. Average *F*-measure performance on DUC 2006.

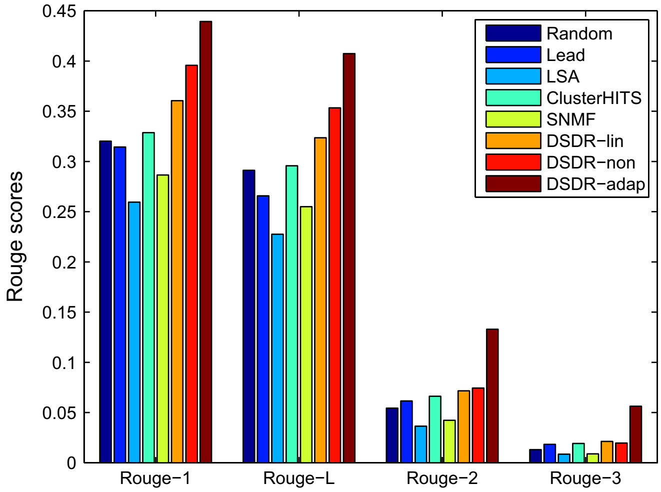


Fig. 4. Average *F*-measure performance on DUC 2007.

[1] http://www.nist.gov/index.html.

4.3. Experimental results

4.3.1. Overall performance comparison

ROUGE can generate three types of scores: recall, precision and *F*-measure. We get similar experimental results using the three types with DSDR taking the lead. In this study, we use *F*-measure to show the experimental results. The *F*-measure of four ROUGE metrics are shown in our experimental results: ROUGE-1, ROUGE-2, ROUGE-3 and ROUGE-L. Figs. 3 and 4 plot the ROUGE scores vs. different evaluation metrics, and show the results on DUC 2006 and DUC 2007 data sets respectively.

As shown in the two figures, the three proposed approaches outperform other compared algorithms in all evaluation metrics. It is worthwhile to notice that DSDR-adap performs especially good since it receives the highest scores which are much greater than the others. By utilizing both the reconstruction relationships and the sentences' manifold structure in the adaptive kernel space, DSDR-adap is able to find those sentences that can reconstruct the original document in the sentence-dependent kernel space. Since DSDR-lin obtains a suboptimal solution and DSDR-non gets the global optimum, the evaluation scores of DSDR-lin is a little lower than that of DSDR-non. The experimental results confirm the

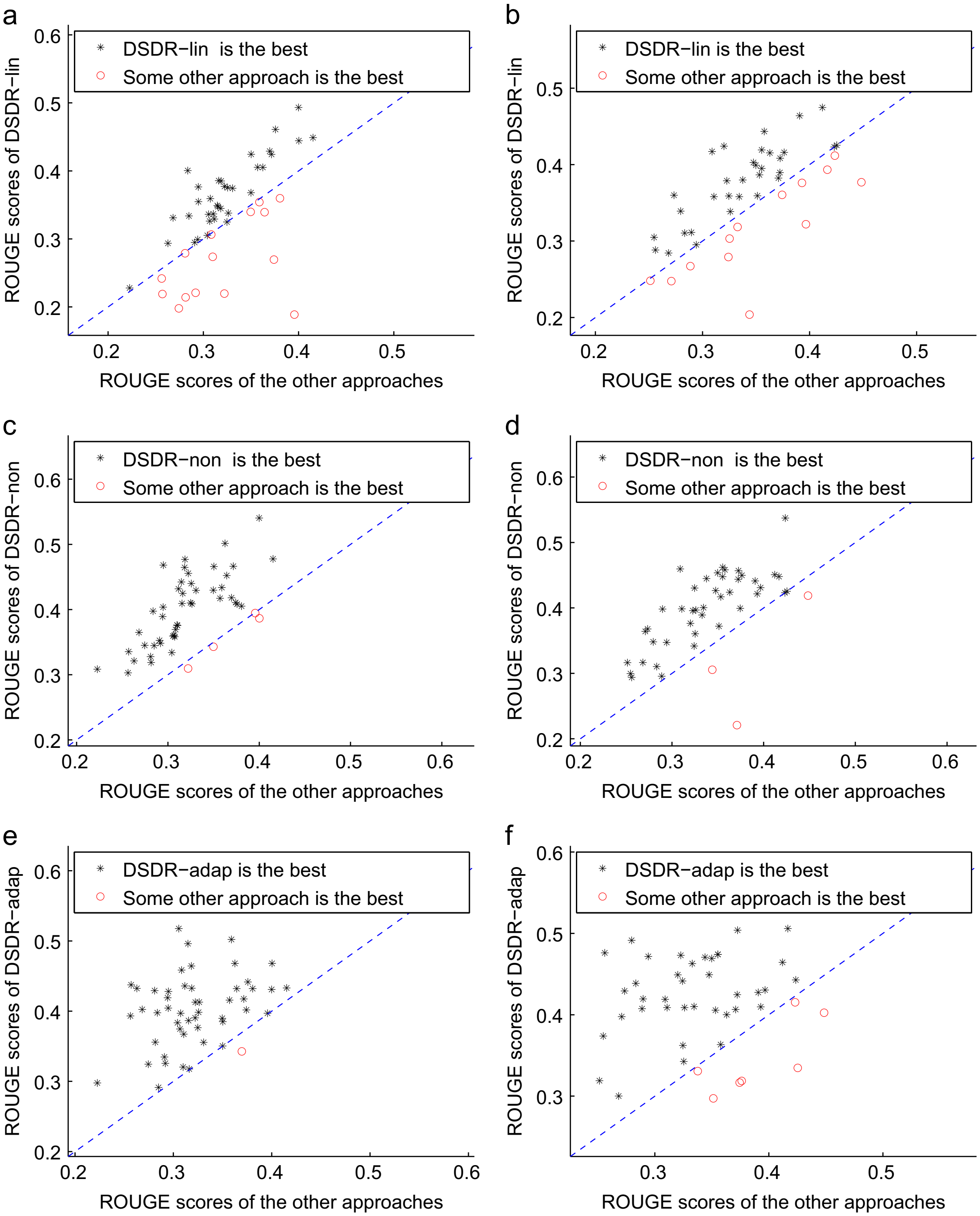


Fig. 5. The scores of all algorithms on each document set of DUC 2006 and DUC 2007, the black stars denote our proposed methods have the greatest scores while the red circles denote otherwise. (a) ROUGE scores on DUC 2006, (b)ROUGE scores on DUC 2007, (c) ROUGE scores on DUC 2006, (d) ROUGE scores on DUC 2007, (e) ROUGE scores on DUC 2006 and (f) ROUGE scores on DUC 2007. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

algorithms we presented in the last section that DSDR-lin is solved greedily and is suboptimal; DSDR-non is solved globally and is better; DSDR-adap considers the non-linear cases in kernel space and is the best. Following our three proposed approaches, ClusterHITS gets the highest scores among the remaining. ClusterHITS considers topics as hubs and sentences as authorities where hubs and authorities can interact with each other. With the interactions between hubs and authorities, the correlations between topics and sentences can be explored to improve the quality of summary. Besides, selecting sentences randomly is a little better than just selecting the leading sentences. Because it is equal for every sentence by selecting them randomly. Among all the summarization algorithms, LSA and SNMF show the poorest performance on both data sets. Directly applying SVD on the terms by sentences matrix, summarization by LSA chooses those sentences with the largest indexes along the orthogonal latent semantic directions. Although SNMF relaxes the orthogonality, it relies on the sentence pairwise similarity. Whereas our DSDR selects sentences which span the intrinsic subspace of the candidate sentence space. Such sentences are contributive to reconstruct the original document, and so are contributive to improve the summary quality.

4.3.2. *Evaluations on different document sets*

In Fig. 5, we illustrate the ROUGE-1 scores for each document set from DUC 2006 and DUC 2007. In each panel, the vertical axis describes the scores of the DSDR approach and the horizontal axis contains the best scores of other methods. The black stars indicate that DSDR gets the best scores on the corresponding document sets while the red circles suggest the best scores are obtained from other methods. It can be obviously observed that the proposed reconstruction approaches are better than others, since the number of black stars are much more than that of red circles in each panel. And again, DSDR-non outperforms DSDR-lin since the numbers of black stars in panel (c) and (d) are more than that in panel (a) and (b). Compared with DSDR-non, DSDR-adap gets more black stars on DUC 2006 but less on DUC 2007. The reason might be that though DSDR-adap can handle the reconstruction problem in nonlinear space, it cannot obtain the global optimum. So it is an interesting future study to extend the non-negative reconstruction in the manifold adaptive space.

In order to check whether the difference between DSDR and other approaches is significant, we perform the paired *t*-test between the ROUGE-1 scores of DSDR and that of other approaches on both data sets. Tables 1 and 2 show the associated *p*-values on DUC 2006 and DUC 2007 data sets, respectively. For example, in Table 1, the value of 4.6×10^{-14} in row two and column two means the associated *p*-value of the paired *t*-test between DSDR-lin and Random. As can be seen, all the tested values are close to zero. So at nearly 100% confidence interval, the test demonstrates that our proposed framework can obtain very encouraging and promising results compared to the others. Moreover, the values in lines 2 and 3 are lower than line 1 in both tables except the last one in Table 2 which further prove that DSDR-non and DSDR-adap can get better results than DSDR-lin.

Table 1
The associated *p*-values of the paired *t*-test on DUC 2006.

	Random	Lead	LSA	ClusterHITS	SNMF
DSDR-lin	4.6×10^{-14}	7.1×10^{-6}	9.2×10^{-14}	4.0×10^{-9}	9.3×10^{-8}
DSDR-non	2.6×10^{-25}	6.7×10^{-17}	2.3×10^{-30}	6.0×10^{-23}	1.8×10^{-25}
DSDR-adap	1.1×10^{-17}	2.7×10^{-19}	1.0×10^{-20}	5.6×10^{-20}	1.4×10^{-18}

Table 2
The associated *p*-values of the paired *t*-test on DUC 2007.

	Random	Lead	LSA	ClusterHITS	SNMF
DSDR-lin	5.2×10^{-14}	1.7×10^{-8}	5.6×10^{-12}	3.4×10^{-10}	1.9×10^{-9}
DSDR-non	2.5×10^{-17}	8.0×10^{-13}	1.4×10^{-14}	7.9×10^{-15}	1.1×10^{-14}
DSDR-adap	1.0×10^{-14}	2.2×10^{-11}	6.1×10^{-16}	1.4×10^{-14}	6.2×10^{-7}

5. Conclusion

In this paper, we propose a novel unsupervised summarization framework called the *Document Summarization based on Data Reconstruction* (DSDR) which selects the most representative sentences that can best reconstruct the entire document. We introduce the linear reconstruction firstly and extend it in two different ways (non-negative and manifold adaptive kernel). The experimental results show that our DSDR framework can outperform other state-of-the-art summarization approaches. DSDR with linear reconstruction is more efficient while DSDR with nonnegative reconstruction has better performance (by generating less redundant sentences). We also show that extending the linear reconstruction in the manifold adaptive kernel space can get excellent summary. Because it models the underlying geometrical structure of the sentences by using the graph Laplacian.

In the future, we are interested in several problems. First, like other kernel based methods, the computational complexity of DSDR-adap scales with the number of sentences. So it might not be suitable to large-scale document with many sentences. Second, it would be expected to efficiently develop DSDR with non-negative reconstruction in the manifold adaptive kernel space.

Acknowledgment

This work was supported in part by National Basic Research Program of China (973 Program) under Grant 2013CB336500, National Natural Science Foundation of China (Grant nos. 61125203, 61173185, 90920303, and 61173186), Zhejiang Provincial Natural Science Foundation of China (Grant nos. Y1101043 and LZ13F020001) and Foundation of Zhejiang Provincial Educational Department under Grant Y201018240.

References

[1] L. Shou, Z. Wang, K. Chen, G. Chen, Sumblr: continuous summarization of evolving tweet streams, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013, pp. 533–542.

[2] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. doi:http://dx.doi.org/10.1016/j.comnet.2012.10.007.

[3] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632. http://dx.doi.org/10.1145/324133.324140.

[4] X. Wan, J. Yang, Collabsum: exploiting multiple document clustering for collaborative single document summarizations, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, ACM, 2007, pp. 143–150. doi:http://dx.doi.org/10.1145/1277741.1277768.

[5] J.M. Conroy, D.P. O'leary, Text summarization via hidden Markov models, in: Proceedings of the 24th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR '01, ACM, 2001, pp. 406–407.

[6] S. Park, J.-H. Lee, D.-H. Kim, C.-M. Ahn, Multi-document summarization based on cluster using non-negative matrix factorization, in: SOFSEM 2007: Theory and Practice of Computer Science, Lecture Notes in Computer Science, vol. 4362, Springer, Berlin, Heidelberg, 2007, pp. 761–770.

[7] D. Shen, J. Sun, H. Li, Q. Yang, Z. Chen, Document summarization using conditional random fields, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007, pp. 2862–2867.

[8] S.E. Palmer, Hierarchical structure in perceptual representation, Cogn. Psychol. 9 (4) (1977) 441–474. http://dx.doi.org/10.1016/0010-0285(77)90016-0.

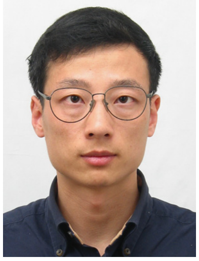
[9] E. Wachsmuth, M. Oram, D. Perrett, Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque, Cereb. Cortex 4 (5) (1994) 509.

[10] D. Cai, X. He, J. Han, T.S. Huang, Graph regularized non-negative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1548–1560.
[11] D. Cai, X. He, Manifold adaptive experimental design for text categorization, IEEE Trans. Knowl. Data Eng. 24 (4) (2012) 707–719. http://dx.doi.org/10.1109/TKDE.2011.104.
[12] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 14.
[13] X. He, P. Niyogi, Locality preserving projections, in: Adv. Neural Inf. Process. Syst. 16 (2003).
[14] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, X. He, Document summarization based on data reconstruction, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, AAAI, 2012, pp. 620–626.
[15] M. Hu, A. Sun, E.-P. Lim, Comments-oriented document summarization: understanding documents with readers' feedback, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, ACM, 2008, pp. 291–298. doi:http://dx.doi.org/10.1145/1390334.1390385.
[16] Y. Gong, X. Liu, 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, SIGIR '01, ACM, 2001, pp. 19–25. doi:http://dx.doi.org/10.1145/383952.383955.
[17] 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, SIGIR '08, ACM, 2008, pp. 307–314. doi: http://dx.doi.org/10.1145/1390334.1390387.
[18] G.A. Miller, Wordnet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41. http://dx.doi.org/10.1145/219717.219748.
[19] Y. Choi, Tree pattern expression for extracting information from syntactically parsed text corpora, Data Min. Knowl. Discov. 22 (2011) 211–231.
[20] A. Nenkova, L. Vanderwende, K. McKeown, A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, ACM, 2006, pp. 573–580. doi:http://dx.doi.org/10.1145/1148170.1148269.
[21] M. Amini, P. Gallinari, The use of unlabeled data to improve supervised learning for text summarization, in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2002, pp. 105–112.
[22] J. Kupiec, J. Pedersen, F. Chen, A trainable document summarizer, in: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1995, pp. 68–73.
[23] T. Nomoto, Y. Matsumoto, A new approach to unsupervised text summarization, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2001, pp. 26–34.
[24] S. Harabagiu, F. Lacatusu, Topic themes for multi-document summarization, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05, ACM, 2005, pp. 202–209. doi:http://dx.doi.org/10.1145/1076034.1076071.
[25] A. Hoerl, R. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.
[26] B. Natarajan, Sparse approximate solutions to linear systems, SIAM J. Comput. 24 (2) (1995) 227–234.
[27] K. Yu, J. Bi, V. Tresp, Active learning via transductive experimental design, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, ACM, 2006, pp. 1081–1088. doi:http://dx.doi.org/10.1145/1143844.1143980.
[28] K. Riedel, A Sherman–Morrison–Woodbury identity for rank augmenting matrices with application to centering, SIAM J. Matrix Anal. Appl. 13 (2) (1992) 659–662.
[29] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc.: Ser. B (Methodol.) (1996) 267–288.
[30] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 68 (1) (2005) 49–67.
[31] K. Yu, S. Zhu, W. Xu, Y. Gong, Non-greedy active learning for text categorization using convex transductive experimental design, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08, ACM, 2008, pp. 635–642. doi:http://dx.doi.org/10.1145/1390334.1390442.
[32] F. Sha, Y. Lin, L. Saul, D. Lee, Multiplicative updates for nonnegative quadratic programming, Neural Comput. 19 (8) (2007) 2004–2031.
[33] X. He, D. Cai, J. Han, Learning a maximum margin subspace for image retrieval, IEEE Trans. Knowl. Data Eng. 20 (2) (2008) 189–201.
[34] B. Settles, M. Craven, An analysis of active learning strategies for sequence labeling tasks, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, Association for Computational Linguistics, 2008, pp. 1070–1079.
[35] A. Berlinet, C. Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, Springer, 2004.
[36] B. Scholkopf, A. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond, The MIT Press, 2002.
[37] V. Sindhwani, P. Niyogi, M. Belkin, Beyond the point cloud: from transductive to semi-supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, ACM, 2005, pp. 824–831. doi: http://dx.doi.org/10.1145/1102351.1102455.
[38] P. Li, J. Bu, C. Chen, Z. He, D. Cai, Relational multi-manifold co-clustering, IEEE Trans. Cybern. (2014), in press. doi:http://dx.doi.org/10.1109/TSMCB.2012.2234108.
[39] P. Li, C. Chen, J. Bu, Clustering analysis using manifold kernel concept factorization, Neurocomputing 87 (2012) 120–131.
[40] C. Lin, Rouge: a package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.
[41] C.-Y. Lin, E. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, NAACL '03, Association for Computational Linguistics, 2003, pp. 71–78. doi:http://dx.doi.org/10.3115/1073445.1073465.
[42] K. Toutanova, C. Brockett, M. Gamon, J. Jagarlamudi, H. Suzuki, L. Vanderwende, The pythy summarization system: microsoft research at duc 2007, in: Proceedings of DUC, 2007.
[43] A. Haghighi, L. Vanderwende, Exploring content models for multi-document summarization, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, Association for Computational Linguistics, 2009, pp. 362–370.
[44] A. Celikyilmaz, D. Hakkani-Tur, A hybrid hierarchical model for multi-document summarization, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, Association for Computational Linguistics, 2010, pp. 815–824.
[45] H. Lin, J. Bilmes, A class of submodular functions for document summarization, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, HLT '11, Association for Computational Linguistics, 2011, pp. 510–520.
[46] M. Wasson, Using leading text for news summaries: evaluation results and implications for commercial summarization applications, in: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 2, ACL '98, Association for Computational Linguistics, 1998, pp. 1364–1368. doi:http://dx.doi.org/10.3115/980691.980791.
[47] 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, SIGIR '08, ACM, 2008, pp. 299–306. doi:http://dx.doi.org/10.1145/1390334.1390386.

Zhanying He received the BS degree in Software Engineering from Zhejiang University, China, in 2009. She is currently a candidate for a PhD degree in computer science at Zhejiang University. Her research interests include information retrieval, data mining and machine learning.

Chun Chen received the BS degree in Mathematics from Xiamen University, China, in 1981, and his MS and PhD degrees in Computer Science from Zhejiang University, China, in 1984 and 1990, respectively. He is a professor in the College of computer science, Zhejiang University. His research interests include information retrieval, data mining, computer vision, computer graphics and embedded technology.

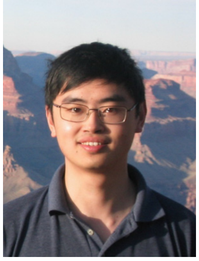
Jiajun Bu received the BS and PhD degrees in computer science from Zhejiang University, China, in 1995 and 2000, respectively. He is a professor in the College of Computer Science, Zhejiang University. His research interests include embedded system, data mining, information retrieval and mobile database.



Can Wang received the BS degree in economics, MS and PhD degrees in computer science from Zhejiang University, China, in 1995, 2003 and 2009, respectively. He is currently a faculty member in the College of Computer Science at Zhejiang University. His research interests include information retrieval, data mining and machine learning.

Lijun Zhang received the BS and PhD degrees in computer science from Zhejiang University, China, in 2007 and 2012, respectively. He worked as a postdoc in Michigan State University, USA, from 2012 to 2014. He is currently an associate professor in computer science at Nanjing University, China. His research interests include machine learning, information retrieval, and data mining.

Deng Cai is a professor in the State Key Lab of CAD&CG, College of Computer Science at Zhejiang University, China. He received the PhD degree in computer science from University of Illinois at Urbana Champaign in 2009. Before that, he received his Bachelor's degree and a Master's degree from Tsinghua University in 2000 and 2003, respectively, both in automation. His research interests include machine learning, data mining and information retrieval.



Xiaofei He received the BS degree in computer science from Zhejiang University, China, in 2000 and the PhD degree in computer science from the University of Chicago, in 2005. He is a professor in the State Key Lab of CAD&CG at Zhejiang University, China. Prior to joining Zhejiang University in 2007, he was a research scientist at Yahoo! Research Labs, Burbank, CA. His research interests include machine learning, information retrieval, and computer vision. He is a senior member of IEEE.