Improved Affinity Graph Based Multi-Document Summarization

Xiaojun Wan, Jianwu Yang
Institute of Computer Science and Technology, Peking University
Beijing 100871, China
{wanxiaojun, yangjianwu}@icst.pku.edu.cn

Abstract

This paper describes an affinity graph based approach to multi-document summarization. We incorporate a diffusion process to acquire semantic relationships between sentences, and then compute information richness of sentences by a graph rank algorithm on differentiated intra-document links and inter-document links between sentences. A greedy algorithm is employed to impose diversity penalty on sentences and the sentences with both high information richness and high information novelty are chosen into the summary. Experimental results on task 2 of DUC 2002 and task 2 of DUC 2004 demonstrate that the proposed approach outperforms existing state-of-the-art systems.

1 Introduction

Automated multi-document summarization has drawn much attention in recent years. Multi-document summary is usually used to provide concise topic description about a cluster of documents and facilitate the users to browse the document cluster. A particular challenge for multi-document summarization is that the information stored in different documents inevitably overlaps with each other, and hence we need effective summarization methods to merge information stored in different documents, and if possible, contrast their differences.

A variety of multi-document summarization methods have been developed recently. In this study, we focus on extractive summarization, which involves assigning saliency scores to some units (e.g. sentences, paragraphs) of the documents and extracting the sentences with highest scores. MEAD is an implementation of the centroid-based method (Radev et al., 2004) that scores sentences based on sentence-level and inter-sentence features, including cluster centroids, position, TF*IDF, etc. NeATS (Lin and Hovy, 2002) selects important content using sentence position, term frequency, topic signature and term clustering, and then uses MMR (Goldstein et al., 1999) to remove redundancy. XDoX (Hardy et al., 1998) identifies the most salient themes within the set by passage clustering and then composes an extraction summary, which reflects these main themes. Harabagiu and Lacatusu (2005) investigate different topic representations and extraction methods.

Graph-based methods have been proposed to rank sentences or passages. Websumm (Mani and Bloedorn, 2000) uses a graph-connectivity model and operates under the assumption that nodes which are connected to many other nodes are likely to carry salient information. LexPageRank (Erkan and Radev, 2004) is an approach for computing sentence importance based on the concept of eigenvector centrality. Mihalcea and Tarau (2005) also propose similar algorithms based on PageRank and HITS to compute sentence importance for document summarization.

In this study, we extend the above graph-based works by proposing an integrated framework for considering both information richness and information novelty of a sentence based on sentence affinity graph. First, a diffusion process is imposed on sentence affinity graph in order to make the affinity graph reflect true semantic relationships between sentences. Second, intra-document links and inter-document links between sentences are differentiated to attach more importance to inter-document links for sentence information richness computation. Lastly, a diversity penalty process is imposed on sentences to penalize redundant sentences. Experiments on DUC 2002 and DUC 2004 data are performed and we obtain encouraging results and conclusions.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 181–184,
New York, June 2006.

2 The Affinity Graph Based Approach

The proposed affinity graph based summarization method consists of three steps: (1) an affinity graph is built to reflect the semantic relationship between sentences in the document set; (2) information richness of each sentence is computed based on the affinity graph; (3) based on the affinity graph and the information richness scores, diversity penalty is imposed to sentences and the affinity rank score for each sentence is obtained to reflect both information richness and information novelty of the sentence. The sentences with high affinity rank scores are chosen to produce the summary.

2.1 Affinity Graph Building

Given a sentence collection $S=\{s_i \mid 1\le i\le n\}$, the affinity weight $aff(s_i, s_j)$ between a sentence pair of s_i and s_j is calculated using the cosine measure. The weight associated with term t is calculated with the tf_t*isf_t formula, where tf_t is the frequency of term t in the corresponding sentence and isf_t is the inverse sentence frequency of term t, i.e. $1+\log(N/n_t)$, where N is the total number of sentences and n_t is the number of sentences containing term t. If sentences are considered as nodes, the sentence collection can be modeled as an undirected graph by generating the link between two sentences if their affinity weight exceeds 0, i.e. an undirected link between s_i and s_j ($i\neq j$) with affinity weight $aff(s_i,s_j)$ is constructed if $aff(s_i,s_j)>0$; otherwise no link is constructed. Thus, we construct an undirected graph G reflecting the semantic relationship between sentences by their content similarity. The graph is called as Affinity Graph. We use an adjacency (affinity) matrix **M** to describe the affinity graph with each entry corresponding to the weight of a link in the graph. $\mathbf{M} = (M_{i,j})_{n\times n}$ is defined as follows:

$$M_{i,j} = aff(s_i, s_j) \tag{1}$$

Then **M** is normalized to make the sum of each row equal to 1. Note that we use the same notation to denote a matrix and its normalized matrix.

However, the affinity weight between two sentences in the affinity graph is currently computed simply based on their own content similarity and ignore the affinity diffusion process on the graph. Other than the direct link between two sentences, the possible paths with more than two steps between the sentences in the graph also convey more or less semantic relationship. In order to acquire the implicit semantic relationship between sentences, we apply a diffusion process（Kandola et al., 2002）on the graph to obtain a more appropriate affinity matrix. Though the number of possible paths between any two given nodes can grow exponentially, recent spectral graph theory (Kondor and Lafferty, 2002) shows that it is possible to compute the affinity between any two given nodes efficiently without examining all possible paths. The diffusion process on the graph is as follows:

$$\widetilde{\mathbf{M}} = \sum_{t=1}^{\infty} \gamma^{t-1} \mathbf{M}^t \tag{2}$$

where $\gamma(0<\gamma<1)$ is the decay factor set to 0.9. $\mathbf{M}^t$ is the t-th power of the initial affinity matrix $\mathbf{M}$ and the entry in it is given by

$$M^t_{i,j} = \sum_{\substack{u\in\{1,\ldots,n\}^t \\ u_1=i,\, u_t=j}} \prod_{\ell=1}^{t-1} M_{u_\ell, u_{\ell+1}} \tag{3}$$

that is the sum of the products of the weights over all paths of length t that start at node i and finish at node j in the graph on the examples. If the entries satisfy that they are all positive and for each node the sum of the connections is 1, we can view the entry as the probability that a random walk beginning at node i reaches node j after t steps. The matrix $\widetilde{\mathbf{M}}$ is normalized to make the sum of each row equal to 1. t is limited to 5 in this study.

2.2 Information Richness Computation

The computation of information richness of sentences is based on the following three intuitions: 1) the more neighbors a sentence has, the more informative it is; 2) the more informative a sentence's neighbors are, the more informative it is; 3) the more heavily a sentence is linked with other informative sentences, the more informative it is. Based on the above intuitions, the information richness score $InfoRich(s_i)$ for a sentence s_i can be deduced from those of all other sentences linked with it and it can be formulated in a recursive form as follows:

$$InfoRich(s_i) = d \cdot \sum_{\text{all } j\neq i} InfoRich(s_j) \cdot \widetilde{M}_{j,i} + \frac{(1-d)}{n} \tag{4}$$

And the matrix form is:

$$\vec{\lambda} = d\widetilde{\mathbf{M}}^T \vec{\lambda} + \frac{(1-d)}{n}\vec{e} \tag{5}$$

where $\vec{\lambda} = [\text{InfoRich}(s_i)]_{n\times 1}$ is the eigenvector of $\widetilde{\mathbf{M}}^T$. $\vec{e}$ is a unit vector with all elements equaling to 1. d is the damping factor set to 0.85.

Note that given a link between a sentence pair of s_i and s_j, if s_i and s_j comes from the same document, the link is an intra-document link; and if s_i and s_j comes from different documents, the link is an inter-document link. We believe that inter-document links are more important than intra-document links for information richness computation. Different weights are assigned to intra-document links and inter-document links respectively, and the new affinity matrix is:

$$\hat{\mathbf{M}} = \alpha\widetilde{\mathbf{M}}_{\text{intra}} + \beta\widetilde{\mathbf{M}}_{\text{inter}} \tag{6}$$

where $\widetilde{\mathbf{M}}_{\text{intra}}$ is the affinity matrix containing only the intra-document links (the entries of inter-document links are set to 0) and $\widetilde{\mathbf{M}}_{\text{inter}}$ is the affinity matrix containing only the inter-document links (the entries of intra-document links are set to 0). α, β are weighting parameters and we let 0≤α, β≤1. The matrix is normalized and now the matrix $\widetilde{\mathbf{M}}$ is replaced by $\hat{\mathbf{M}}$ in Equations (4) and (5).

2.3 Diversity Penalty Imposition

Based on the affinity graph and obtained information richness scores, a greedy algorithm is applied to impose the diversity penalty and compute the final affinity rank scores of sentences as follows:

1. Initialize two sets A=Ø, B={s_i | i=1,2,…,n}, and each sentence's affinity rank score is initialized to its information richness score, i.e. ARScore(s_i) = InfoRich(s_i), i=1,2,…n.
2. Sort the sentences in B by their current affinity rank scores in descending order.
3. Suppose s_i is the highest ranked sentence, i.e. the first sentence in the ranked list. Move sentence s_i from B to A, and then a diversity penalty is imposed to the affinity rank score of each sentence linked with s_i as follows:
   For each sentence s_j in B, we have
   $$\text{ARScore}(s_j) = \text{ARScore}(s_j) - \omega \cdot \widetilde{M}_{j,i} \cdot \text{InfoRich}(s_i) \tag{7}$$
   where ω>0 is the penalty degree factor. The larger ω is, the greater penalty is imposed to the affinity rank score. If ω=0, no diversity penalty is imposed at all.
4. Go to step 2 and iterate until B= Ø or the iteration count reaches a predefined maximum number.

After the affinity rank scores are obtained for all sentences, the sentences with highest affinity rank scores are chosen to produce the summary according to the summary length limit.

3 Experiments and Results

We compare our system with top 3 performing systems and two baseline systems on task 2 of DUC 2002 and task 4 of DUC 2004 respectively. ROUGE (Lin and Hovy, 2003) metrics is used for evaluation[1] and we mainly concern about ROUGE-1. The parameters of our system are tuned on DUC 2001 as follows: ω=7, α=0.3 and β=1.

We can see from the tables that our system outperforms the top performing systems and baseline systems on both DUC 2002 and DUC 2004 tasks over all three metrics. The performance improvement achieved by our system results from three factors: diversity penalty imposition, intra-document and inter-document link differentiation and diffusion process incorporation. The ROUGE-1 contributions of the above three factors are 0.02200, 0.00268 and 0.00043 respectively.

System	ROUGE-1	ROUGE-2	ROUGE-W
Our System	**0.38125**	**0.08196**	**0.12390**
S26	0.35151	0.07642	0.11448
S19	0.34504	0.07936	0.11332
S28	0.34355	0.07521	0.10956
Coverage Baseline	0.32894	0.07148	0.10847
Lead Baseline	0.28684	0.05283	0.09525

Table 1. System comparison on task 2 of DUC 2002

System	ROUGE-1	ROUGE-2	ROUGE-W
Our System	**0.41102**	**0.09738**	**0.12560**
S65	0.38232	0.09219	0.11528
S104	0.37436	0.08544	0.11305
S35	0.37427	0.08364	0.11561
Coverage Baseline	0.34882	0.07189	0.10622
Lead Baseline	0.32420	0.06409	0.09905

Table 2. System comparison on task 2 of DUC 2004

Figures 1-4 show the influence of the parameters in our system. Note that α: β denotes the real values α and β are set to. "w/ diffusion" is the system with the diffusion process (our system) and "w/o diffusion" is the system without the diffusion proc-

[1] We use ROUGEeval-1.4.2 with "-l" or "-b" option for truncating longer summaries, and "-m" option for word stemming.

ess. The observations demonstrate that "w/ diffusion" performs better than "w/o diffusion" for most parameter settings. Meanwhile, "w/ diffusion" is more robust than "w/o diffusion" because the ROUGE-1 value of "w/ diffusion" changes less when the parameter values vary. Note that in Figures 3 and 4 the performance decreases sharply with the decrease of the weight β of inter-document links and it is the worst case when inter-document links are not taken into account (i.e. α:β=1:0), while if intra-document links are not taken into account (i.e. α:β=0:1), the performance is still good, which demonstrates the great importance of inter-document links.

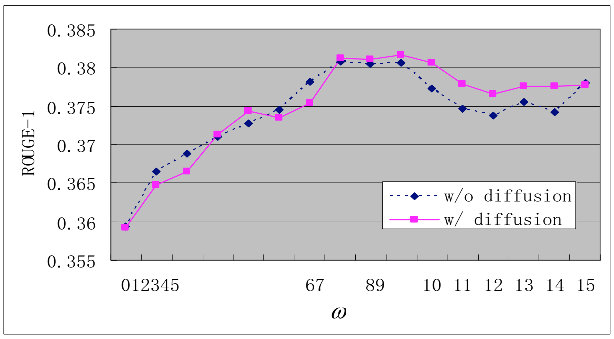

Figure 1. Penalty factor tuning on task 2 of DUC 2002

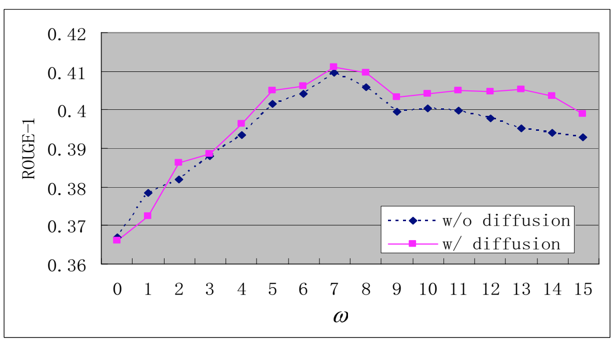

Figure 2. Penalty factor tuning on task 2 of DUC 2004

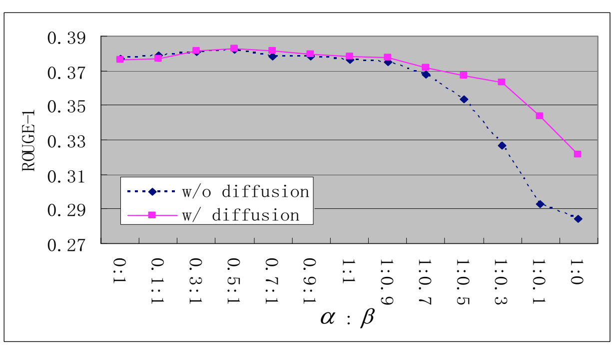

Figure3. Intra- & Inter-document link weight tuning on task 2 of DUC 2002

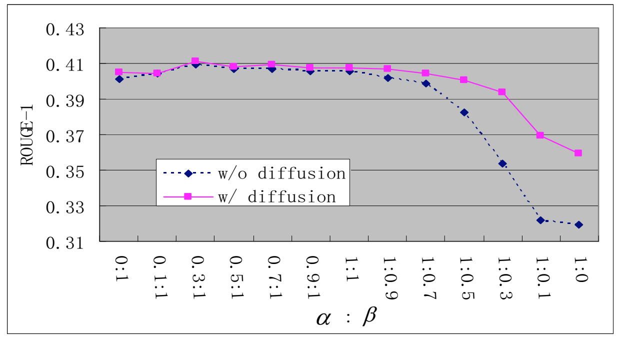

Figure 4. Intra- & Inter-document link weight tuning on task 2 of DUC 2004

References

G. Erkan and D. Radev. LexPageRank: prestige in multi-document text summarization. In Proceedings of EMNLP'04

J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. Proceedings of SIGIR-99.

S. Harabagiu and F. Lacatusu. Topic themes for multi-document summarization. In Proceedings of SIGIR'05, Salvador, Brazil, 202-209, 2005.

H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise, and X. Zhang. Cross-document summarization by concept classification. In Proceedings of SIGIR'02, Tampere, Finland, 2002.

J. Kandola, J. Shawe-Taylor, N. Cristianini. Learning semantic similarity. In Proceedings of NIPS'2002.

K. Knight and D. Marcu. Summarization beyond sentence extraction: a probabilistic approach to sentence compression, Artificial Intelligence, 139(1), 2002.

R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In Proceedings of ICML'2002.

C.-Y. Lin and E.H. Hovy. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In Proceedings of HLT-NAACL 2003.

C.-Y. Lin and E.H. Hovy. From Single to Multi-document Summarization: A Prototype System and its Evaluation. In Proceedings of ACL-2002.

I. Mani and E. Bloedorn. Summarizing Similarities and Differences Among Related Documents. Information Retrieval, 1(1), 2000.

R. Mihalcea and P. Tarau. A language independent algorithm for single and multiple document summarization. In Proceedings of IJCNLP'2005.

D. R. Radev, H. Y. Jing, M. Stys and D. Tam. Centroid-based summarization of multiple documents. Information Processing and Management, 40: 919-938, 2004.