GSPSummary: A Graph-Based Sub-topic Partition Algorithm for Summarization

Jin Zhang, Xueqi Cheng, and Hongbo Xu

Institute of Computing Technology, Chinese Academy of Sciences,
Beijing, P.R. China
zhangjin@software.ict.ac.cn,
{cxq,hbxu}@ict.ac.cn

Abstract. Multi-document summarization (MDS) is a challenging research topic in natural language processing. In order to obtain an effective summary, this paper presents a novel extractive approach based on graph-based sub-topic partition algorithm (GSPSummary). In particular, a sub-topic model based on graph representation is presented with emphasis on the implicit logic structure of the topic covered in the document collection. Then, a new framework of MDS with sub-topic partition is proposed. Furthermore, a novel scalable ranking criterion is adopted, in which both word based features and global features are integrated together. Experimental results on DUC2005 show that the proposed approach can significantly outperform existing approaches of the top performing systems in DUC tasks.

Keywords: Multi-document Summarization, Sub-topic, Graph Representation.

1 Introduction

With the rapid increasing of online information and fast development of science and technology, a lot of research efforts have been made on web mining, text mining, information extraction, and information retrieval (IR). However, the conventional IR technologies are becoming more and more insufficient for obtaining useful information effectively. Which makes how to summarize documents with all kinds of information increasingly urgent. Therefore, MDS - capable of summarizing either complete documents sets, or a series of documents in the context of previously ones - is likely to be essential in such situations. The goal of text summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's application needs [3]. If the summarization system can make an effective summary, which can be a substitute of the original documents, the retrieval effectiveness or efficiency can be improved and the user can save the reading time.

Usually, the topic of a document collection is composed of some aspects of information, each aspect is named sub-topic of the document collection. In order to

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 321–334, 2008.

model the sub-topics, many cluster-based approaches have been proposed. These approaches employ a clustering method to model the logic structure of the topic based on the structure of the topic covered in the document collection in the first, follows by a sentence selection method in a a specified cluster. However, the implicit logic structure of the topic covered in the document collection is not only represented by the explicit distribution of features (statistical features in usual), but also represented by the implicit distribution of features (centrality, etc).

In this paper, we argue that information selection in a MDS can be based on the implicit logic structure of the topic covered in the document collection. Using the relationship information with graph representation, we investigate the use of sub-topics as a model of the document collection for the purpose of producing summaries. Furthermore, unlike the two-step cluster-based approaches, we aim to obtain an approach can select important information when modeling sub-topics.

It would be worthwhile to highlight several aspects of our proposed algorithm here:

1. Presenting a new framework of MDS with sub-topic model, according to the implicit logic structure of the topic covered in the document collection.
2. Proposing a scalable criterion to rank the salience of sentences, in which both the word based and global features are modeled explicitly and effectively.
3. Proposing a novel MDS algorithm to determine the sub-topics in global space of a document collection.

The rest of this paper is organized as follows. Section 2 relates a review of the previous work. In section 3, we present the proposed graph-based summarization approach using sub-topic partition. The experimental methodologies and results are reported in section 4 and 5, followed by the conclusion and future work in section 6.

2 Related Work

Generating an effective summary requires the summarizer to select, evaluate, order and aggregate items of information according to their relevance to a particular subject or purpose. These tasks can either be approximated by IR techniques or done in great depth with full natural language processing (NLP). Most previous work in summarization has attempted to deal with the issues by focusing more on a related, but simple problem. Most of the work in sentence extraction applied statistical techniques (frequency analysis, variance analysis, etc.) to linguistic units such as tokens, names, anaphora, etc. (e.g., [9]). Other approaches include the utility of discourse structure [10], the combination of information extraction and language generation [1], and using machine learning to find patterns in text [6][7].

Several researchers have extended various aspects of the single document summarization approach to look at MDS [12][13]. These include comparing templates filled in by extracting information - using specialized, domain specific knowledge sources - from the document, and then generating natural language summaries

from the templates, comparing named-entities - extracted using specialized lists - between documents and selecting the most relevant section, finding co-reference chains in the document collection to identify common sections of interest, or building activation networks of related lexical items (identity mappings, synonyms, hypernyms, etc.) to extract text spans from the document collection [13].

Many of recent researches put emphasis on the comprehensiveness while keeping readability of summaries or maximizing the coverage and the anti-redundancy to keep the comprehensiveness and readability to some extent. For example, Radev et al. [14] proposed a method that classifies given documents into some clusters and made one sub-summary for each cluster, then placed them in an order.

Carbnell [1] proposed the Maximal Marginal Relevance (MMR) criterion for combining query relevance with information novelty in the context of text retrieval and summarization. Goldstein et al. [11] proposed a method called MMR-MD (Maximal Marginal Relevance - Multi-Document), which collects passages related to the query from newspaper articles retrieved by an IR system and arranged them into one summary.

As first proposed in [17], the central to the MDS approach has been gained a lot of interest. In 2005, Harabagiu et al. [15] proposed a topic themes method that a MDS can be based on the structure of the topic covering in the document collection.

3 Graph-Based Sub-topic Partition Algorithm

Although the document collection used to generate a summary may be relevant to the same general topic, they do not necessarily include the same information. Extracting all similar sentences would produce a verbose and repetitive summary, while extracting some similar sentences could produce a summary biased towards some sources, as it was noted in [8]. However, the graph-based extractive summarization algorithm succeeds in identifying the most important sentences in a document collection based on information exclusively drawn from the collection itself. In this section, we propose a graph-based algorithm - GSPSummary - to obtain the important sub-topics. GSPSummary starts from the assumption that capturing sub-topic structure of document collection is essential for summarization. It firstly creates a graph representation of the document collection, then selects the salient (or more central) sentences with GSPRank and obtains the most important sub-topics in global graph space iteratively, finally forms the summary supported by the salient sentences of different sub-topics. We will give the definition of graph-based sub-topic representation, the GSPRank criterion, and the GSPSummary algorithm in more details in the subsections below.

3.1 Problem Formalization

Let $G = (V, E)$ be an undirected graph with the set of nodes V and set of edges E, where E is a subset of $V \times V$. Then a graph G of a related document collection can be represented by the set of sentences V, and the similarities to each other

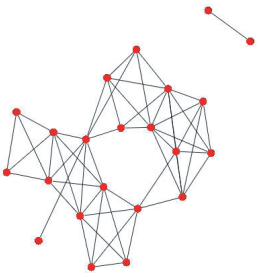

Fig. 1. Sentences distance graph for a small document collection with 22 sentences

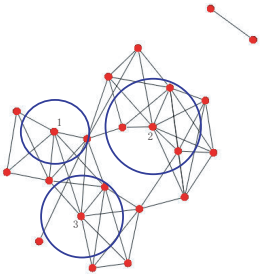

Fig. 2. Sentences distance graph with three sub-topics marked by three circles around node 1, 2, 3 respectively

E. Figure 1 is a sentences distance undirected graph representing a document collection with 22 sentences (the nodes in Figure 1), the edge stands for the distance $dist(u_i, u_j)$ of every pair of sentences u_i, u_j in the document collection. A threshold is defined to eliminate the edges whose distance is higher, since we are interested in significant similarities. This reduction in the graph also provides us with computational savings. To define distance, we use the bag-of-words model to represent each sentence as an N-dimension vector, where N is the number of all possible words. Formally, the distance between two sentences is then defined by the following equation:

$$dist(u_i, u_j) = 1.0 - \frac{\boldsymbol{u_i}.\boldsymbol{u_j}}{|\boldsymbol{u_i}||\boldsymbol{u_j}|} \tag{1}$$

where $\frac{\boldsymbol{u_i}.\boldsymbol{u_j}}{|\boldsymbol{u_i}||\boldsymbol{u_j}|}$ is the traditional cosine similarity between two sentences u_i and u_j using $tfidf$ weighting method. The effectiveness and robustness of the above measure has been proven in IR and NLP.

Suppose there are three sub-topics in this document collection (the three circles marked in Figure 2), and node 1, node 2 and node 3 are the salient sentences of the three sub-topics.

In order to generate the sub-topics set, we need a ranking method which can generate sub-topics using combined criterion of relevance to the given topic and its centrality. Here, relevance and centrality are not two conflicting concepts while belong to two different dimension. Relevance is the relationship of the topic with retrieved sentences set, centrality is based on the relationship among its similar sentences.

We used the following notation throughout this paper:

- $subtopic(S|T)$: the sub-topic coverage of the document collection corresponding to a topic. Given a certain topic T, we may substitute the notation by $subtopic(S)$ which is clear under certain context.
- u: a single node in the document collection, in usual, is a sentence.
- $p_c(u)$: the salient node u of a certain sub-topic S.
- $neighbor(u)$: the nodes near to the salient node u, also these nodes belong to the same sub-topic S.

More precisely, we define the sub-topics $subtopic(S|T)$ as:

$$subtopic(S|T) = p_c(u) + neighbor(u) \tag{2}$$

Then the summarization problem can be formulated as a graph partition problem:

Given a sentences distance graph G of a document collection of a certain topic T, composed of a set of N nodes $U = u_1, u_2, ..., u_N$, and a length l, partition K sub-graphs of nodes $S_i \subseteq U$ as K sub-topics such that: (1) each sub-graph has a salient node u and its neighborhood $neighbour(u)$ and (2) using u as a representative node of S_i; (3) sum of the length of all the K salient nodes should not be more than l.

The key for our task here is to find the appropriate salient node $p_c(u)$ and its neighborhood $neighbour(u)$ in G.

3.2 GSPRank Criterion

Many existing approaches explore the most important units (clauses/ sentences/ paragraphs) in texts [4] with statistics scoring methods and other higher semantic/syntactic structure such as rhetorical analysis, lexical chains, co-reference chains [6]. Unfortunately, these methods are still hard to obtain the really important units, for the important units are not only decided by the statistical features, but also decided by the semantic features and other fields' features. To explore the most important units or assess the salient nodes in graph, we propose a new sentence ranking criterion - GSPRank - served as basis for our GSPSummary method. This criterion has inspired by the ideas in information retrieval and feature selection. Since the summarization is controlled by choosing the central sentences, which we call "salient sentences", it is in principle possible for the salient sentences to be scored according to the word based features - the statistical features or semantic features according to words or phrases - and the global features.

$$g(u) = f_1(u) \cdot f_2(u) \tag{3}$$

where $g(u)$ is the salience score of sentence u, $f_1(u)$ is the score of word based features, and $f_2(u)$ is the score of global features. We can use the product of the two classes of features to assess the salience of sentence u, for they belong to two different feature spaces.

Word Based Features Metrics. Among the word based features proposed previously, the $tfidf$ score of word is the most widely used approach. In the course of our investigation, the word based features can be presented with a linear combination as the following:

$$f_1(u) = \sum_{i=1}^{m} \lambda_i f_{wi}(u) \tag{4}$$

s.t.

$$\lambda_i \geq 0$$

where $f_{wi}(u)$ is a single word based feature, and the parameter λ_i is the factor to adjust different word based features. Normally, we can express these m word based features with a linear combination. In practice, we use the following word based features:

$$f_1(u) = \lambda_1 f_{w1}(u) + \lambda_2 f_{w2}(u, T) + \lambda_3 f_{w3}(D(u), T) \tag{5}$$

That is, $f_{w1}(u)$ is the centrality score of sentence u, $f_{w2}(u, T)$ is the relevance score between sentence u and the document collection's topic T, and $f_{w3}(D(u), T)$ is the relevance score between the document $D(u)$ where sentence u located and the topic T.

To compute the overall centrality $f_{w1}(u)$ of a sentence given to other sentences, Radev et al. [5] proposed a LexRank approach based on the concept of graph-based centrality. The LexRank value of a sentence gives the limiting probability that such a random walk will visit that sentence in the long run. By LexRankthe score of sentence u can be computed as:

$$f_{w1}(u) = l(u) = \frac{d}{N} + (1 - d) \sum_{v \in adj[u]} \frac{w(u, v)}{\sum_{z \in adj[v]} w(z, v)} l(v) \tag{6}$$

where $l(u)$ is the LexRank value of sentence u, N is the total number of nodes in the graph, d is a damping factor for the convergence of method, and $w(u, v)$ is the weight of the link form sentence u to sentence v. Equation 6 can be written in the matrix form as

$$l = [d\mathbf{U} + (1 - d)\mathbf{B}]^T l \tag{7}$$

where U is a square matrix with all elements being equal to $1/N$. The transition kernel $[d\mathbf{U} + (1 - d)\mathbf{B}]$ of the resulting Markov chain is a mixture of two kernels U and B.

Global Features Metrics. Here, global features mainly consider the length, the position, the text pattern of a sentence, and so on. A simple fact is that short sentences cannot carry enough information corresponding to the topic. Thus, they are not appropriate candidates of summary sentences. And due to the constraint of summary length, too long sentences are not appropriate, either. There are some patterns which are unsuitable for appearing in the summary. The sentences which have these patterns will be discounted for summary sentence. Normally, we can consider the global features are independent, then the global features can illustrated in a form of conditional probability in Equation 9.

$$f_2(u) = P(F_g|u) = \prod_{i=1}^{k} p(f_{gi}|u) \tag{8}$$

where F_g are the global features, and $P(F_g|u)$ is the probability of sentence u in global features space, and $P(F_g|u)$ equals to the product of k global features. In

our work, global feature space involves three salient phases: the sentence length, sentence position, and sentence pattern.

$$\begin{cases} p(f_{g1}|u) = p(length|u) \\ p(f_{g2}|u) = p(position|u) \\ p(f_{g3}|u) = p(pattern|u) \end{cases} \tag{9}$$

That is, $p(f_{g1}|u)$ is the probability that the observation of length feature was generated by the training data set, $p(f_{g2}|u)$ is the probability of position feature of u, and $p(f_{g3}|u)$ is the probability of sentence pattern of u. What's more, the global features can be exploited from a supervised way by using a machine learning method based on a training corpus of documents, such as HMM.

GSPRank. As mentioned above, we can obtain the new sentence ranking criterion - GSPRank - combining with word based features and global features. From the Equation 3, 5, 6, and 9, we can induce

$$GSPRank(u) = g(u) = j(u) \cdot l(u) \tag{10}$$

s.t.

$$j(u) = (1 + \lambda_1' \frac{f_{w2}(u,T)}{l(u)} + \lambda_2' \frac{f_{w3}(D(u),T)}{l(u)}) \prod_{i=1}^{k} p(f_{gi}|u)$$

where $g(u)$ is the salience score of sentence in Equation 3, $j(u)$ is a feed function for sentence u, and $l(u)$ is the centrality score of sentence u, which is same to $f_{w1}(u)$ noted in Equation 5. As the Equation 6 mentioned, $l(u)$ can be calculated as a Markov chain model. The convergence property of Markov chains provides a simple iterative algorithm, called Power Method[1], to compute the stationary distribution. The algorithm starts with a uniform distribution. At each iteration, the eigenvector is updated by multiplying with the transpose of the stochastic matrix. Since the Markov chain is irreducible and aperiodic, the algorithm is guaranteed to terminate.

Based on these, we can write Equation 3 in the matrix notation as Equation 11. Here, the salience scores of the sentences set U can be formulated with the product of a feed matrix J and a vector L as the following equation.

$$R = J \cdot L \tag{11}$$

where $\mathbf{R}$ is the vector of GSPRank scores of the sentences set U, J is the feed matrix corresponding to U, each diagonal element in J is a feed function for sentence u in Equation 12, and L is the centrality score vector of U, which can be calculate with the Power Method. Since the procedure of calculating L is iterative, the procedure of calculating R can also be presented as an iterative method with the Markov model.

[1] http://math.fullerton.edu/mathews/n2003/PowerMethod- Mod.html

$$\mathbf{J} = \begin{pmatrix} j(u_1) \ldots & 0 & \ldots & 0 \\ \ldots \quad \ldots & \ldots & \ldots & \ldots \\ 0 \quad \ldots & j(u_i) & \ldots & 0 \\ \ldots \quad \ldots & \ldots & \ldots & \ldots \\ 0 \quad \ldots & 0 & \ldots & j(u_n) \end{pmatrix} \tag{12}$$

3.3 GSPSummary Algorithm

In section 3.2, we proposed a novel ranking criterion - GSPRank - to assess salience of sentence. The GSPRank can be expressed as an iterative way. Equivalently, our procedure of sub-topic partition algorithm can be described iteratively. This way, a GSPSummary algorithm (in Algorithm 1) should include the following stages as the Fig.3 illustrates:

1. Partition a sub-topic: Generate a ranked list G of U with GSPRank, select the most salient node $p_c(u)$, then obtain $neighbor(u)$ with graph searching or graph partition algorithms.
2. Modify adjacency matrix for next partition: Reduce all the nodes of subgraph S from M (in Fig.3(3)), and generate the next salient node $p_c(u')$ and its neighborhood $neighbor(u')$ until the algorithm can be terminated.

A brief sketch of our GSPSummary algorithm by looking at the graphs in Fig.3 is to find the salient node $p_c(u)$ and its neighborhood $neighbor(u)$ in graph G of the document collection based on sentences relation. Suppose M is the adjacency matrix of G (in Fig.3(1)), M_0 is the initial matrix of G, and each circle is an element. As seen in Fig.3(2), we can use GSPRank to obtain the salient node $p_c(u)$ in the global space of M_0, then the neighborhood of $p_c(u)$ can

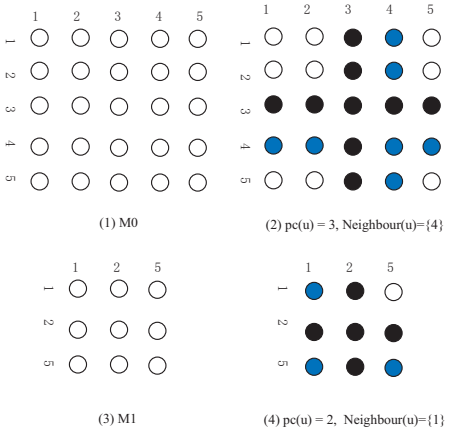

Fig. 3. The procedure of GSPSummary algorithm, (1) illustrates a adjacency matrix of graph, which has 5 nodes, and a circle means an element of the matrix

be generated with graph partition or searching algorithms - e.g, BFS - with a specified neighborhood threshold. At the first iteration, the salient node 3 and its neighborhood node 4 denote the first sub-topic S_1. After eliminated the elements corresponding to S_1, the matrix is be adjusted into a lower dimension one M_1 in Fig.3(3), follows by the next iteration to find the next sub-topic in Fig.3(4). In order to rank the scores of salient nodes, we used the GSPRank method (in Algorithm 2) in each iteration. ζ is the convergence factor for Power Method.

Input: A document collection D about the topic T
Output: An array of summary sentences S

```
1  repeat
2  InitGraphMatrix(&M,D);
3  ArrayRS;
4  i = GSPRankMethod(M, DistThre, ζ, RS);
5  ArrayNeighbours = NeighbourSearch(i, M, NeighbourThre);
6  iLen = LengthOfSentence(i);
7  if ((iLen + iSummaryLen) > SelectThre) then
8  |  break;
9  end
10 InsertIntoSelectedArray(S,i);
11 iSummaryLen+ = iLen;
12 UpdateRemainGraph(RS);
13 until (RS.size() >= MINGRAPHSIZE);
```

Algorithm 1. GSPSummary Algorithm

The following GSPRank Method (Algorithm 2) describes how to select a salient sentence for a given set of sentences with GSPRank. Note that the centrality score vector L is also computed as a side product of the algorithm, and ϵ is the distance threshold used to eliminate some high distance.

4 Experimental Setup

In order to evaluate our GSPSummary approach, we use the ROUGE[2] metrics on DUC2005 data sets for comparison. And the ROUGE score of the DUC2005 start-of-the-art systems came from Hoa's overview of DUC2005 in [2].

4.1 DUC Task Description

Every year, Document Understanding Conferences (DUC[3]) evaluates competing research group's summarization systems on a set of summarization tasks. In DUC2005, the task is to produce summaries of sets of documents in response to short topic statements that define what the summaries should address. The

[2] ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation, http://haydn.isi.edu/ROUGE/
[3] Document Understanding Conferences, http://duc.nist.gov/

Input: An array S of n sentences, distance threshold ϵ, a size N feed matrix J
Output: The ID of the salient sentence of S

```
Array DistMatrix[n][n];
Array Degree[n];
maxDist = -INFINITE;
for i ← 0 to n do
    for j ← 0 to n do
        DistMatrix[i][j]=dist(S[i],S[j]);
        if DistMatrix[i][j] < ε then DistMatrix[i][j]=1;
        Degree[i]++;
        else DistMatrix[i][j]=0;
    end
end
Normalization of matrix DistMatrix[i][j];
L = PowerMethod(DistMatrix);
R = J · L;
return the ID with maximal score from R;
```

Algorithm 2. GSPRank Method for obtaining salient sentence in global space

summaries are limited to 250 words in length. The DUC 2005 task was a complex question-focused summarization task that required summaries to piece together information from multiple documents to answer a question or set of questions as posed in a DUC topic. NIST[4] Assessors developed a total of 50 DUC topics to be used as test data. For each topic, the assessor selected 25-50 related documents from the Los Angeles Times and Financial Times of London and formulated a DUC topic statement, which was a request for information that could be answered using the selected documents. The topic statement could be in the form of a question or set of related questions and could include background information that the assessor thought would help clarify his/her information need. The assessor also indicated the granularity of the desired response for each DUC topic. That is, they indicated whether they wanted the answer to their question(s) to name specific events, people, places, etc., or whether they wanted a general, high-level answer. Only one value of granularity was given for each topic, since the goal was not to measure the effect of different granularity on system performance for a given topic, but to provide additional information about the user's preferences to both human and automatic summarization.

4.2 ROUGE

Automatic text summarization has drawn a lot of interest in the NLP and IR communities in the past years. Recently, a series of government-sponsored evaluation efforts in text summarization have taken place in both the United States and Japan. The most famous DUC evaluation is organized yearly to compare the summaries created by systems with those created by humans. Following

[4] National Institute of Standard and Technology, http://www.nist.gov/

the recent adoption of automatic evaluation techniques by the machine translation community, a similar set of evaluation metrics - known as ROUGE [16] - were introduced for both single and multi-document summarization. ROUGE includes four automatic evaluation methods that measure the similarity between summaries: ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S.

5 Experimental Results

5.1 Threshold Selection Results

In order to obtain an appropriate sub-topic, the threshold selection of neighborhood is also significant. The higher the threshold, the less the informative; while the lower the threshold, the higher redundancy. On the extreme point where we have a very high threshold or a very low threshold, the GSPSummary algorithm would be of no expected use. Fig.4 demonstrates the effect of threshold for GSPSummary on DUC2005 data set with ROUGE-2 and ROUGE-SU4 metrics. We have experimented with 13 different thresholds - from 0.09 to 0.81 with step 0.06. It is apparent in the figure that the threshold of 0.21 can produce the best summaries together. When the threshold is too lower, the ROUGE scores are decreased for no node in the neighborhood. Similarly, when the threshold is too higher, the ROUGE scores are rapidly decreased for too many nodes in the neighborhood. Therefore, the curves less than 0.08 and higher than 0.81 were not plotted in Fig.4.

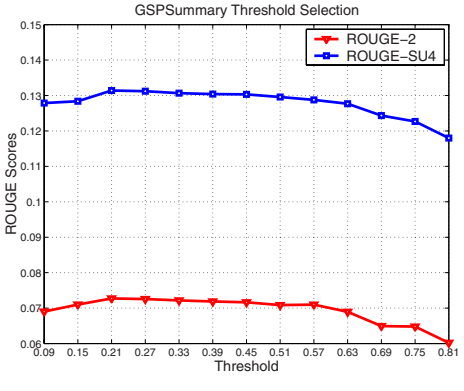

Fig. 4. The thresholds selection of GSPSummry, and the ROUGE scores change when the neighborhood threshold varies

5.2 Performance Comparison

We evaluated the performance of our system in terms of both comparison with LexRank and comparison with DUC2005 results. These comparisons indicate their applicability for real data, DUC2005.

Comparison of GSPSummary and LexRank. In the experiment, the proposed approach was compared with LexRank. With a unit matrix replaced the feed matrix J in Equation 12, our system will degenerate to a hierarchical LexRank system. In practice, we can introduce a hierarchical LexRank method to obtain the most important sub-topics. Unfortunately, for centrality only in LexRank, it is hard to measure the real salience of a sub-topic. Table 1 shows the results of two systems with two different ranks - LexRank and GSPRank. The ROUGE scores of Table 1 illustrates that our system with GSPRank can be quite more effective than the system with LexRank.

Table 1. Performance comparison between systems with LexRank and GSPRank. Columns 2-3 are the ROUGE-2 and ROUGE-SU4 scores of these two systems, and compared with LexRank, the results of GSPRank increase by 48% in ROUGE-2, 26% in ROUGE-SU4 respectively.

Approach	ROUGE-2	ROUGE-SU4
LexRank	0.04943	0.10541
GSPRank	0.07311	0.13231

Table 2. Evaluation results on DUC2005 dataset, IIITH-Sum, PolyU, NUS3 are the state-of-the-art systems competing in DUC2005, PolyU is the rank 2 system in ROUGE-2 and ROUGE-SU4, and NUS3 is the best system in ROUGE-2 and ROUGE-SU4. The last two rows are our two systems GSP-S1 and GSP-S2, and the ROUGE scores of DUC2005 can be highly improved with our GSP-Summary algorithm GSP-S2.

MDS Systems	ROUGE-2	ROUGE-SU4
Baseline	0.04160	0.08946
IIITH-Sum	0.06963	0.12525
PolyU	0.07174	0.12973
NUS3	0.07251	0.13163
GSP-S1	0.06964	0.12923
GSP-S2	**0.07311**	**0.13231**

Comparison of GSPSummary and DUC2005 Results. Table 2 shows the results of our two summarization systems GSP-S1, GSP-S2 on the data set of DUC2005 with ROUGE-2, ROUGE-L, and ROUGE-SU4. The baseline is the result provided by NIST, NUS3 is the best system in the two NIST official ROUGE scores: ROUGE-2 and ROUGE-SU4 recall. The GSP-S1 is used the GSPRank without consideration global features, while the GSP-S2 is used the GSPRank with the global features consideration. The score of our GSP-S1 in ROUGE-2 can obtain the 3rd rank, and the score of ROUGE-SU4 can obtain the 3rd place in DUC2005. Furthermore, comparing with IIITH-Sum - the third ranked system in ROUGE-2 and the 5th ranked system in ROUGE-SU4 - our GSP-S1 system has significant superiority in performance. The scores of our GSP-S2 can both obtain the 1st place in DUC2005. In comparison with the scores in GSP-S1, the ROUGE-2 score and the ROUGE-SU4 score increase 5.0%

and 2.4% respectively, which demonstrates the influence of the global features in the proposed approach. The results confirm that our graph-based sub-topic partition summarizer performs well as comparing to the state-of-the-art systems competing in DUC.

6 Conclusions and Future Work

Summarization is a product of electronic document explosion, and can be seen as the condensation of the document collection. As summary is concise, accurate and explicit, it became more and more important. In this paper, we present a new sub-topic representation model for MDS, and a new rank criterion is presented to obtain sub-topics. Furthermore, a new procedure and algorithm for generic and topic-oriented summarization is proposed. With the representation of graph, our algorithm can obtain the appropriate sub-topic with an iterative procedure in global space. We test our algorithm with DUC2005 data set, and the results suggest that our algorithm is effective in MDS.

The study has three main contributions: (1) we propose a new framework of MDS with sub-topic representation model, according to the logic structure of the topic covered in the document collection. (2) we propose a new ranking criterion GSPRank, in which both the word based and global features are modeled explicitly and effectively. (3) we present a new graph-based sub-topic partition algorithm GSPSummary for MDS.

As future work, we plan to explore in how to generate neighborhood with some other graph searching and partition algorithms. To some extent, our GSPSummary approach can be viewed as a simple version of hierarchical Markov model, with the scalable ranking criterion GSPRank, our algorithm can be further improved. Thus, in future work, we will study how to deal with such issues, and use fitful neighborhood searching or partition algorithms to model sub-topics.

Acknowledgments. The work is supported by the National Grand Fundamental Research 973 Program of China "Large-Scale Tex Content Computing" under Grand NO.2004CB318109 an Grand NO.2007CB311100.

References

1. Carbonell, J., Goldstein, J.: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In: Proceedings of SIGIR 1998 (August 1998)
2. Dang, H.T.: Overview of DUC 2005 (2005), `http://duc.nist.gov/pubs/2005papers/`
3. Mani, I.: Recent developments in text summarization. In: Proceedings of CIKM 2001, Atlanta, Georgia, USA, pp. 529–531 (2001)
4. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)
5. Erkan, G., Radev, D.R.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research (JAIR) (July 2004)

6. Barzilay, R., Elbadad, M.: Using Lexical Chains for Text Summarization. In: Proceedings of the ACL Intelligent Scalable Text Summarization Workshop, pp. 86–90 (1997)
7. Teufel, S., Moens, M.: Sentence Extraction as a Classification Task. In: Proceedings of the ACL Intelligent Scalable Text summarization Workshop (July 1997)
8. Barzilay, R., McKeown, K.R., Elhadad, M.: Information Fusion in the Context of Multi-Document Summarization. In: Proceedings of ACL 1999, June 16-20 (1999)
9. Mitra, M., Singhal, A., Buckley, C.: Automatic text summarization by paragraph extraction. In: ACL/EACL-1997 Workshop on Intelligent Scalable Text Summarization, July 1997, Madrid, Spain (1997)
10. Marcu, D.: From discourse structures to text summaries. In: Proceedings of the ACL 1997/EACL 1997 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain (1997)
11. Goldstein, J., Mittal, V.O., Carbonell, J.G., Callan, J.P.: Creating and Evaluating Multi-Document Sentence Extract Summaries. In: Proceedings of CIKM 2000 (2000)
12. Radev, D.R., McKeown, K.R.: Generating natural language summaries from multiple online sources. Computational Linguistics 24(3) (1998)
13. Mani, I., Bloedern, E.: Multi-document summarization by graph search and merging. In: Proceedings of AAAI-1997, pp. 622–628 (1997)
14. Radev, D.R., Jing, H., Budzikowska, M.: Summarization of multiple documents: clustering, sentence extraction, and evaluation. In: Proceedings, ANLP-NAACL Workshop on Automatic Summarization, April 2000, Seattle, WA (2000)
15. Harabagiu, S., Lacatusu, F.: Topic themes for multi-document summarization. In: Proceedings of SIGIR 2005 (2005)
16. Lin, C.-Y.: ROUGE: a Package for Automatic Evaluation of Summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain (2004)
17. Lin, C.-Y., Hovy, E.: The automated acquisition of topic signatures for text summarization. In: Proceedings of the 18th COLING Conference, Saarbrucken, Germany (2000)