Co-HITS-Ranking Based Query-Focused Multi-document Summarization

Po Hu[1,2], Donghong Ji[1], and Chong Teng[1]

[1] Computer School, Wuhan University,
430072 Wuhan, China
[2] Department of Computer Science, Huazhong Normal University,
430079 Wuhan, China
geminihupo@163.com, donghong_ji2000@yahoo.com.cn,
tchong616@126.com

Abstract. Graph-based ranking methods have been successfully applied to multi-document summarization by adopting various link analysis algorithms such as PageRank and HITS to incorporate diverse relationships into the process of sentence evaluation. Both the homogeneous relationships between sentences and the heterogeneous relationships between sentences and documents have been investigated in the past. However, for query-focused multi-document summarization, the other three kinds of relationships (i.e. the relationships between documents, the relationships between the given query and documents, and the sentence-to-document correlation strength) are seldom considered when computing the sentence's importance. In order to address the limitations, this study proposes a novel Co-HITS-Ranking based approach to query-biased summarization, which can fuse all of the above relationships, either homogeneous or heterogeneous, in a unified two-layer graph model with the assumption that significant sentences and significant documents can be self boosted and mutually boosted. In the model, the manifold-ranking algorithm is employed to assign the initial biased information richness scores for sentences and documents individually only based on the local recommendations between homogeneous objects. Then by adopting the Co-HITS-Ranking algorithm, the initial biased information richness scores of sentences and documents are naturally incorporated in a mutual reinforcement framework to co-rank heterogeneous objects jointly. The final score of each sentence can be obtained through an iteratively updating process. Experimental results on the DUC datasets demonstrate the good effectiveness of the proposed approach.

Keywords: Query-focused multi-document summarization, graph model, Co-HITS ranking.

1 Introduction

The growing availability of text in electronic formats has created an urgent need for the effective technologies that can help users cope with information overload problem. A revival of interest on multi-document summarization is spurred in the circumstances, because it can reduce information overload by synthesizing contents from a

P.-J. Cheng et al. (Eds.): AIRS 2010, LNCS 6458, pp. 121–130, 2010.

large collection of documents to produce a short text that can be read more quickly and digested more conveniently.

Multi-document summarization aims to provide a highly comprehensive overview of a document set. As a particular kind of multi-document summarization, query-focused summarization exhibits high practicability in many demand-driven applications and a great amount of research has been concerned. In [1], a query-biased summary was created by incorporating the content similarity between each sentence and the given query into a generic multi-document summarizer. In [2], a novel query expansion method was presented to improve the sentence ranking result. Wei et al. proposed a cluster-sensitive graph model and the corresponding iterative algorithm for query-focused multi-document summarization [3]. A variety of graph-based sentence ranking approaches have also been proposed recently [4,5,6,7].

However, these methods either make uniform use of inter-sentence recommendation to evaluate the sentence's significance without considering the influence of the document-level information or only divide the links between sentences into intra-document relationship and inter-document relationship without considering the sentence-to-document correlation strength. So in this study, a novel Co-HITS-Ranking based extractive approach is proposed to extend the existing work by naturally fusing three kinds of relationships between sentences and documents, either homogeneous or heterogeneous, in a unified two-layer graph model. Experiments have been performed on the DUC benchmark datasets, and the results demonstrate that the proposed Co-HITS-Ranking based approach can outperform both the lead baseline method and the sentence-based manifold-ranking method on the sentence affinity graph over three ROUGE metrics.

The rest of this paper is organized as follows: the proposed Co-HITS-Ranking based approach is presented in Section 2. The experiments and results are shown in Section 3. Section 4 presents our conclusion.

2 The Proposed Co-HITS-Ranking Based Approach

2.1 Overview

The proposed Co-HITS-Ranking based approach is intuitively based on the following assumptions:

Assumption 1: A sentence should be significant if it is heavily linked with the given query and other significant sentences. A document should be significant if it is heavily linked with the given query and other significant documents.

Assumption 2: A sentence should be significant if it has high correlation strength with the significant documents. A document should be significant if it has high correlation strength with the significant sentences.

Based on the assumptions, we develop a two-layer graph model to fuse three kinds of relationships (i.e. the homogeneous relationships between sentences or documents, and the heterogeneous relationships between sentences and documents), where the

significance of a sentence is not only determined by the significances of its related sentences and query, but also the significances of its closely related documents.

2.2 Two-Layer Graph Model

The two layer graph model is denoted as $G=< V_{SD}, E_{SS}, E_{DD}, E_{SD} >$, in which three sub-graphs are involved (i.e. G_{SS}, G_{DD}, and G_{SD}). $G_{SS} = (V_{SS}, E_{SS})$ is the undirected affinity graph of sentences. $V_{SS} = \{S_i \mid 1 \leqslant i \leqslant N\}$ is the set of sentences in a set, while $E_{SS} = \{(S_i, S_j) \mid S_i, S_j \in V_{SS}\}$ includes all possible links between pairs of sentences with the link weight $w(S_i, S_j)$ denoting the pair-wise content similarity between two sentences S_i and S_j . $G_{DD} = (V_{DD}, E_{DD})$ is the undirected affinity graph of documents, where $V_{DD} = \{D_j \mid 1 \leqslant j \leqslant M\}$ is the set of documents, and $E_{DD} = \{(D_i, D_j) \mid D_i, D_j \in V_{DD}\}$ includes all relationships between pairs of documents with the weight $w(D_i, D_j)$ representing the pair-wise similarity between document D_i and D_j. $G_{SD} = (V_{SD}, E_{SD})$ is the bipartite graph denoting the sentence-to-document correlations. $V_{SD} = V_{SS} \cup V_{DD}$. $E_{SD} = \{(S_i, D_j) \mid S_i \in V_{SS}, D_j \in V_{DD}\}$. The element's weight $w(S_i, D_j)$ of E_{SD} represents the correlation strength between the sentence S_i and document D_j . Figure 1 gives an illustration of the two layer graph G and its sub-graphs.

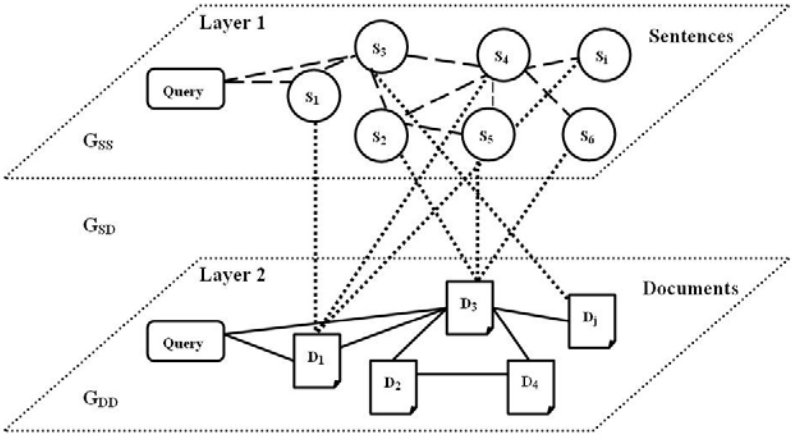

Fig. 1. Illustration of the two layer graph G and its sub-graphs

In Figure 1, the thin dotted lines linking the different kinds of objects from two layers demonstrate the correlation between sentences and documents. The upper layer expresses both the relationships among all the sentences and the relationships between the given query and the sentences. The relationships among all the documents and the relationships between the given query and the documents in the lower layer have been further investigated in this study. The given query q is treated as a pseudo-sentence when building the affinity graph G_{SS} of sentences, which can be processed in the same way as other sentences. Similarly, when building the affinity

graph G_{DD} of documents, the same query q is treated as a short pseudo-document, which can then be processed in the same way as other documents.

In the graph model, the link weight $w(S_i, S_j)$ can be computed by adopting the cosine similarity measure between the corresponding term vectors $\overrightarrow{V_{S_i}}$ and $\overrightarrow{V_{S_j}}$ of sentence S_i and S_j, whose element's value can be computed by the $TF_{sk}*ISF_k$ formula, where TF_{sk} is the frequency of term T_k in the corresponding sentence and ISF_k is the inverse sentence frequency of term T_k, i.e.1+log(N/N_k), where N is the total number of the sentences in a set and N_k is the number of the sentences containing term T_k.

$$w(S_i, S_j) = \frac{\overrightarrow{V_{S_i}} \cdot \overrightarrow{V_{S_j}}}{|\overrightarrow{V_{S_i}}| \times |\overrightarrow{V_{S_j}}|} \tag{1}$$

Likewise, we can compute the link weight $w(D_i, D_j)$ by adopting the cosine similarity measure between a pair of documents' term vectors $\overrightarrow{V_{D_i}}$ and $\overrightarrow{V_{D_j}}$, whose element's value can be computed by the $TF_{dk}*IDF_k$ formula, where TF_{dk} is the frequency of term T_k in the corresponding document and IDF_k is the inverse document frequency of term T_k.

$$w(D_i, D_j) = \frac{\overrightarrow{V_{D_i}} \cdot \overrightarrow{V_{D_j}}}{|\overrightarrow{V_{D_i}}| \times |\overrightarrow{V_{D_j}}|} \tag{2}$$

The correlation strength between the corresponding sentence S_i and document D_j can be computed by the cosine similarity measure between the sentence's term vector and the document's term vector.

$$w(S_i, D_j) = \frac{\overrightarrow{V_{S_i}} \cdot \overrightarrow{V_{D_j}}}{|\overrightarrow{V_{S_i}}| \times |\overrightarrow{V_{D_j}}|} \tag{3}$$

Where $\overrightarrow{V_{S_i}}$ and $\overrightarrow{V_{D_j}}$ are the term vectors of sentence S_i and document D_j.

2.3 Ranking Homogeneous Objects

Manifold-ranking algorithm is a general graph-based ranking method [8], which takes advantage of local recommendations among the neighboring nodes to rank nodes. In this study, when the sentence set S and the document set D are provided, two subgraphs G_{SS} and G_{DD} can be correspondingly constructed on S and D respectively, where $S = V_{SS} \cup \{q\}$, $D = V_{DD} \cup \{q\}$, and q denotes the given query. In the model, the manifold ranking algorithm [7,8] is further used to assign initial ranking scores (i.e. the biased information richness scores) for two kinds of homogeneous objects (i.e. sentences and documents) individually, which is performed as follows:

Table 1. The manifold ranking algorithm for ranking homogeneous objects

Input:

G_{SS}: The affinity graph of the sentence set S.

G_{DD}: The affinity graph of the document set D.

N: The number of the total sentences in the document set to be summarized.

M: The number of the total documents in the document set to be summarized.

Output:

The limit value $V_{S_i}^*$ of the sentence ranking function V_S: $S \rightarrow \Re$, which can be represented as a vector $V_S=[V_{S_0},..., V_{S_N}]^T$ with each element V_{S_i} denoting the biased information richness score of the corresponding sentence.

The limit value $V_{D_i}^*$ of the document ranking function V_D: $D \rightarrow \Re$, which can be represented as a vector $V_D=[V_{D_0},..., V_{D_M}]^T$ with each element V_{D_i} denoting the biased information richness score of the corresponding document.

Process:

Step 1: Define a prior sentence vector $Y_S=[Y_{S_0},..., Y_{S_N}]^T$ and a prior document vector $Y_D=[Y_{D_0},..., Y_{D_M}]^T$ respectively, in which Y_{S_0} and Y_{D_0} are set to 1 since they both correspond to the given query which can be regarded as the only labeled seed on both of the affinity graphs, and other vector elements in Y_S and Y_D are set to 0.

Step 2: Define the affinity matrix $W_S=(w_{S_{i,j}})_{(N+1)\times(N+1)}$ with each element $w_{S_{i,j}}$ denoting the affinity weight $w(S_i, S_j)$ between the sentences S_i and S_j. Define the affinity matrix $W_D=(w_{D_{i,j}})_{(M+1)\times(M+1)}$ with each element $w_{D_{i,j}}$ denoting the affinity weight $w(D_i, D_j)$ between the documents D_i and D_j.

Step 3: Normalize W_S and W_D by $N_S = D_S^{-1/2} \cdot W_S \cdot D_S^{-1/2}$ and $N_D = D_D^{-1/2} \cdot W_D \cdot D_D^{-1/2}$ respectively, where D_S is the diagonal matrix whose entry (i, j) equals to the sum of the i-th row of W_S and D_D is the diagonal matrix whose entry (i, j) equals to the sum of the i-th row of W_D.

Step 4: Iterate according to the following equations until convergence.

$V_S(t+1)= \lambda_s N_S V_S(t)+ (1- \lambda_s) Y_S$, $V_D(t+1)= \lambda_D N_D V_D(t)+ (1- \lambda_D) Y_D$

Where the parameter $\lambda_s, \lambda_D \in [0, 1]$ specifies the relative contribution to the ranking scores from the neighborhood homogeneous objects and the initial scores.

Step 5: Let $V_{S_i}^*$ and $V_{D_i}^*$ denote the limit of the sequence $\{V_{S_i}(t)\}$ and $\{V_{D_i}(t)\}$ respectively, each sentence S_i gets its ranking score $V_{S_i}^*$ and each document D_i gets its ranking score $V_{D_i}^*$.

In the fourth step of the algorithm, all nodes spread their ranking scores to their neighbors via the corresponding affinity graph, and the whole spreading process is repeated until a stable state is achieved.

2.4 Co-ranking Heterogeneous Objects

In Section 2.3, the initial ranking scores are only determined by homogeneous objects. However, the interactions between heterogeneous objects are not considered. To leverage the above information, the Co-HITS-Ranking algorithm is adopted to rank sentences and documents jointly [9], which can be summarized as follows.

Table 2. The Co-HITS-Ranking algorithm for co-ranking heterogeneous objects

Input:

G_{SD}: The bipartite graph denoting the sentence-to-document correlations.

N: The number of the total sentences in the document set to be summarized.

M: The number of the total documents in the document set to be summarized.

Output:

The limit value $Z_{S_i}^*$ of the sentence ranking function Z_S: $V_{SS} \rightarrow \Re$, which can be represented as a vector Z_S=[Z_{S_1},..., Z_{S_N}]^T with each element Z_{S_i} denoting the significance score of the corresponding sentence.

The limit value $Z_{D_i}^*$ of the document ranking function Z_D: $V_{DD} \rightarrow \Re$, which can be represented as a vector Z_D=[Z_{D_1},..., Z_{D_M}]^T with each element Z_{D_i} denoting the significance score of the corresponding document.

Process:

Step 1: Initialize $Z_S(0)$ and $Z_D(0)$ respectively by set each entry in them to the initial biased information richness score of the corresponding sentence or document.

Step 2: Define the affinity matrix $W_{SD}=(w_{SD_{i,j}})_{N\times M}$ of G_{SD} with each element $w_{SD_{i,j}}$ denoting the affinity weight $w(S_i, D_j)$ (i.e the sentence-to-document correlation strength) between sentence S_i and document D_j by the cosine similarity measure. Then normalize $Z_S(0)$, $Z_D(0)$ and W_{SD} respectively.

Step 3: Iterate according to the following equations until convergence.

$$Z_{S_i}(t+1) = (1-\mu_S)Z_{S_i}(t) + \mu_S \sum_{D_j \in V_{DD}} w_{SD_{i,j}} Z_{D_j}(t) \tag{4}$$

$$Z_{D_j}(t+1) = (1-\mu_D)Z_{D_j}(t) + \mu_D \sum_{S_i \in V_{SS}} w_{SD_{i,j}} Z_{S_i}(t) \tag{5}$$

Where the parameter μ_S, $\mu_D \in [0, 1]$ specifies the relative contribution to the ranking scores from the correlated heterogeneous objects and the objects' latest scores.

Step 4: Let $Z_{S_i}^*$ and $Z_{D_i}^*$ denote the limit of the sequence $\{Z_{S_i}(t)\}$ and $\{Z_{D_i}(t)\}$ respectively, each sentence S_i gets its final significance score $Z_{S_i}^*$ and each document D_i gets its final significance score $Z_{D_i}^*$.

In the study, the interaction information between sentences and documents is encoded by the bipartite graph G_{SD}, which reflects the sentence-to-document correlations in essence. We believe that the direct links and the corresponding correlation strength between sentences and documents may have significant effect on the sentence ranking, so the Co-HITS-Ranking algorithm is used to incorporate the bipartite graph G_{SD} with the content information from both layers to co-rank sentences and documents more effectively. Here the content information from both layers refers to the initial ranking scores of sentences and documents determined by each layer alone.

The final ranking scores of every sentence and document can be got through the above iteratively updating process, which can take into account the mutual influences between documents and sentences and retain their initial scores to an extent at the same time. Therefore, the significance of a sentence is determined ultimately by both its initial significance and the document's significance that is related with it closely.

After the significance score of each sentence has been obtained, a variant of MMR algorithm is employed to remove redundancy and extract summary sentences.

3 Experimental Evaluation

3.1 Dataset and Evaluation Metrics

DUC is a series of evaluation workshops that have been supported by NIST to further progress in automatic summarization. Query-focused multi-document summarization has become the main task since DUC 2005 with the aim to synthesize from a set of documents a well-organized summary to meet the information need. In this study, we use the DUC 2005 dataset for evaluation. Table 3 gives a brief summary of the dataset.

Table 3. The brief summary of the DUC 2005 dataset

Data Source	TREC collections (Los Angeles Times and Financial Times of London)
Number of Topics	50
Number of Relevant Documents Associated with Each Topic	25-50 documents
Number of Human Model Summaries for Each Topic	either 4 or 9
Summary length	250 words

We use ROUGE toolkit [10] as the evaluation utility, which has provided multiple recall-oriented metrics to evaluate the quality of a candidate summary automatically. In our experiments, documents and queries were firstly segmented into sentences. The stop words in both documents and queries were removed. And the average recall scores of the above ROUGE metrics are demonstrated in the experimental results at a confidence level of 95%.

3.2 Experimental Results

In the experiment, the optimized parameters of our approach are set by empirically. The proposed Co-HITS-Ranking based approach (denoted as "CoHR") was firstly compared with the systems participating in DUC 2005. Table 4 lists the ROUGE scores of our summaries and those of the DUC 2005 runs.

Table 4. The ROUGE scores of our summaries and those of the DUC 2005 runs

	ROUGE-1	ROUGE-2	ROUGE-SU4
CoHR	*0.37011*	*0.07012*	*0.12899*
DUC 2005 Median	0.33612	0.05842	0.11205
DUC 2005 Best	0.37515	0.07251	0.13163
DUC 2005 Worst	0.17935	0.02564	0.05569

From Table 4, we can find that the proposed Co-HITS-Ranking based approach can achieve comparable performance to the state-of-the-art systems on the DUC 2005 dataset. The results also demonstrate the effectiveness of the proposed methods, as compared with many different summarization approaches. In addition, to gain a better insight into the proposed approach, we compared it with two baseline methods. One is the lead baseline method (denoted as “LeadBase”), which simply took the first 250 words of the most recent document for each topic in the final summary. The other is the sentence-based manifold-ranking method (denoted as "SenMR"), which makes use of the sentence-to-sentence relationships and the sentence-to-query relationships in a manifold-ranking process to computes each sentence’s information richness score in the documents. Then the same MMR like algorithm with same parameter configuration is applied to reduce redundancy in the ranked sentence list and choose those sentences with highest ranking scores and minimum duplicate information to create the summary according to the length limit.

The "SenMR" method is performed only on the sentence affinity graph G_{SS} without considering the influence of the document-level information that is encoded by G_{DD} and G_{SD}. Specifically, the important information ignored in the "SenMR" includes each document’s significance and the sentence-to-document correlation strength. For the purpose of comparison and simplicity, we use the same value of λ_s in the "SenMR" as in the "CoHR". Table 5 shows the comparison results with the above two baseline methods on the DUC 2005 dataset.

Table 5. Comparison results with two baseline methods on DUC 2005 dataset

	ROUGE-1	ROUGE-2	ROUGE-SU4
CoHR	*0.37011*	*0.07012*	*0.12899*
SenMR	0.36102	0.06128	0.12045
LeadBase	0.27523	0.04026	0.08716

The experimental results shown in Table 5 demonstrate that the proposed Co-HITS-Ranking based approach can outperform both the lead baseline method and the sentence-based manifold-ranking method over three ROUGE metrics. The encouraging performance can be attributed to the following major factors.

Factor 1: Evaluating the initial significance of sentences and documents via the local recommendations between homogeneous objects
To collaboratively evaluate single object's importance via the local recommendations within the homogeneous objects, we make use of the manifold-ranking algorithm to integrate the relationships between homogeneous objects as well as the information about the given query in a unified graph-based score propagation process, which has been proved to be effective in the previous research [7].

Factor 2: Updating the significance scores of sentences and documents via the global mutual reinforcement between heterogeneous objects
In the model, by adopting the Co-HITS-Ranking algorithm, the initial biased information richness scores of sentences and documents can be naturally incorporated in a mutual reinforcement framework to co-rank heterogeneous objects jointly and updating their significance scores adaptively. The updating process can be regarded as a bipartite-graph-based score propagation process based on the heterogeneous relationships between sentences and documents, which can be used to co-rank sentences and documents more effectively.

In summary, the proposed Co-HITS-Ranking based approach can benefit from the integration of two single layer's own information as well as the interaction information between both layers into a unified two-layer graph model, which has been investigated in our preliminary experiment and has shown its superiority to the method that only considers the information from one layer.

4 Conclusion

In this paper, we propose a novel approach to query-focused multi-document summarization, which can extend the existing work by incorporating all kinds of relationships between sentences and documents in a unified two-layer graph model. The main feature of the proposed approach is its ability to evaluate sentences comprehensively by making use of local recommendations within homogeneous objects as well as global mutual reinforcement between heterogeneous objects. Preliminary experimental results on the DUC2005 dataset demonstrate the effectiveness of the proposed approach.

Acknowledgments. This work was supported by the Major Research Plan of National Natural Science Foundation of China (90820005, 90920005), National Natural Science Foundation of China (60773011, 60773167) and Wuhan University 985 Project (985yk004).

References

1. Saggion, H., Bontcheva, K., Cunningham, H.: Robust Generic and Query-Based Summarization. In: 10th Conference of the European Chapter of the Association for Computational Linguistics, pp. 235–238 (2003)

2. Zhao, L., Wu, L.D., Huang, X.J.: Using Query Expansion in Graph-Based Approach for Query, Focused Multi, Document Summarization. Information Processing and Management 45, 35–41 (2009)
3. Wei, F.R., Li, W.J., Lu, Q., He, Y.X.: A Cluster-Sensitive Graph Model for Query-Oriented Multi-Document Summarization. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 446–453. Springer, Heidelberg (2008)
4. Erkan, G., Radev, D.R.: LexRank: Graph-Based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22, 457–479 (2004)
5. Mihalcea, R., Tarau, P.: TextRank–Bringing Order into Text. In: Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
6. Haveliwala, T.H.: Topic-Sensitive PageRank. In: 11th International Conference on World Wide Web, pp. 517–526. ACM, New York (2002)
7. Wan, X.J., Yang, J.W., Xiao, J.G.: Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In: 20th International Joint Conference on Artificial Intelligence, pp. 2903–2908. Morgan Kaufmann Publishers Inc, San Francisco (2007)
8. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on Data Manifolds. In: Advances in Neural Information Processing Systems, vol. 16, pp. 169–176. MIT Press, Cambridge (2004)
9. Deng, H.B., Lyu, M.R., King, I.: A Generalized Co-HITS Algorithm and Its Application to Bipartite Graphs. In: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 239–248. ACM, New York (2009)
10. Lin, C.Y., Hovy, E.: Automatic Evaluation of Summaries Using N-Gram Cooccurrence Statistics. In: Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 71–78 (2003)