Social Summarization via Automatically Discovered Social Context

Po Hu[1][2], Cheng Sun[1], Longfei Wu[1], Donghong Ji[1] and Chong Teng[1]
[1]Computer School, Wuhan University, China
[2]Department of Computer Science, Huazhong Normal University, China

phu@mail.ccnu.edu.cn, gensun.cc@gmail.com,
lonfee88@gmail.com, donghong_ji2000@yahoo.com.cn,
tchong616@126.com

Abstract

Heavy research has been done in recent years on tasks of traditional summarization. However, social context, which is critical in building high-quality social summarizer for web documents, is usually neglected. To address this issue, we propose a novel summarization approach based on social context. In this approach, social summarization is implemented by first employing the tripartite clustering algorithm to simultaneously discover document context and user context for a specified document. Then sentence relationships intra and inter documents plus intended user communities are taken into account to evaluate the significance of each sentence in different context views. Finally, a few sentences with highest overall scores are selected to form the summary. Experimental results demonstrate the effectiveness of the proposed approach and show the superior performance over several baselines.

1 Introduction

Now, an increasing number of Web 2.0 applications (e.g. Del.icio.us, CiteULike, LinkedIn) are allowing users to play more active roles such as annotating online documents with free-form tags, submitting opinions on the items they are interested in, or participating in social networks, etc.

Despite their focus on different resources, all of them have the common purpose of helping users organize, retrieve and share knowledge. The rich information offered by these systems provides additional clues for collaboratively summarizing online documents in a social context. However, most existing methods generate a summary based only on the information within each document or its neighboring documents, while the social context is usually ignored. This is an important issue that has not been exten-sively investigated in the summarization's literature.

In this study, different user's social tagging history and their tagged documents are employed and a novel social summarization approach is proposed by considering both document context and user context in the sentence evaluation process. The intuition is that if an appropriate social context is available, then integration of social context knowledge into the existing sentence evaluation process will improve the performance of traditional summarization methods.

The proposed approach takes into account both the mutual influences between documents and the impact from different user communities, which consists of the following phases.

Firstly, the tripartite clustering algorithm is adopted to discover document context and user context by clustering documents, users and tags simultaneously, which is based on the inherent structure of the three kinds of objects. In the clustering result, each document cluster is regarded as a context for each document in the cluster and all the user clusters are regarded as user communities with diverse information preferences.

Secondly, the context-sensitive ranking algorithm is applied for evaluating the significance of each sentence in different context views respectively.

Thirdly, a few significant sentences with highest overall scores are extracted from the specified document to generate the summary.

The main contribution of this paper is summarized as follows:

1) We examine an important factor called social context for document summarization.

2) We propose a novel social summarization approach by incorporating both document context and user context in the sentence evaluation process.

3) We conduct experiments to validate the effectiveness of the proposed approach on the

dataset sampled from del.icio.us and investigate how the parameters influence the summarization performance.

The paper is organized as follows: Section 2 introduces related work. Section 3 describes the details of the proposed social summarization approach. Section 4 presents experimental results and analysis. Lastly we conclude our paper in Section 5.

2 Related Work

Document summarization aims to automatically create a concise representation of a given document that delivers the main content of it. To date, a variety of methods have been developed, which can be roughly divided into two types: extractive approach and abstractive approach.

The former directly assigns each sentence a significance score and extracts a few sentences of highest scores from the original document, which depends on the combination of implicit or explicit statistical or linguistic features, while the latter usually makes use of advanced natural language understanding or generation technique to fuse, compress or reformulate information. In this paper, we focus on the extractive approach.

Much work has been done on extractive summarization including classification-based methods (Conroy and O'leary, 2001), regression-based methods (You et al., 2011), NMF-based methods (Lee et al., 2009), MMR-based methods (Carbonell and Goldstein, 1998), clustering-based methods (Nomoto and Matsumoto, 2001), etc. Recently, graph-based ranking methods are becoming more and more popular. LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), mutual reinforcement based ranking (Zha, 2002), and manifold ranking (Wan et al., 2007) are such methods using algorithms similar to PageRank and HITS to compute sentence's significance.

When summarizing a specified document, most methods employ only the information contained in the document while ignore its context. One exception is the collaborative approach proposed by Wan and Yang (2007), which improves news document summarization by use of the content from neighboring documents. This motivates us to further consider how social context knowledge might be incorporated in the sentence evaluation process to improve the performance of traditional summarization systems.

To date, much work on summarization tends to focus on a specific type of document, such as news articles (McKeown and Radev, 1995), academic papers (Qazvinian and Radev, 2010) or medical records (Afantenos et al., 2005). With the rapid growth of documents over the Internet, a large number of web documents need to be summarized. However, the content contained in a web document is observed to be sparser and noisier, and it is difficult for the traditional summarization methods that only focus on the local content of a document to capture the true meaning of web documents in a richer context environment. So it is more reasonable to summarize the web document by taking advantage of its social context (i.e. document context and user context).

Relevant work on web document summarization includes harnessing the search engine's click-through data to guide summarization (Sun et al., 2005), producing summaries by using query-result selection pairs according to their relative importance (Boydell and Smyth, 2007), etc.

The work described in this paper is concerned with producing an extractive summary for a Web document. Different from existing summarization methods that use document content or document context alone, the novelty of our approach stems from the integration of an important factor (i.e. social context) for sentence evaluation. We focus on how the richer information from social context can be utilized to improve the summarization performance, which is an interesting issue that needs to be carefully investigated.

3 The Proposed Social Summarization Approach

3.1 Overview

The main idea of the proposed summarization approach is to incorporate an important factor into sentence evaluation process by discovering social context knowledge from online bookmarking services and utilizing the discovered knowledge to collaboratively evaluate each sentence's significance.

Given a document to be summarized, the approach first clusters three types of objects (i.e. documents, users, and tags) simultaneously in a unified framework so that social context can be identified automatically. The identified document context is a cluster of documents, which are topically close to the specified document and are tagged by like-minded users. The identified user contexts include multiple user clusters with each

representing a unique user community with different information preference. The discovered social context knowledge is deemed beneficial to evaluating sentences' significance from diverse views since they can provide richer external knowledge and more complementary clues to help rank sentences comprehensively.

Then the context-sensitive ranking algorithm is adopted to score all the sentences in the document context that the specified document belongs to by differentiating inter-sentence relationships in document context view and considering the impact of different user communities in user context view.

Finally, each sentence's overall significance is computed by fusing their scores from different views and a few sentences with highest overall scores are extracted from the specified document to generate the final summary.

3.2 Social Context Discovery

Social context discovery aims to find not only a set of neighboring documents similar to the specified document but also a group of intended user communities with different information preferences. Figure 1 shows an example of the social context for document d.
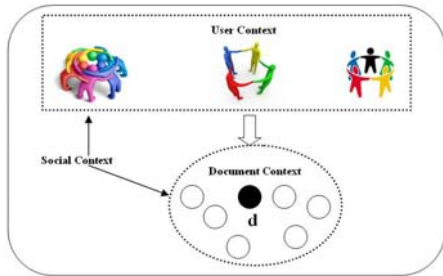

Figure 1. An Example of the Social Context for Document d.

A major characteristic of the social context knowledge acquired from bookmarking services is the tripartite relationships formed through users' social tagging behaviors, which can be illustrated through the example shown in Figure 2.
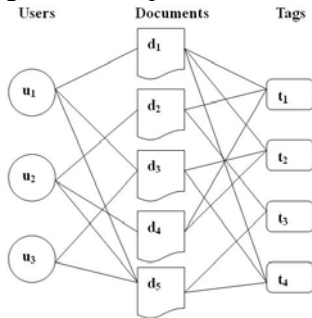

Figure 2. An Example of Tripartite Relationships among Users, Documents, and Tags.

In Figure 2, whenever a user annotates a document with a tag, a ternary relationship is built among the three kinds of objects. The tripartite nature among documents, users, and tags provides valuable information for discovering the topic-related documents, the intended user communities, and the semantics of tags. Besides, in the del.icio.us bookmarking service, it can be observed that topic-related documents are usually annotated with semantically-related tags by like-minded users with common interests, so in this study, the tripartite clustering algorithm (Lu et al., 2009) is employed to cluster documents, users, and tags simultaneously based on the inherent structure of the three kinds of objects to automatically discover the social context for a specific document.

The tripartite relationships among documents, users, and tags can be formally represented by a graph denoted as: $G=(D, U, T, E_{UD}, E_{UT}, E_{DT})$, where D, U, and T are the sets of documents, users and tags respectively, and E_{UD}, E_{UT}, and E_{DT} denote the relationships between user-document, user-tag, and document-tag respectively.

In the tripartite graph, each kind of object can be represented by a combined vector. For example, a document is naturally related to the users who have tagged it and the tags which have been used to tag it, so a document d_i can be represented by a combined vector D_i consisting of two components with one denoting user link vector and the other denoting tag link vector. i.e. $D_i=(D_i^{(U)}, D_i^{(T)})$, $D_i^{(U)}= (y_{ih}^{(U)} \mid h=1,2,...,|U|)$, $D_i^{(T)} =(y_{ij}^{(T)} \mid j=1,2,...,|T|)$, where $y_{ih}^{(U)}$ denotes the times that d_i is annotated by user u_h, $|U|$ denotes the total number of users, $y_{ij}^{(T)}$ denotes the times that d_i has been annotated with tag t_j, and $|T|$ denotes the total number of tags. Likewise, user and tag can be represented in the similar way.

Assuming $C_m(D)$ represent a document cluster, we can calculate the value of the centroid vector at user dimension by formula (1).

$$Centroid_{mu}^{(U)} = \frac{\sum_{d_i \in C_m^{(D)}, u_h \in C_l^{(U)}} y_{ih}^{(U)}}{|C_m^{(D)}| * |C_l^{(U)}|}, (u_u \in C_l^{(U)}) \qquad (1)$$

where $C_l^{(U)}$ is a user cluster which user u_u belongs to, u_h is any user in $C_l^{(U)}$, and d_i is any document in $C_m^{(D)}$. It can be seen that the value of a document cluster's centroid at user dimension does not only depends on the links from u_u to all the cluster's documents, but also relies on the links from other users belonging to the same cluster as u_u to the cluster's documents. Likewise, the value of the centroid vector at tag dimension can be

calculated in the similar way and the similarity value between a document d_i and the centroid of a document cluster $C_m^{(D)}$ can be calculated as follows,

$$Similarity(d_i, Centroid_m^{(D)}) = \alpha * Similarity(D_i^{(U)}, Centorid_m^{(U)}) \quad (2)$$
$$+ (1-\alpha) * Similarity(D_i^{(T)}, Centroid_m^{(T)})$$

where $Simlarity(D_i^{(U)}, Centorid_m^{(U)})$ denotes the similarity between d_i and the centroid of $C_m^{(D)}$ based on the user link vector, $Similarity(D_i^{(T)}, Centroid_m^{(T)})$ denotes the similarity of them based on the tag link vector, and α is a weighting adjusting parameter usually set to 0.5.

In this study, social context can be discovered by the tripartite clustering algorithm described in Table 1 (Lu et al., 2009).

Algorithm: Tripartite clustering based social context discovery

Input: The tripartite graph G=(D, U, T, E_{UD}, E_{UT}, E_{DT}) encoding the relationships among documents, users, and tags, the predefined number of document clusters N_{dc}, the predefined number of user clusters N_{uc}, and the predefined number of tag clusters N_{tc}.

Output: The social context, which includes not only the document context but also the user context.

1: Assign each document (user or tag) to a random document cluster (user cluster or tag cluster);

2: **Repeat:**

3:     **For** each type of objects **do**

4:         Compute the centroid of each cluster based on the link features of its cluster members and the cluster structures of other two types of objects from last iteration;

5:     **For** each document (user or tag) **do**

6:         Compute the similarity between the object and the centroid of each document (user or tag) cluster;

7:         Reassign the current object to the closest cluster based on the similarity.

8:     **End For**

9:     **End For**

10: **Until** the assignments no longer change.

Table 1. Tripartite Clustering Based Social Context Discovery

After that, we can discover all the document clusters. Each document cluster is regarded as a document context for any document in the cluster. At the same time, all user communities with varied information preferences can be also found. Since the 'true' numbers of document clusters, user cluster, and tag clusters are hard to predict, in this study we simply set them to the square root of the total number of documents, users, and tags respectively.

The potential benefit of adopting the tripartite clustering algorithm for social context discovery is that it can incorporate the social tagging information in a unified framework and all social context information can be simultaneously obtained by making use of the interactions among the cluster structures of different types of objects.

3.3 Social Context Based Summarization

In this study, for summarizing a single document, all the documents in its document context are firstly segmented into sentences and each sentence is evaluated in document context view and user context view respectively. Then the two scores are fused to evaluate the overall significance of a sentence. Lastly, a few sentences with highest overall scores are extracted from the specified document to generate the summary.

Sentence Scoring in Document Context View

The document-context-sensitive ranking algorithm is applied on the document context of the specified document for scoring sentences collaboratively (Wan, 2008).

In document context view, inter-sentence relationships are described by sentence affinity graph G_s with each vertex s_i representing a sentence and each edge e_{ij} representing the relationship between sentence s_i and s_j ($i \neq j$) whose weight is the similarity between the pair of sentences. G_s can be encoded by either the matrix M_{intra} or the matrix M_{inter} with each entry corresponding to the edge's weight of G_s's sub-graph G_{intra} and G_{inter}, which describe either the within-document relationships or the cross-document relationships among sentences. Then M_{intra} and M_{inter} are normalized to $\widetilde{M}_{intra}$ and $\widetilde{M}_{inter}$ by making the sum of each row equal to 1.

The document-context-sensitive score of sentence s_i is denoted as $DCScore(s_i)$ that can be computed as follows:

$$DCScore(s_i) = \lambda * DCScore_{intra}(s_i) + (1-\lambda) * DCScore_{inter}(s_i) \quad (3)$$

$$DCScore_{intra}(s_i) = \delta * \sum_{all\ j \neq i} DCScore_{intra}(s_j) * (\widetilde{M}_{intra})_{j,i} + \frac{(1-\delta)}{n} \quad (4)$$

$$DCScore_{inter}(s_i) = \delta * \sum_{all\ j \neq i} DCScore_{inter}(s_j) * (\widetilde{M}_{inter})_{j,i} + \frac{(1-\delta)}{n} \quad (5)$$

where n is the number of sentences in the document context, s_j is any other sentence linked with s_i, $DCScore_{intra}$ and $DCScore_{inter}$ are the sentence scores by considering either the within-document relationship or the cross-document relationship. δ is a damping factor usually set to 0.85 as in PageRank and λ is a weight adjusting parameter specifying the relative contribution to the score from the within-document relationship and the cross-document relationship. Since the previous research (Wan, 2008) has demonstrated

that the use of cross-document relationships between sentences can much improve the performance of summarization, in this study λ is set to 0.4 to enhance the contribution from cross-document relationship.

Sentence Scoring in User Context View

Since the discovered user contexts represent different user communities, when evaluating the sentence's significance within the specified document, we should take into account the recommendation from diverse user communities. However, how to evaluate the recommendation strength becomes a difficult problem. In this paper, we propose a relevance measurement to evaluate it by computing the affinity between the document context of the specified document and the profile of each user community. The reason of using the document context instead of the document is that the expanded document context includes richer information than the single document, which can be used to match the community profile better.

In delicious, the documents annotated by a user can reflect the user's information preference to certain extent. Therefore, for a user community and two documents x and y, if the number of users in the community who annotate document x is greater than that of users who annotate document y, we may assume that the community is more interested in the content of document x than that of document y. Based on this assumption, we model the user community profile by choosing a certain number of representative documents that have been annotated by the most of users in this community. In this study, twenty percent of documents have been selected.

For scoring sentence in the user context view, each user context UC_k is firstly transformed into a pseudo-query q_k that is represented by the centroid vector of all the sentences in the user community profile. The affinity graph G_{uk} is constructed in which the vertexes include all the sentences in the specified document's context and the kth pseudo-query associated with the kth user context, and the edges encode both the relationships among the sentences and the relationship between the kth pseudo-query and the sentences. Here q_k can be processed in the same way as other sentences.

The user-context-sensitive score $UCScore_k (s_i)$ for sentence s_i in the kth user context view can be deduced from those of other sentences linked with it and the kth user context, which can be computed by the query-sensitive ranking algorithm as follows:

$$UCScore_k(s_i) = \beta * \sum_{all\ j \neq i} UCScore_k(s_j) * (\widetilde{M}_{uk})_{j,i} + (1-\beta) * Rel(s_i, q_k) \quad (6)$$

where $\widetilde{M}_{uk}$ is the normalized affinity matrix of G_{uk}, $Rel(s_i, q_k)$ denotes the Cosine relevance of the sentence s_i to the pseudo-query q_k, and β is a damping factor usually set to 0.85. The user-context-sensitive score for sentence s_i in the rest of user contexts can be deduced in the same way.

The final score of sentence s_i assigned in the user context view can be denoted as $UCScore(s_i)$, which is calculated by the combination of all scores from different user contexts.

$$UCScore(s_i) = \frac{\sum_{k=1}^{N_{uc}} RS(UC_k, DC_{s_i}) * UCScore_k(s_i)}{\sum_{k=1}^{N_{uc}} RS(UC_k, DC_{s_i})} \quad (7)$$

where N_{uc} is the number of user contexts, $RS(UC_k, DC_{s_i})$ denotes the recommendation strength of the user context UC_k to the document context DC_{s_i} of sentence s_i.

Summary Generation

In order to evaluate the overall score of each sentence s_i in the social context view, we fuse both document-context-sensitive score $DCScore(s_i)$ and user-context-sensitive score $UCScore(s_i)$ in a unified way as follows:

$$Score(s_i) = \eta * UCScore(s_i) + (1-\eta) * DCScore(s_i) \quad (8)$$

where $\eta \in [0,1]$ is a weight adjusting parameter, specifying the relative contribution to the overall scores from user context view and document context view. If $\eta = 1$, only the user context's impact is considered and the score of sentence s_i equals to $UCScore(s_i)$; if $\eta = 0$, only the document context's impact is considered and the score of sentence s_i equals to $DCScore(s_i)$; and if $\eta = 0.5$, the two context view's impacts are considered equally.

Finally, a few sentences with highest overall scores and least redundancy are chosen into the summary according to the summary length limit.

4 Experiments

4.1 Experimental Setup

Because there is no existing benchmark dataset for social summarization, we construct a real-world dataset to evaluate the proposed method by downloading 200 bookmarked CNN news web documents from del.icio.us website on diverse topics (e.g. financial crisis, accidents and natural disasters, health, sports, etc). The "Story

Highlights" texts are extracted from each CNN news document to form the gold-standard (model) summary, which contains about 50-100 words.

The detailed statistical result of the dataset is shown in Table 2.

Summary of the Dataset	
Number of documents	200
Number of users	1194
Number of tags	2186

Table 2. The Statistical Result of the Dataset

Both intrinsic and extrinsic methods are proposed for summarization evaluation. In this paper, we employ the intrinsic method to evaluate the proposed summarization approach and all the baselines.

To date, various intrinsic evaluation methods such as ROUGE (Lin and Eduard, 2003) and Pyramid (Nenkova and Passonneau, 2004) have been proposed. In the study, The ROUGE-1.5.5 toolkit is adopted because it was officially adopted by DUC (Now TAC) for automatic summarization evaluation and has been shown to correlate with human evaluations well. ROUGE metrics measure a summary's content quality by counting overlapping units such as n-gram, word sequences, and word pairs between the automatically generated summary and the gold-standard summaries. Formally, ROUGE-N is an n-gram recall based measurement between a candidate summary and a set of reference summaries, which is computed as follows (Lin and Eduard, 2003):

$$ROUGE-N = \frac{\sum_{S \in \{reference\ summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{reference\ summaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (9)$$

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

A few recall-oriented ROUGE metrics have been employed such as ROUGE-1 (unigram based metric), ROUGE-2 (bigram based metric), and ROUGE-SU4 (skip bigram and unigram based metric with maximum skip distance 4), etc. We report the metric scores of ROUGE-1, ROUGE-2 and ROUGE-SU4 at the confidence level of 95% in the following experiments.

4.2 Experimental Results

As a preprocessing step, in the following experiments, all the documents were segmented into sentences, stop-words were removed and the remaining words were stemmed by the Porter Stemmer. All the sentences were represented as the term vectors according to TF*ISF scheme. The process of redundancy removing and the setup of the corresponding parameters of the following baselines are also the same as that of the proposed approach.

For comparison, given a document and its document context discovered by the tripartite clustering algorithm, we implement the following methods as the baselines and each method generates a summary for each document in accordance to the length of the corresponding model summary. Since the good performance of the tripartite clustering algorithm adopted in this paper has been validated in the previous study (Lu et al., 2009), which shows that it can be applied to cluster different types of objects simultaneously and significantly outperforms the content-based K-means algorithm. In this study, we don't compare it again with the K-means algorithm in the discovery of document context.

Baseline 1 (RANDSum): RANDSum selects sentences randomly from the specified document to generate a summary.

Baseline 2 (DCISum): DCISum is a document-context-independent method which computes the significance score of a sentence based only on the within-document relationships while ignoring the document context. In this study, it is realized according to formula (4).

Baseline 3 (DCDSum): DCDSum is a document-context-dependent method which computes the significance score of a sentence based on both the cross-document relationships and the within-document relationships. In this study, it is realized according to formula (3).

Note that all the above baseline methods depend either on the internal information of the specified document or the external information of the document context, yet the user context information is entirely neglected. The proposed social context based approach (*abbr.* **SCSum**) considers not only document context but also user context in a unified framework.

We show the summarization evaluation results of different approaches in Tables 3.

Method	ROUGE-1	ROUGE-2	ROUGE-SU4
RANDSum	0.25532	0.02174	0.05634
DCISum	0.33961	0.05576	0.09039
DCDSum	0.34273	0.05522	0.09113
SCSum	**0.35880**	**0.07130**	**0.11417**

Table 3. The Summarization Evaluation Results of Different Approaches

In Table 3, the best result of our approach is achieved when the weight adjusting parameter η specifying the relative contribution from user context and document context is set to 0.4.

Seen from Table 3, our proposed approach SCSum using the discovered social context knowledge achieves the best performance on all ROUGE metrics comparable to that of the baseline approaches (i.e. RANDSum, DCISum, and DCDSum), which demonstrates that both document context and user context are very important for improving the performance of document summarization if richer context information is available for a specific document.

We also observe that the DCDSum that uses the document context information performs better than the DCISum, and RANDSum that use only the local information within the specified document. It shows the expanded document context from like-minded users can benefit the sentence's evaluation process by proving more external document information related to the specified document.

To discover how the relative contribution from user context and document context influences the summarization performance, we set the weight adjusting parameter η from 0.2 to 0.9, and Figure 3 shows the ROUGE-1 evaluation results of the proposed approach with different η value.
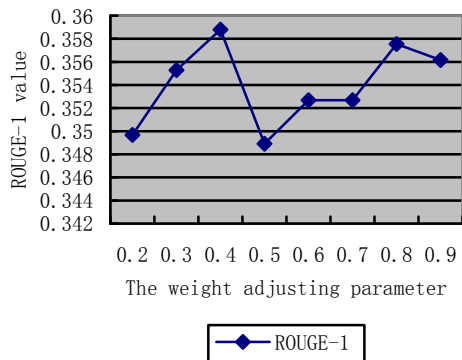

Figure 3. The ROUGE-1 evaluation results of the proposed approach with different η value

Seen from Figure 3, it is clear that the summarization performance on ROUGE-1 first increases with η, when η is larger than 0.4, the performance tends to decrease. It shows that considering appropriate user context knowledge is critical for improving summarization performance.

The reason underlying the above observations that the proposed social context summarization approach can improve the performance of document summarization is that there are many different documents on the Internet to discuss the same topic from various perspectives, and the discovered appropriate social context would guarantee that the influences through the different documents and different user communities are reliable.

5   Conclusion and Future Work

This paper examines an important factor called social context and proposes a novel social summarization approach for incorporating both document context and user context for collaborative generation of summaries. Experimental results on the dataset sampled from del.icio.us demonstrate the effectiveness of our method and show the superior performance over several baselines.

In future work, it would be interesting to investigate the performance of the approach on larger data sets with richer social annotation information. Besides, we will explore the optimization-based estimation strategy to automatically determine the parameters of our approach in an adaptive way. We also plan to make use of more implicit or explicit user feedback information, meta-content information, and hyperlink information to acquire richer social context knowledge to improve the summarization performance.

References

Ani Nenkova and Rebecca J. Passonneau. 2004. *Evaluating content selection in summarization: the pyramid method. Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '04)*, 145-152.

Caimei Lu, Xin Chen, and Park E. K.. 2009. *Exploit the tripartite network of social tagging for web clustering. Proceeding of the 18th ACM conference on Information and knowledge management (CIKM '09)*, 1545-1548.

Chin-Yew Lin and Hovy Eduard. 2003. *Automatic evaluation of summaries using n-gram co-occurrence statistics. Proceedings of the 2003*

Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03).

Gunes Erkan and Dragomir R. Radev. 2004. *Lex-PageRank: prestige in multi-document text summarization. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP' 04).*

Hongyuan Zha. 2002. *Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02)*, 113-120.

Jaime G. Carbonell, Jade Goldstein. 1998. *The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98)*, 335-336.

Jiantao Sun, Dou Shen, Huajun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. *Web-page summarization using clickthrough data. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05) *, 194–201.

John M. Conroy and Dianne P. O'leary. 2001. *Text summarization via hidden Markov models. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*, 406-407.

Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. 2009. *Automatic generic document summarization based on non-negative matrix factorization. Information Processing & Management*, 45(1): 20-34.

Kathleen McKeown and Dragomir R. Radev. 1995. *Generating summaries of multiple news articles. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '95)*, 74–82.

Oisin Boydell and Barry Smyth. 2007. *From social bookmarking to social summarization: An experiment in community-based summary generation. Proceedings of the 12th international conference on Intelligent user interfaces (IUI '07)*, 42–51.

Ouyanga You, Wenjie Li, Sujian Li, and Qin Lu. 2011. *Applying regression models to query-focused multi-document summarization. Information Processing & Management*,47(2): 227-237.

Rada Mihalcea, Paul Tarau. 2004. *TextRank: bringing order into texts. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP' 04).*

Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. *Summarization from medical documents: a survey. Artificial Intelligence in Medicine*, 33(2): 157–177.

Tadashi Nomoto and Yuji Matsumoto. 2001. *A new approach to unsupervised text summarization. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*, 26-34.

Vahed Qazvinian and Dragomir R. Radev. 2010. *Identifying non-explicit citing sentences for citation-based summarization. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, 555-564.

Xiaojun Wan. 2008. *Using only cross-document relationships for both generic and topic-focused multi-document summarizations. Information Retrieval*, 11(1): 25-49.

Xiaojun Wan and Jianwu Yang. 2007. *CollabSum: Exploiting multiple document clustering for collaborative single document summarizations. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07)*, 143-150.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. *Manifold-ranking based topic-focused multi-document summarization. Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI'07)*, 2903–2908.

Xiaojun Wan and Jianwu Yang. 2007. *Single document summarization with document expansion. Proceedings of the 22nd national conference on Artificial intelligence (AAAI'07)*, 931-936.