Affinity-Preserving Random Walk for Multi-Document Summarization

Kexiang Wang, Tianyu Liu, Zhifang Sui and **Baobao Chang**
Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
Collaborative Innovation Center for Language Ability, Xuzhou 221009 China
{wkx,tianyu0421,szf,chbb}@pku.edu.cn

Abstract

Multi-document summarization provides users with a short text that summarizes the information in a set of related documents. This paper introduces affinity-preserving random walk to the summarization task, which preserves the affinity relations of sentences by an absorbing random walk model. Meanwhile, we put forward adjustable affinity-preserving random walk to enforce the diversity constraint of summarization in the random walk process. The ROUGE evaluations on DUC 2003 topic-focused summarization task and DUC 2004 generic summarization task show the good performance of our method, which has the best ROUGE-2 recall among the graph-based ranking methods.

1 Introduction

Multi-document summarization provides users with summary that reflects the main information in a set of given documents. The documents are often related and talk about more than one topics. Generic multi-document summarization and topic-focused multi-document summarization are two typical kinds of summarization. The former is a summarization delivering the main information of the documents with no bias while the latter is a one delivering the main information biased to a given topic description (a few sentences or even phrases). Most existing summarization systems are designed for these two kinds of summarization.

There are two goals for generic multi-document summarization. The first one is **saliency**. Summary sentences should be central sentences that capture the majority of information in a document cluster. The sentences with little information about the document cluster should not be included in the summary. The second one is **diversity**. The information overlap between summary sentences should be as minimal as possible due to the length limit of summary. In other words, the information coverage of summary is a determinant, which requires that the summary sentences should cover diverse aspects of information. Besides the two goals, there is another goal for the topic-focused summarization and that is **relevancy**. It requires that the summary sentences be relevant to the topic description. A series of conferences and workshops on automatic text summarization (e.g. NTCIR, DUC), special topic sessions in ACL, EMNLP and SIGIR have advanced the techniques to achieve these goals and many approaches have been proposed so far.

In this paper, we focus on the extractive summarization methods, which extract the summary sentences from the input document cluster. We propose affinity-preserving random walk for multi-document summarization. The method is a graph-based ranking method, which takes into account the global information collectively computed from the entire sentence affinity graph. Different from the previous graph-based ranking methods, our method adopts "global normalization" to transform sentence affinity matrix into sentence transition matrix and formulates the sentence ranking process in an absorbing random walk model. Meanwhile, the adjustable affinity-preserving random walk is proposed to facilitate the diversity of summary by adjusting the sentence transition matrix after each iteration of random walk. Experimental results on DUC generic and topic-focused multi-document summarization tasks show the competitive performance of our method. To our best knowledge, our system has the best ROUGE-2 recall on both tasks among all existing graph-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 210–220
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

based ranking methods, which defeats most other methods.

We summarize our contributions as follows. (1) We preserve the original affinity relations between sentences in a novel affinity-preserving random walk view for multi-document summarization. The preservation of affinity leads to a more salient summary. And it is applicable to both generic and topic-focused summarization. (2) We propose adjustable affinity-preserving random walk to enforce the diversity constraint of summarization in the random walk process. (3) Experiments on DUC 2003 and DUC 2004 tasks demonstrate the competitive performance of our method.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 describes traditional random walk model for summarization. Section 4 proposes affinity-preserving random walk for the saliency goal of summarization and this section also proposes adjustable affinity-preserving random walk to produce both salient and diverse summary. Section 5 gives our evaluation results and the conclusion is made in Section 6.

2 Related Work

Our method belongs to the graph-based ranking methods to select sentences in the documents to produce the summary. Erkan and Radev (2004) proposed LexPageRank to compute the sentence saliency based on the concept of eigenvector centrality. It constructs the sentence affinity graph and computes the sentence saliency based on an algorithm similar to PageRank (Page et al., 1999). Like PageRank, the affinity matrix is converted into the row-stochastic matrix, which is used as the transition matrix of random walk on the weighted graph. Wan (2007) proposed manifold ranking for topic-focused multi-document summarization. It makes full use of both the relationships among all sentences in the documents and the relationships between the given topic description and the sentences. Manifold ranking conducts a different normalization on the sentence affinity matrix to guarantee the algorithm's convergence. GRASSHOPPER (Zhu et al., 2007) ranks sentences with an emphasis on the diversity constraint of summarization. It turns already ranked sentences into absorbing states, which effectively prevents redundant sentences from receiving a high rank. The algorithm is based on an absorbing random walk and produces only one summary sentence after one particular random walk becomes stationary. And the normalization from sentence affinity matrix to sentence transition matrix is the same as PageRank. DivRank (Mei et al., 2010) is a method to balance the saliency and diversity of the top ranked sentences in a reinforced random walk model. Also, the normalization in DivRank from affinity matrix to transition matrix is the same as PageRank. Another notable diversified graph-based ranking method GCD (Dubey et al., 2011) relies on large amounts of training data to learn edge conductances.

Our method formulates the multi-document summarization as an affinity-preserving random walk and uses the "global normalization" to transform sentence affinity matrix into sentence transition matrix, which is different from all those proposed methods. And the adjustable transition matrix in our method balances the saliency and diversity goals of summarization. Like GRASSHOPPER, our method relies on the absorbing random walk model. The difference is that our method does not turn the sentence vertex into absorbing state but add an absorbing vertex to the original sentence affinity graph. And all summary sentences are extracted after the random walk reaches a stationary state in our method. Like DivRank, the sentence transition matrix is adjustable in our method to enforce the diversity constraint of summarization. However, our method differs from DivRank in the mechanism to adjust the transition matrix.

3 Traditional Random Walk for Summarization

Suppose $G = (S, E)$ is a graph with vertex set S and edge set $E \subset S^2$. Suppose there is a conductance $c(s_i, s_j) > 0$ associated with each edge $(s_i, s_j) \in E$ and $c(s_i, s_j) = 0$ associated with the set $S^2 - E$ (the conductance of nonexistent edge is zero). Let

$$C(s_i) = \sum_{s_j \in S} c(s_i, s_j),\ s_i \in S \tag{1}$$

so that $C(s_i)$ is the total conductance of the edges coming from s_i. And the traditional random walk on graph is defined as

Definition 3.1. *The discrete-time Markov chain* $X = (X_0, X_1, X_2, ...)$ *with state space and tran-*

sition probability matrix $\mathbf{P}$ *given by*

$$P(s_i, s_j) = \frac{c(s_i, s_j)}{C(s_i)},\ (s_i, s_j) \in S^2 \quad (2)$$

is called a random walk on the graph G.

This chain governs a particle moving along the vertices of G. If the particle in the state X_m is at vertex $s_i \in S$, it will be at a neighbor of s_i in the next state X_{m+1}, which is chosen randomly in proportion to the conductance. We can prove that $\sum_{s_j \in S} P(s_i, s_j) = 1$ for any $s_i \in S$ ($C(s_i) \neq 0$) so $\mathbf{P}$ is a row-stochastic matrix by $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with entries $\mathbf{D}_{ii} = C(s_i)$ and $\mathbf{W}$ is the adjacency matrix of G where $\mathbf{W}_{ij} = c(s_i, s_j)$.

For the summarization task, G is the sentence affinity graph. The vertex set $S = \{s_1, s_2, ..., s_n\}$ contains every sentence in the document cluster and the edge set E contains the pairwise affinity between sentences. We use the tf*isf formula to calculate the weight associated with each term occurring in the sentence, where tf is the term frequency in the sentence and isf is the inverse sentence frequency of the term among all sentences. isf is often calculated as $1+log(n/n_t)$, where n is the total number of sentences and n_t is the number of sentences containing the term t. $\mathbf{W}_{ij}$ is computed using the standard cosine measure (Baeza-Yates et al., 1999).

$$\mathbf{W}_{ij} = sim_{cosine}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \times \|\mathbf{v}_j\|_2} \quad (3)$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the corresponding term vectors of s_i and s_j. Two vertices are connected if their affinity is larger than 0 and $\mathbf{W}_{ii}$ is set as 0 to avoid self transition. We get an undirected graph G as well as a symmetric sentence affinity matrix $\mathbf{W}$ in this way. Then we transform $\mathbf{W}$ into $\mathbf{P}$ by $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$ and use the stationary distribution of random walk as sentence ranking scores. The traditional random walk model is a simple practice of PageRank algorithm for multi-document summarization.

4 Affinity-Preserving Random Walk for Summarization

4.1 Prior of Multi-Document Summarization

In the above traditional random walk on graph, the normalization from affinity matrix $\mathbf{W}$ to transition matrix $\mathbf{P}$ is to make $\mathbf{P}$ a row-stochastic matrix. This can be interpreted as a democratic normalization because the surfer of a traditional random walk visits neighbors of a vertex with probability 1. The surfer has to choose a neighbor to visit next although it is a random choice w.r.t. the conductance distribution of the vertex. However this democratic normalization is not suitable for multi-document summarization due to the fact that most sentences are not salient and should not be normalized democratically as the few salient ones. The prior here is that the number ratio of good candidate sentences over bad candidate sentences is very low due to the summary length limit. Good candidate sentences are the sentences highly overlapping with sentences in the reference summary written by humans. And the remaining sentences are bad candidate sentences. The democratic normalization of $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$ will extend the adverse effect of bad candidate sentences and suppress the effect of good candidate sentences, because the total conductance of the bad candidate sentence is usually smaller than that of the good candidate sentence. In this case, the random surfer has to choose a neighbor to visit even when she is currently at a bad candidate sentence, which will direct her to visit other bad candidate sentences neighboring to the current sentence. The invariant behavior of the surfer at all vertices in the graph is not consistent with the prior which makes a distinction between good and bad candidate sentences. It may pervert the random walk process to get an ideal distribution in which only few sentences are assigned with a high ranking score.

It is worth noting here that manifold ranking (Wan et al., 2007) for summarization uses a different normalization: $\mathbf{P} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$. It is a symmetric normalization on both endpoints of an edge, which makes $\mathbf{P}$ a suitable choice in the manifold ranking process to smooth the scores of neighboring vertices. The symmetric normalization can be deduced from the objective function of manifold ranking (Zhou et al., 2003) and does not make a distinction between the good and bad candidate sentences. It is also not consistent with the prior. We can conclude that existing graph-based ranking methods can not well characterize the prior of multi-document summarization.

4.2 Affinity-Preserving Random Walk

We need a new normalization method that distinguishes between good and bad candidate sen-

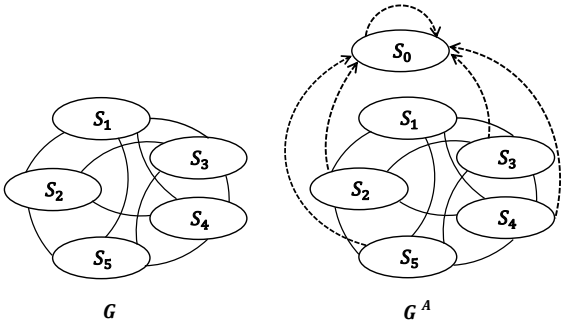

Figure 4.1: Sentence graphs for summarization. G: sentence affinity graph constructed from the document cluster. G^A: sentence augmented graph with an absorbing vertex s_0. C_{max} equals to $C(s_1)$ indicating that sentence s_1 has the maximum conductance, so there is no edge (s_1, s_0).

tences to satisfy the prior of multi-document summarization. Affinity-preserving random walk has an intrinsic mechanism that preserves the original affinity relations between sentences in the documents, which is proposed and defined as follows.

Definition 4.1. *For the graph G, the vertex set S has $(n+1)$ vertices: $s_0, s_1, s_2, ..., s_n$. The maximum conductance $C_{max} = max_{i=1,2,...n}C(s_i)$. Of the $(n+1)$ vertices, s_0 is an absorbing vertex with $c(s_i, s_0) \geqslant 0, c(s_0, s_i) = 0$, and $c(s_0, s_0) = C_{max}$ for $i = 1, 2, ..., n$. The remaining vertices are the normal vertices with $c(s_i, s_j) \geqslant 0$ for $i, j = 1, 2, ..., n$. The discrete-time Markov chain $X = (X_0, X_1, X_2, ...)$ with state space and transition probability matrix $\boldsymbol{P}$ given by*

$$\begin{aligned} &P(s_i, s_j) = \frac{c(s_i, s_j)}{C_{max}} \\ &P(s_i, s_0) = 1 - \frac{C(s_i)}{C_{max}}\,,\ P(s_0, s_i) = 0 \\ &P(s_0, s_0) = 1 \\ &\textit{for } i, j = 1, 2, ..., n \end{aligned} \tag{4}$$

is called an affinity-preserving random walk on the graph G.

For our summarization task, we construct a sentence augmented graph G^A (as shown in Figure 4.1) by adding an absorbing vertex s_0 to the sentence affinity graph G. The unabsorbed vertices s_i $(i = 1, 2, ..., n)$ represent sentences in the documents. The affinity-preserving random walk process as defined above is implemented on G^A to rank sentences in the documents. In the affinity-preserving random walk, once the surfer reaches the absorbing vertex, she cannot walk out of there. Because $P(s_i, s_0)$ is small for the vertex s_i with a large conductance, it is less likely for the surfer at s_i to walk into s_0. As for the vertex with a small conductance, the surfer has a tendency to be absorbed by s_0. The absorbing vertex here plays a role of soaking unreliable ranking scores from large numbers of bad candidate sentences and highlighting the few good candidate sentences. The affinity matrix $\mathbf{W}$ is normalized by its first norm (equivalent to C_{max}) in the affinity-preserving random walk, which results in a kind of "*soft*" stochastic matrix for n unabsorbed vertices. "*soft*" here means that the sum of row elements in the matrix can be less than 1. By contrast, $\mathbf{P}$ in the traditional random walk is a "*hard*" stochastic matrix in which every sum of row elements has to be 1. Meanwhile, $\mathbf{P}$ in this absorbing Markov chain (Seneta, 2006) preserves the original affinity relations in $\mathbf{W}$ as all sentences are globally normalized by the same factor (i.e. C_{max}). We call this approach an "affinity-preserving random walk" as it is used in (Cho et al., 2010), which deals with a graph matching problem that aims at assigning 1-vs-1 correspondences between two graphs. The similar idea is also applied in the case of ontology matching (Xiang et al., 2015). Transition matrix $\mathbf{P}$ including the absorbing vertex is formulated in (Cho et al., 2010) as follows

$$\mathbf{P} = \begin{pmatrix} 1 & \mathbf{0}^T \\ \mathbf{e} - \mathbf{c}/\|\mathbf{W}\|_1 & \mathbf{W}/\|\mathbf{W}\|_1 \end{pmatrix} \tag{5}$$

where $\mathbf{0}^T$ is a $1 \times n$ vector with all elements 0, $\mathbf{e}$ is an $n \times 1$ vector with all elements 1, $\mathbf{c} = [C(s_1), C(s_2), ..., C(s_n)]^T$ is a vector containing the conductances of n sentences and $\mathbf{W}/\|\mathbf{W}\|_1$ is the $n \times n$ soft substochastic matrix. However, the stationary distribution of such a random walk on graph with one absorbing vertex is always $\begin{pmatrix} 1 & \mathbf{0}^T \end{pmatrix}$, which is not a good characterization of the sentence ranking distribution. We turn to the quasi-stationary distribution $\bar{\mathbf{x}}$ (Cho et al., 2010; Darroch and Seneta, 1965) of absorbing random walk for ranking sentences. $\bar{\mathbf{x}}^{(K)}$ is defined as

$$\begin{aligned} \bar{\mathbf{x}}_i^{(K)} &= Prob(X^{(K)} = s_i | X^{(K)} \neq s_0) \\ &= \frac{\mathbf{x}_i^{(K)}}{\sum_j \mathbf{x}_j^{(K)}} \end{aligned} \tag{6}$$

where $X^{(K)}$ denotes the position of random surfer at time K. It can be proven that $\bar{\mathbf{x}}^{(K)}$ has its

stationary distribution $\bar{\mathbf{x}}$ if $\mathbf{W}$ is irreducible (Darroch and Seneta, 1965). We remove the sentences that have the total conductance 0 (i.e. the isolated sentences) when constructing the sentence affinity graph G. In this way, G will be strongly connected and has an irreducible adjacency matrix $\mathbf{W}$.

We introduce the teleport vector $\mathbf{y}$ as used in personalized PageRank (Page et al., 1999; Haveliwala, 2002; Jeh and Widom, 2003) for the summarization task. In the generic summarization case, we define the vector $\mathbf{y}$ in a way that reflects the position of each sentence in a document. If the sentence s_{i+1} is right after the sentence s_i in the same document, then

$$\frac{\mathbf{y}_{i+1}}{\mathbf{y}_i} = \lambda^{-1}, \lambda > 1 \tag{7}$$

where λ is the decay factor. In the topic-focused summarization case, we incorporate the topic description as a vertex in the random walk process, which is a standard way of achieving the relevancy goal of this kind of summarization. Vector $\mathbf{y}$ is defined to be $[y_1, y_2, ..., y_n, y_{n+1}]^T$ in which $y_i = 0 (1 \leqslant i \leqslant n)$ and $y_{n+1} = 1$ when the first n elements represent sentences in the documents and the last one represents the topic description. We normalize $\mathbf{y}$ by its first norm to get a prior sentence ranking for multi-document summarization. Based on $\mathbf{W}$ and $\mathbf{y}$, sentence ranking scores in affinity-preserving random walk can be formulated in a recursive form as follows

$$\mathbf{x} = \frac{\mu \mathbf{W}^T / \|\mathbf{W}\|_1 \mathbf{x} + (1-\mu)\mathbf{y}}{\|\mu \mathbf{W}^T / \|\mathbf{W}\|_1 \mathbf{x} + (1-\mu)\mathbf{y}\|_1} \tag{8}$$

where $\mathbf{x} = [Score(s_i)]_{n\times 1}$ is the vector of sentence ranking scores. μ is the damping factor that trades off between two actions: the transition according to $\mathbf{W}^T / \|\mathbf{W}\|_1$ and the teleport specified by $\mathbf{y}$. Transpose operation in Eq.(8) can be removed because of symmetry of $\mathbf{W}$. The final transition matrix of affinity-preserving random walk is given by $\mathbf{A} = \mu \mathbf{W} / \|\mathbf{W}\|_1 + (1-\mu)\mathbf{y} \cdot \mathbf{e}^T$ and $\mathbf{x}$ should be normalized by its first norm after each iteration of random walk. Like PageRank, the quasi-stationary distribution is obtained by the normalized principal eigenvector of $\mathbf{A}$.

For implementation, the initial ranking scores of all sentences are set to $1/n$ and the iterative process in Eq.(8) is adopted to compute new ranking scores of sentences. Usually convergence of the iterative algorithm is achieved when the difference between scores computed at two successive iterations falls below a given threshold.

4.3 Adjustable Affinity-Preserving Random Walk for Summarization

Affinity-preserving random walk preserves the affinity relations between sentences and gives high ranking scores to the salient sentences. However, the diversity constraint of summarization has not been taken into account. The surfer of affinity-preserving random walk has no knowledge about what a diverse summary should be. If we just take redundancy removing as the post-processing separate step to improve diversity, sentences that highly overlap with other summary sentences may be chosen and sentences that include information about different topics may be submerged. This phenomenon can be explained by the theorem as follows.

Theorem 4.1. *Suppose $\bar{\boldsymbol{x}}$ is the quasi-stationary distribution of affinity-preserving random walk as defined in Sec.4.2 and $\boldsymbol{x}$ is the solution of a continuous quadratic optimization problem $argmax(\boldsymbol{x}^T \boldsymbol{A}\, \boldsymbol{x})$ s.t. $\boldsymbol{x} \in [0,1]^n$, $\|\boldsymbol{x}\|_2 = 1$ and $\boldsymbol{A}$ has definition in Sec.4.2. The following equation holds*

$$\bar{\boldsymbol{x}} = \boldsymbol{x} / \|\boldsymbol{x}\|_1 \tag{9}$$

when $\mu = 1$.

Proof. In mathematics, for a given symmetric real matrix $\mathbf{A}$ (when $\mu = 1$) and nonzero real vector $\mathbf{x}$, the Rayleigh quotient $R(\mathbf{A}, \mathbf{x})$ is defined as

$$R(\mathbf{A}, \mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

and it reaches its maximum value when $\mathbf{x}$ is the principal eigenvector of $\mathbf{A}$. If $\|\mathbf{x}\|_2 = 1$, $R(\mathbf{A}, \mathbf{x})$ is equivalent to $\mathbf{x}^T \mathbf{A}\, \mathbf{x}$. So the solution $\mathbf{x}$ is the principal eigenvector of $\mathbf{A}$. From Section 4.2, $\bar{\mathbf{x}}$ is the normalized principal eigenvector of $\mathbf{A}$. $\bar{\mathbf{x}}$ and $\mathbf{x}$ have the relation in Eq.(9). The conclusion is made. □

From Theorem 4.1, affinity-preserving random walk tends to produce a stationary distribution in which the total sum of affinity between sentences (i.e. $\mathbf{x}^T \mathbf{A} \mathbf{x}$) is large if there is a subtle teleporting effect. It will lead to a summary consisting of many sentences overlapping with each other, which clearly violates the diversity constraint of summarization. Good candidate sentences may not have high affinity with other sentences and are likely to be submerged by affinity-preserving random walk. Conversely, some bad candidate sentences could have high affinity with others and will

be highlighted by the random walk process. An extreme example is that a cluster of sentences will all get high ranking scores if they are very similar to each other. However, only one sentence in this cluster should be included in the summary and the others should be suppressed. The random surfer is caught in a trap of larger sentence cluster, which operates against the exploration of good candidates in smaller cluster.

We introduce adjustable affinity-preserving random walk to enforce the diversity constraint of summarization. In the original affinity-preserving random walk, sentence transition matrix $\mathbf{A}$ is fixed and set as $\mu\mathbf{W}/\|\mathbf{W}\|_1+(1-\mu)\mathbf{y}\cdot\mathbf{e}^T$. Edge (s_i, s_0) (if $C(s_i) \neq C_{max}$) always exists and has an invariant conductance $c(s_i, s_0)$, which means that the surfer at s_i walks into s_0 in the same manner for the entire random walk process. The random surfer makes her decision only based on invariant $\mathbf{A}$ to select salient sentences and form the summary. To equip the surfer with knowledge about what a diverse summary should be, we propose to adjust sentence transition matrix $\mathbf{A}$ in each iteration of random walk. The key point is that good candidate sentences should be normalized *locally* while bad ones should be normalized *globally* in the transformation from affinity matrix $\mathbf{W}$ to transition matrix $\mathbf{A}$.

A "virtual" summary V is produced based on $\mathbf{x}$ in each iteration of affinity-preserving random walk. "Virtual" here means that V is a summary based on transient distribution $\mathbf{x}$, which differs from the final summary based on quasi-stationary distribution $\bar{\mathbf{x}}$. The method of diversity penalty imposition (Wan et al., 2007) is used to produce V, which is denoted by the *producingSummary* function in Algorithm 4.1. It is a simple greedy algorithm to select sentences that are both salient and diverse, which often plays a role of greedy post-processing step to produce the final summary. Rather, we use it to produce virtual summary V that satisfies both the saliency and diversity constraints based on a specific iteration. V is an indicator for the diversity constraint of summarization.

The sentence transition matrix $\mathbf{A}$ in the iteration $(K+1)$ is then constructed with help of the virtual summary V in the iteration K. Here, different normalization methods are used to transform $\mathbf{W}$ into $\mathbf{A}$. If V includes the sentence s_i, elements in the corresponding i-th row of $\mathbf{W}$ will be normalized by the sum of the row (i.e. $C(s_i)$). Otherwise, elements will be normalized by the maximum sum of row elements in $\mathbf{W}$ (i.e. C_{max}). In this way, "local normalization" is adopted for the sentences in V while "global normalization" is adopted for the sentences not in V. We differentiate the normalization methods to lead the surfer of affinity-preserving random walk to explore more in the neighborhood of the sentences in V rather than end in the absorbing vertex s_0. As a result, the sentences that satisfy the saliency and diversity constraints will be highlighted even though they are in a small sentence cluster. We characterize different normalizations in the diagonal matrix $\mathbf{D}$. $\mathbf{D}_{ii}$ is $C(s_i)$ if $s_i \in V$ and $\mathbf{D}_{ii}$ is C_{max} if $s_i \notin V$,which is denoted by the *adjustingNormalization* function in Algorithm 4.1. $\mathbf{D}$ in the current iteration is here dependent on $\mathbf{x}$ in the previous iteration. We will have different sentence augmented graphs G^A in each iteration. Figure 4.2 shows an example of $G^A(K)$ and $G^A(K+1)$ in the respective iterations K and $(K+1)$. The probability distribution in the adjustable affinity-preserving random walk is updated as follows.

Algorithm 4.1: Adjustable Affinity-Preserving Random Walk for Multi-Document Summarization

Input: The sentence affinity matrix, $\mathbf{W}$; The starting and maximum number of iteration, B and M; The teleport vector, $\mathbf{y}$; The damping factor, μ;

Output: The multi-document summary, V;

```
A ← μW/C_max + (1 − μ)y · e^T
Initialize the starting distribution x as uniform
for i ← 1, 2, ..., M do
    if i > B then
        V ← producingSummary(x)
        D ← adjustingNormalization(V)
        A ← μ(D^{-1}W)^T + (1 − μ)y · e^T
    x̄ ← Ax
    x̄ ← x̄/‖x̄‖_1
    if ‖x̄ − x‖_1 < ε then
        break
    x ← x̄
V ← producingSummary(x)
Return V
```

$$\mathbf{x} = \frac{\mu(\mathbf{D}^{-1}\mathbf{W})^T\mathbf{x} + (1-\mu)\mathbf{y}}{\|\mu(\mathbf{D}^{-1}\mathbf{W})^T\mathbf{x} + (1-\mu)\mathbf{y}\|_1} \tag{10}$$

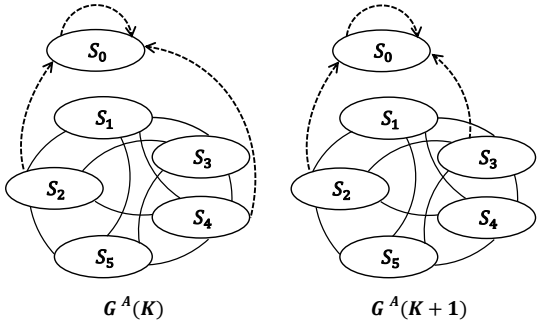

Figure 4.2: Sentence augmented graphs for summarization in two successive iterations. $G^A(K)$: augmented graph in the iteration K. Virtual summary $V = \{s_1, s_3, s_5\}$, which is constructed from $\mathbf{x}$ in the iteration $(K-1)$ by *producingSummary*. $\mathbf{D} = diag([C(s_1), C_{max}, C(s_3), C_{max}, C(s_5)])$. $G^A(K+1)$: augmented graph in the iteration $(K+1)$. $V = \{s_1, s_4, s_5\}$, which is constructed from $\mathbf{x}$ in the iteration K by *producingSummary*. $\mathbf{D} = diag([C(s_1), C_{max}, C_{max}, C(s_4), C(s_5)])$. In both cases, C_{max} equals to $C(s_1)$ indicating that sentence s_1 has the maximum conductance.

This is an adjustable Markov chain for which the transition matrix $\mathbf{A}$ is $\mu(\mathbf{D}^{-1}\mathbf{W})^T + (1-\mu)\mathbf{y}\cdot\mathbf{e}^T$. In this setting, $\mathbf{A}$ is dependent on the transient distribution $\mathbf{x}$ in the previous iteration, which differs from the invariant transition matrix in Eq.(8). As the diversity constraint is embedded in $\mathbf{A}$, subsequent random walks move to the solution that induces a better summary. The algorithm of the adjustable affinity-preserving random walk for multi-document summarization is demonstrated in Algorithm 4.1.

The parameter B in Algorithm 4.1 is used to produce a transient distribution which is reliable enough to adjust the transition matrix. To get the final multi-document summary, we use the same *producingSummary* function.

5 Experiments

5.1 Data Sets

Generic and topic-focused multi-document summarization have been the main tasks in DUC[1]. Task 2 of DUC 2004 is a generic summarization task and task 3 of DUC 2003 is a topic-focused summarization task. Both tasks are used for performance evaluation of our method. In the experiments, task 2 of DUC 2003 is used for the parameter tuning of our method. We preprocess the document data sets by removing stopwords from each sentence and stemming the remaining words using the Porter's stemmer[2]. Also, the sentences containing the said clause (if a said, says, told, tells word and quotation marks appear simultaneously) are filtered out.

For evaluation, four reference summaries generated by human judges for each document cluster are provided by DUC as the ground truth. A brief summary over the evaluation datasets is shown in Table 5.1. According to (Hong et al., 2014), we adjust the length limit of summary in DUC 2004 from 665 bytes to 100 words as it provides the same setting for system evaluations.

	DUC 2003	DUC 2003	DUC 2004
Task	Task 2	Task 3	Task 2
Type	Generic	Topic-focused	Generic
Cluster numbers	30	30	50
Data source	TDT	TREC	TDT
Summary length	100 words	100 words	100 words

Table 5.1: Summary of data sets used in our experiments.

5.2 Evaluation Metric

We use the ROUGE-1.5.5 (Lin and Hovy, 2003) toolkit for evaluation, which has been officially adopted by DUC for automatic summarization evaluation. The toolkit measures summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE-N is an n-gram based measure and the ROUGE-N recall is computed as follows

$$ROUGE\text{-}N_R = \frac{\sum_{S\in\{RefSum\}}\sum_{n\text{-}gram\in S} Count_{match}(n\text{-}gram)}{\sum_{S\in\{RefSum\}}\sum_{n\text{-}gram\in S} Count(n\text{-}gram)} \quad (11)$$

where n stands for the length of the n-gram, and $Count_{match}(n\text{-}gram)$ is the maximum number of n-grams co-occurring in the candidate summary and the set of reference summaries. $Count(n\text{-}gram)$ is the number of n-grams in the reference summaries.

We conduct our ROUGE experiments following the recommended standard in (Owczarzak et al.,

[1] http://www-nlpir.nist.gov/projects/duc/intro.html

[2] https://tartarus.org/martin/PorterStemmer/

2012; Hong et al., 2014)[3]. We compute ROUGE-2 recall with stemming and stopwords not removed, which provides the best agreement with manual evaluations. We also compute ROUGE-1 recall which has the highest recall of ability to identify the better summary in a pair, and ROUGE-4 recall which has the highest precision of ability to identify the better summary in a pair (Owczarzak et al., 2012).

5.3 Experimental Results

In the experiments, the parameters of our method are set as follows: the decay factor λ is 2, the maximum number of iteration M is 100, the number of starting iteration B is 30, the damping factor μ is 0.85 and the minimum error ε is 1E-30.

System	R-1	R-2	R-4
Cont. LexPageRank*	35.95	7.47	0.82
FreqSum	35.30	8.11	1.00
CLASSY 04	37.62	8.96	1.51
CLASSY 11	37.22	9.20	1.48
GRASSHOPPER*	37.20	9.26	1.50
DivRank*	37.60	9.30	1.52
GCD*	38.68	9.31	1.45
Submodular	39.18	9.35	1.39
APRW*	38.10	9.39	1.35
DPP	39.79	9.62	1.57
ICSISumm	38.41	9.78	1.73
AAPRW*	38.92	10.06	1.61
WFS-NMF	39.24	10.94	1.65

Table 5.2: System comparisons on task 2 of DUC 2004 (%). *: Graph-based ranking methods.

Table 5.2 shows the performance of our method and other eleven well-known systems on task 2 of DUC 2004 according to ROUGE-1,2,4 recall, sorted by ROUGE-2 recall in the ascending order. Some of the results are from (Hong et al., 2014). *Cont. LexPageRank* (Erkan and Radev, 2004) is a graph-based ranking method and a representative of traditional random walk approach. Here we employ the continuous version of LexPageRank. *FreqSum* (Nenkova et al., 2006) is a simple approach to approximate the importance of words with their probability in the input and then select sentences with high average word probability. *CLASSY 04* (Conroy et al., 2004) was the participant of the official DUC 2004 evaluation with the best evaluation score. It employs a Hidden Markov Model using topic signature feature and requires a linguistic preprocessing component. *CLASSY 11* (Conroy et al., 2011) is the successor of *CLASSY 04* and selects the non-redundant sentences using the non-negative matrix factorization algorithm. In the *Submodular* system (Lin and Bilmes, 2011), multi-document summarization is formulated as a submodular set function maximization problem. *DPP* (Lin and Bilmes, 2011) combines a sentence saliency model with a global diversity model encouraging non-overlapping information. *ICSISumm* (Gillick and Favre, 2009) aims at finding the globally optimal summary by formulating the summarization task in Integer Linear Programming. *WFS-NMF* (Wang et al., 2010) extends the non-negative matrix factorization algorithm and provides a good framework for weighting different terms and documents. *GRASSHOPPER*, *DivRank* and *GCD* are the three graph-based ranking models mentioned in Section 2. *APRW* and *AAPRW* are our methods. *APRW* is the method of affinity-preserving random walk described in Section 4.2 and *AAPRW* is the method of adjustable affinity-preserving random walk described in Section 4.3.

System	R-1	R-2	R-4
S17	31.81	4.98	0.47
S13	31.99	5.83	0.73
S16	35.00	7.31	1.04
Manifold Ranking	37.33	7.68	1.26
APRW	35.72	7.72	1.34
AAPRW	36.36	8.21	1.40

Table 5.3: System comparisons on task 3 of DUC 2003 (%).

Table 5.3 shows the evaluation results on task 3 of DUC 2003 according to ROUGE-1,2,4 recall, sorted also by ROUGE-2 recall in the ascending order. *S13*, *S16* and *S17* are the system IDs of the top performing systems in the official DUC 2003 evaluation, whose details are described in DUC publications (Zhou and Hovy, 2003; Chali et al., 2003). *Manifold Ranking* is the method proposed in (Wan et al., 2007) to make use of both the relationships among all sentences in the documents and the relationships between the given topic de-

[3]ROUGE-1.5.5 with the parameters: -n 4 -m -a -l 100 -x -c 95 -r 1000 -f A -p 0.5 -t 0

scription and the sentences. *APRW* and *AAPRW* are our methods.

From Tables 5.2 and 5.3, our method has the best ROUGE-2 score among all graph-based ranking methods for generic multi-document summarization, and it also has the best ROUGE-2 score for topic-focused multi-document summarization. *AAPRW* has the ROUGE-2 score 10.06% on DUC 2004 task 2, which is 0.28% higher than the best system *ICSISumm* reported by (Hong et al., 2014) and 1.1% higher than the official best system *CLASSY 04*. *WFS-NMF* has the overall best score on DUC 2004 task 2 due to the sentence feature selection and the weights on the document side, which is reported by (Wang et al., 2010; Alguliev et al., 2013). *AAPRW* has the ROUGE-2 score 8.21% on DUC 2003 task 3, which is 0.53% higher than *Manifold Ranking* and 0.9% higher than the official best system *S16*. In DUC 2004 *AAPRW* has 0.67% more ROUGE-2 score than *APRW* and the gap is 0.49% in DUC 2003, which proves the effectiveness of the adjustable transition matrix in the random walk process. It is worth mentioning that our method has the best ROUGE-4 score on the DUC 2003 topic-focused summarization task.

We conducted the two-sided Wilcoxon signed-rank tests between each pair of *AAPRW* and other methods. For the generic summarization in DUC 2004, our method provides a significant improvement over the official best system *CLASSY 04* on ROUGE-2 (with p-value lower than 0.05). For the query-focused summarization in DUC 2003, our method also provides a significant improvement over *S17*, *S13* and *S16* on ROUGE-2.

In order to further investigate the influences of the parameter in our proposed method, the damping factor μ is varied from 0 to 1. Figures 5.1 and 5.2 show the ROUGE-1 and ROUGE-2 recall curves of our method on the two data sets, respectively. We can see from the figures that the damping factor has an effect on the performance of multi-document summarization.

6 Conclusion and Future Work

In this paper we propose the adjustable affinity-preserving random walk for generic and topic-focused multi-document summarization, which deals with the saliency and diversity goals in a unified framework. Experiments demonstrate the effectiveness of our method.

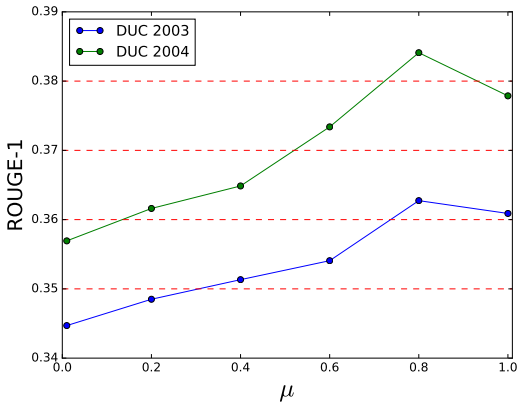

Figure 5.1: ROUGE-1 recall scores vs. μ of our method.

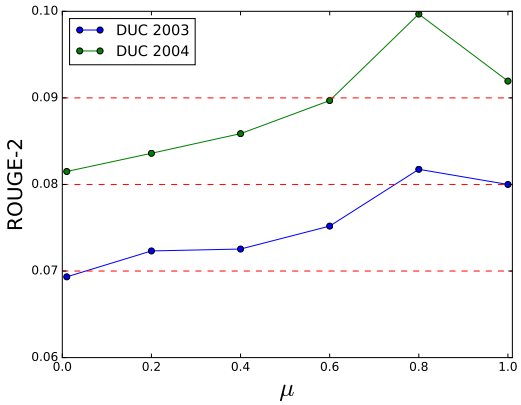

Figure 5.2: ROUGE-2 recall scores vs. μ of our method.

In the future work, we will focus on the self transition of adjustable affinity preserving random walk, which could be used to remove the redundancy between summary sentences.

Acknowledgements

We would like to thank our three anonymous reviewers for their helpful advice on various aspects of this work. This research is supported by National Key Basic Research Program of China (No.2014CB340504) and National Natural Science Foundation of China (No.61375074,61273318). The contact authors for this paper are Zhifang Sui and Baobao Chang.

References

Rasim M Alguliev, Ramiz M Aliguliyev, and Nijat R Isazade. 2013. Multiple documents summarization

based on evolutionary optimization algorithm. *Expert Systems with Applications*, 40(5):1675–1689.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Yllias Chali, Maheedhar Kolla, Nanak Singh, and Zhenshuan Zhang. 2003. The university of lethbridge text summarizer at duc 2003. In *the Proceedings of the HLT/NAACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2003)*.

Minsu Cho, Jungmin Lee, and Kyoung Lee. 2010. Reweighted random walks for graph matching. *Computer Vision–ECCV 2010*, pages 492–505.

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)*.

John M Conroy, Judith D Schlesinger, Jeff Kubina, Peter A Rankel, and Dianne P O'Leary. 2011. Classy 2011 at tac: Guided and multi-lingual summaries and evaluation metrics. *TAC*, 11:1–8.

John N Darroch and Eugene Seneta. 1965. On quasi-stationary distributions in absorbing discrete-time finite markov chains. *Journal of Applied Probability*, 2(1):88–100.

Avinava Dubey, Soumen Chakrabarti, and Chiranjib Bhattacharyya. 2011. Diversity in ranking via resistive graph centers. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 78–86. ACM.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics.

Taher H Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM.

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*, pages 1608–1616.

Glen Jeh and Jennifer Widom. 2003. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279. ACM.

Chin-Yew Lin and Eduard Hovy. 2003. 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-Volume 1*, pages 71–78. Association for Computational Linguistics.

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.

Qiaozhu Mei, Jian Guo, and Dragomir Radev. 2010. 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*, pages 1009–1018. Acm.

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. 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*, pages 573–580. ACM.

Karolina Owczarzak, John M Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9. Association for Computational Linguistics.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Eugene Seneta. 2006. *Non-negative matrices and Markov chains*. Springer Science & Business Media.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI*, volume 7, pages 2903–2908.

Dingding Wang, Tao Li, and Chris Ding. 2010. Weighted feature subset non-negative matrix factorization and its applications to document understanding. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 541–550. IEEE.

Chuncheng Xiang, Baobao Chang, and Zhifang Sui. 2015. An ontology matching approach based on affinity-preserving random walks. In *IJCAI*, pages 1471–1478.

Dengyong Zhou, Jason Weston, Arthur Gretton, Olivier Bousquet, and Bernhard Schölkopf. 2003. Ranking on data manifolds. In *NIPS*, volume 3.

Liang Zhou and Eduard Hovy. 2003. Headline summarization at isi. In *Document Understanding Conference (DUC-2003), Edmonton, Alberta, Canada*.

Xiaojin Zhu, Andrew B Goldberg, Jurgen Van Gael, and David Andrzejewski. 2007. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL*, pages 97–104.