Context-Enhanced Personalized Social Summarization

Po Hu[1,2], Donghong Ji[1], Chong Teng[1] and Yujing Guo[1]
(1) Computer School, Wuhan University, China
(2) Computer School, Central China Normal University, China
phu@mail.ccnu.edu.cn, donghong_ji2000@yahoo.com.cn,
tchong616@126.com, yujingguo.ximo@gmail.com

ABSTRACT

This work investigates an interesting and challenging task in summarization, i.e., personalized social summarization, which aims to adapt summarization result of a specified document to an intended user based on his interests inferred from social context implicitly. Most existing summarization systems generate a uniform version of summary for different users no matter who is reading or generate personalized summaries employing only the local information in the document and the user profile. This paper proposes a novel unsupervised approach by making use of enhanced social context to aid personalized summary generation. In the proposed method, document expansion, user expansion, and implicit induction of the intended user's interest aspects are achieved simultaneously by adopting a fuzzy tripartite clustering algorithm. And both the informativeness of sentences and the user’s interest aspects are incorporated in a unified ranking process. Preliminary experimental results on a social tagging dataset validate the effectiveness of the proposed approach.

KEYWORDS: Personalized social summarization, social context, fuzzy tripartite clustering

1 Introduction

With the dramatic growth of the Internet, people are overwhelmed by a large number of accessible documents. In recent years, document summarization has become one of the most important research topics, which aims to address such dilemma by automatically capturing the essential content from document(s) and presenting it to a human reader in a succinct and friendly form. However, most existing summarization methods generate the same summary for different users, regardless of the interests of the readers for whom they are intended. These "one size fits all" methods may perform well in general but may not meet the needs of individuals.

Now with the rapid growth of social networking services like Delicious[1], CiteULike[2], and Flickr[3], users are no longer passive consumers of web contents. They can create contents and add metadata. Similarly, web documents no longer exist on their own and they are naturally associated with other documents and diverse users. All these information can be considered as the potential data source for document understanding and personalization.

For generating a personalized summary, traditional methods usually require that a user explicitly provides his interest aspects, such as specifying the categories he prefers (Díaz and Gervás, 2007) or clicking a subset of sentences in a document according to his interests (Yan et al., 2011). However, most users are reluctant to provide such information, thus it is more meaningful to infer a user's interests implicitly.

To address these concerns, we present an unsupervised approach for personalized summarization. The underlying assumption is that it is beneficial to understand both a single document and a single user better if appropriate social context can be leveraged under some constraints. In this work, the expanded social context used to infer users' interests and enrich document's content is highly selective, which comes from the most similar users and documents. We explored how the size of social context influences the summarization performance, and further demonstrated that appropriate contextual information can ensure better quality and personalization of summaries.

To the best of our knowledge, implicitly exploiting social contextual information to collaboratively summarize single document in a personalized way has been rarely investigated in the summarization community. In this work, we propose a novel personalized summarization approach which benefits from three important elements: the interests of like-minded users, the contents of topic-related documents, and semantically-related tags. In the approach, a fuzzy tripartite clustering algorithm is proposed and a multi-manifold ranking algorithm is adopted to generate personalized summary by considering both the informativeness of sentences and the intended user's interests.

The main contribution of this paper is summarized as follows:

1. we investigate an interesting and challenging summarization task, i.e., personalized social summarization.
2. we propose a novel approach making use of expanded social context to capture the intended user's interests, enrich the target document's content, and collaboratively summarize the document in a personalized way.

[1] http://delicious.com/
[2] http://www.citeulike.org/
[3] http://www.flickr.com/

3. we conduct preliminary experiments to validate the effectiveness of the proposed approach on a social tagging dataset and investigate how the expanded social context improves the performance of personalized summarization.

The remainder of the paper is organized as follows. The related work is introduced in Section 2. The proposed summarization approach is described in Section 3. Experimental results are shown in Section 4. Section 5 is our conclusion and future work.

2 Related work

Document summarization has been widely studied for many years. To date, various approaches have been proposed, and our work is under the framework of extractive summarization.

The vast majority of extractive methods identify which sentences are important by making use of unsupervised or supervised learning techniques. In unsupervised methods, feature-based ranking methods are usually based on a combination of linguistic and statistical features such as term frequency, sentence position, cue words, stigma words, lexical chains, rhetorical structure, topic signatures (Luhn, 1969; Lin and Hovy, 2000), etc. Clustering-based methods usually select one or more representative sentences from each subtopic to produce a summary with minimized redundancy and maximized coverage (Nomoto and Matsumoto, 2001). Graph-based methods have been shown to work well and are becoming more and more popular. LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) are representative methods adopting models like PageRank and HITS to estimate the importance of sentences via the computation of the stationary distribution of a Markov chain or a mutual reinforcement process (Zha, 2002).

For supervised methods, summarization is often regarded as a classification task or a sequence labeling task at sentence level, and many supervised learning algorithms have been investigated including Hidden Markov Models (Conroy and O'leary, 2001), Support Vector Regression (You et al., 2011), Factor Graph Model (Yang et al., 2011), etc. However, such a supervised learning paradigm often requires a large amount of labeled data, which are not available in most cases.

With the rapid growth of online information, some work has began to employ context to aid summarization, such as contents from external documents (Wan and Yang, 2007) or cited papers (Mei and Zhai, 2008; Qazvinian and Radev, 2010), click-through data or search logs (Sun et al., 2005), and social tags (Qu and Chen, 2009; Hu et al., 2011), comments (Hu et al., 2008) or discussing tweets (Yang et al., 2011), etc.

However, such methods so far are usually designed for generic summarization and do not take into account the impact of users' interests on summary generation. Besides, in the existing studies, personalized summarization is often conducted with the help of a query (Sun, 2008; You et al., 2011) or a static user profile (Díaz and Gervás, 2007), and most studies only use the local content from target document(s) or the user profile, with little attention paid to the rich social contextual information affiliated with them.

Currently, an increasing number of social websites allow users to enrich the source content. Many documents are now presented together with various feedback information in the form of social tags, comments, or ratings, etc. These usage data can be exploited for personalized summarization since they provide a natural channel to reveal users' interests implicitly.

Based on the analysis above, we investigate a challenging task in summarization, i.e., personalized social summarization, and propose an unsupervised approach for this task. The characteristic of our proposed approach is that it can leverage topic-related documents, like-minded users, and semantically-related tags to infer the intended user's interests implicitly and collaboratively summarize the target document in a personalized context-aware way.

3 Personalized social summarization

3.1 Overview

Given a user u (u∈U), a document d (d∈D), and related social tagging data G (G = (D, U, T, R)), personalized social summarization aims to generate a tailored summary of d for u. Here D, U, and T are documents, users, and tags respectively. R is a ternary relation between them, which denotes the set of annotations of each tag in T to a document in D by a user in U.

In most social tagging sites, many documents have been annotated by few tags and most users have only annotated few documents. In this case, existing tag-based summarization methods will fail to produce a personalized summary (Boydell and Smyth, 2007; Zhu et al., 2009), since the user-related tags may be absent for that document. To address it, we propose to expand both the target document and the intended user with appropriate social context so that the important parts in the document that the intended user may care about can be identified from context.

The general framework of our proposed approach consists of three major steps.

Step1. Social context identification by document expansion and user expansion

In this step, the given document d is expanded to a small document set $D_d^{(c)}$ by adding a small number of topic-related documents, and the intended user u is expanded to a small user community $U_u^{(c)}$ by adding a small number of like-minded users. Here $D_d^{(c)}$ and $U_u^{(c)}$ are identified as the expanded social context, which is based on the intuition that we would better know a user if we know more like-minded users close to him and we would better understand a document if we read more topic-related documents close to it.

Step2. User interest discovery

In this step, the interest aspects of the intended user u are inferred from the social context $D_d^{(c)}$ and $U_u^{(c)}$ by making use of the social tagging information that the like-minded users gave to the topically related documents.

Step3. Personalized summary generation

In this step, given the expanded document context $D_d^{(c)}$ and the inferred interest aspects, the relationships of all sentences in $D_d^{(c)}$ against each interest aspect are incorporated in a unified ranking process to extract personalized informative sentences from document d.

3.2 Social context identification

In this study, the related social tagging data G, which the target document d and the intended user u belong to, is firstly collected. Then it is used to identify the social context, which can be demonstrated by the example in Figure 1.

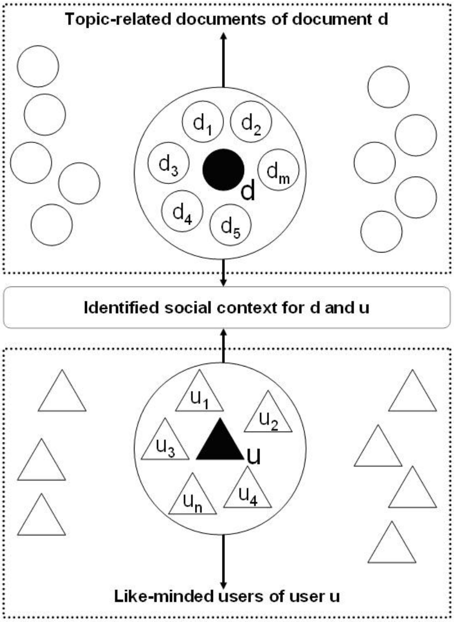
Topic-related documents of document d
d1
d2
d3
d
dm
d4
d5
Identified social context for d and u
u1
u2
u3
u
u4
un
Like-minded users of user u

FIGURE 1 – Social context for document d and user u.

Since content-related documents are usually annotated with semantically-related tags by users with similar interests, it is feasible to find topic-related documents, like-minded users, and semantically-related tags simultaneously by clustering them collaboratively (Lu et al., 2009). Therefore, we propose a fuzzy tripartite clustering algorithm to solve the fuzzy partition issues peculiar in personalized social summarization: a document may cover different subtopics, a user may have diverse interest aspects, and a tag may be a polysemy. The potential benefit of our algorithm is that it can make use of the inherent cluster structure and interactions among the different types of objects to cluster them simultaneously and flexibly. By the algorithm, an object can have a fuzzy membership across clusters and each cluster can be represented by a committee, i.e., a small number of objects with the highest membership for the cluster.

Before clustering, each type of object (e.g., document, user, and tag) is first represented by a combined vector. A document d_i is represented by D_i consisting of two components with one denoting user link vector and the other denoting tag link vector. $D_i = (D_i^{(U)}, D_i^{(T)})$, $D_i^{(U)} = (x_{ij}^{(U)} \mid j=1,2,...,|U|)$, $D_i^{(T)} = (x_{ik}^{(T)} \mid k=1,2,...,|T|)$, where $x_{ij}^{(U)}$ denotes the times that d_i is annotated by user u_j, $|U|$ denotes the total number of users, $x_{ik}^{(T)}$ denotes the times that d_i has been annotated with tag t_k, and $|T|$ denotes the total number of tags. User and tag can be represented likewise. Accordingly, the similarity between any two objects of the same type can then be computed by the linear combination of the similarity between their combined vectors. Our proposed fuzzy tripartite clustering algorithm is shown as follows.

Algorithm 1: The fuzzy tripartite clustering algorithm.

```
Input:
    G=(D, U, T, R ): the related social tagging data that document d and user u belong to;
    $N_{dc}$, $N_{uc}$, $N_{tc}$: the predefined number of document clusters, user clusters, and tag clusters.
Output:
    The fuzzy cluster assignments of documents, users, and tags: $M_d^*$, $M_u^*$, and $M_t^*$, where each
    object is affiliated with a list of membership values with respect to various clusters.
Method:
    Initialize the fuzzy partition matrices of documents, users, and tags $M_d^{(0)} = \left[u_{(d)_{i,j}}\right]_{|D|\times N_{dc}}$,
    $M_u^{(0)} = \left[u_{(u)_{i,j}}\right]_{|U|\times N_{uc}}$ and $M_t^{(0)} = \left[u_{(t)_{i,j}}\right]_{|T|\times N_{tc}}$ randomly, such that $0 \leqslant u_{(d)_{i,j}}$, $u_{(u)_{i,j}}$, $u_{(t)_{i,j}} \leqslant 1$ and
    $\sum_p u_{(d)_{i,p}} = 1$, $\sum_p u_{(u)_{i,p}} = 1$, $\sum_p u_{(t)_{i,p}} = 1$. And then generate the initial committee of each cluster
    and set k = 1.
    Repeat:
        For each type of object (e.g. document, user, and tag) do
            Calculate the centroid vector $c^{(k)}$ of each cluster based on the current committee of this
            cluster according to formula (1).
            For each object do
                Update the object's membership values $u_{i,j}^{(k)}$ to $u_{i,j}^{(k+1)}$ by the normalized Cosine
                similarity value between the i-th object and the centroid of the j-th fuzzy object cluster
                formed in the k-th iteration. Here the computation of the similarity value can be
                considered as the membership function.
            End For
            Regenerate the committee of each cluster.
        End For
        k = k + 1
    Until $\max_{i,j}\{|u_{i,j}^{(k+1)} - u_{i,j}^{(k)}|\} < \varepsilon$ or k>specified threshold.
```

In the algorithm, $u_{i,j}$ denotes the membership value for the i-th object in the j-th cluster, ε is the termination criterion, which is set as 0.01 in this study. The threshold of maximum iteration number is set at k=50. Since the 'true' numbers of document clusters, user clusters, and tag clusters are hard to predict in advance, we simply set N_{dc}, N_{uc}, and N_{tc} to the square root of the total number of documents, users, and tags in the related social tagging data respectively. The committee of each cluster is determined by selecting 30 percent of objects which have the highest membership values for the cluster from all the objects of the same type. In the following demonstration, we will take documents as examples of objects.

Let C_d represent a fuzzy document cluster and $C_d(c)$ represent the committee of C_d ($C_d(c) \subset C_d$). Since each document can be represented by a user link vector and a tag link vector, we will first consider the user link vectors of these documents. The value of the centroid vector of the document cluster C_d at the user dimension u_μ can be calculated by formula (1).

$$Centroid_{C_{d_u}}^{(U)} = \frac{\sum_{d_i \in C_d(c), u_j \in C_u(c)} x_{ij}^{(U)}}{|C_d(c)| * |C_u(c)|}, \ (u_u \in C_u(c)) \qquad (1)$$

where $C_u(c)$ is the committee of a fuzzy user cluster for which user u_u has its highest membership value, u_j is any user in $C_u(c)$, and d_i is any document in $C_d(c)$. $x_{ij}^{(U)}$ denotes the times that d_i is tagged by u_j. The value of the centroid vector at the tag dimension can be calculated similarly. Accordingly, the similarity between a document d_i and the centroid of a fuzzy document cluster C_j can be calculated by the linear combination of the Cosine similarity between their user link vectors and tag link vectors.

After clustering, we get the cluster assignments of documents, users, and tags, where each document (user, tag) gets a membership value for each cluster. Next, the given document d is expanded to the document context $D_d^{(c)} = \{d, d_1, d_2, ..., d_m\}$ by adding m topic-related documents with highest membership value for the cluster that d belongs to most likely. Similarly, the intended user u is expanded to the user context $U_u^{(c)} = \{u, u_1, u_2, ..., u_n\}$ by adding n like-minded users with highest membership value for the cluster that u belongs to most likely. $D_d^{(c)}$ and $U_u^{(c)}$ are identified as the expanded social context, aiming to boost information shared by topic-related documents and users with similar interests for personalized summary generation. We will further discuss the variation of performance with different assignment of m and n in Section 4.

3.3 User interest discovery

As a common form of users' online behavior, users' social tagging activities are good at reflecting their interests about document's contents and expressing the general concepts of documents. Previous work has studied the utility of social tags for user interest modeling (Li et al., 2008) and confirmed that a set of semantically related tags can characterize users' interests well (Zhou et al., 2010).

Considering that a user may have diverse interest aspects on a given document and the combination of topic-related documents and like-minded users can provide rich global contextual clues, we propose to model the interests of a user u about a document d by the social tags which have been used to annotate the documents in the document context $D_d^{(c)}$ by the users from the user context $U_u^{(c)}$. The intuitive idea is that users who annotate similar documents may have common interests on the topic shared by these documents, so the tags used by these like-minded users may reveal the latent interests of the intended user about this kind of topic.

According to the output of the fuzzy tripartite clustering algorithm, the tags on $D_d^{(c)}$ annotated by $U_u^{(c)}$ may belong to different tag clusters with varying degrees of membership. So we assign these tags into the clusters for which they have highest membership values, and then we can model the intended user's interests by the tag clusters with each indicating one unique interest aspect of the user. Here each cluster consists of one or more semantically-related tags, corresponding to the committee of the relevant tag cluster.

Formally, the intended user's interests on the given document can be represented as $UM_{u,}$ which can be regarded as multiple subtopics for modelling user's interest aspects. $UM_u = \{p_i \mid 1 \leq i \leq N_{tu}\}$, where N_{tu} is the number of interest aspects for user u, and p_i is the user's i-th interest aspect for the given document.

3.4 Personalized summary generation

Based on the identified interest aspects, we further adopt multi-manifold ranking algorithm to fuse the sentence relationships against different aspects in a unified ranking process, which has performed successfully in the multi-subtopic summarization task (Wan, 2009). In this study, we collaboratively summarize the target document by multiple topic-related documents within the document context, since topic-related documents can provide more clues from global context to aid extracting salient summary sentences from the specified document.

Formally, given the sentence set $S=\{s_i \mid 1\leq i\leq n\}$ of the document context $D_d^{(c)}$ for document d and the k-th interest aspect p_k of user u, an affinity matrix $W_k=\left[w_{(k)_{i,j}}\right]_{(n+1)\times(n+1)}$ can be built firstly to represent both the relationships among all the n sentences in $D_d^{(c)}$ and the relationship between each sentence and p_k. Then W_k is symmetrically normalized by $S_k=D_k^{-1/2}\cdot W_k\cdot D_k^{-1/2}$. Here $w_{(k)_{i,j}}$ is computed by the Cosine similarity between the i-th sentence and the j-th sentence. D_k is the diagonal matrix with the (i,i)-element equal to the sum of the i-th row of W_k. In this study, there will be N_{tu} affinity matrices in total since the number of discovered interest aspects for user u is N_{tu}.

Let F represent a ranking function that assigns each element s_i ($0\leq i\leq n$) a ranking score f_i. It can be regarded as a vector $F=[f_0,\ldots,f_n]^T$. We also define a prior vector $Y=[y_0,\ldots,y_n]^T$, in which $y_0=1$ for the k-th interest aspect p_k and $y_i=0$ ($1\leq i\leq n$) for all the remaining sentences.

Next, we can rank all the sentences by adopting the multi-manifold ranking algorithm (Wan, 2009), in which the ranking function F is to be learned from W_k ($1\leq k\leq N_{tu}$) and Y. In this study, the constraints from S_k ($1\leq k\leq N_{tu}$) and Y are naturally fused in a regularized optimization framework defined by the following cost function.

$$Q(F)=\sum_{k=1}^{N_{tu}}\left[u_k\cdot\sum_{i,j=0}^{N_{tu}}\left(w_{(k)}\right)_{ij}\left|\frac{1}{\sqrt{(D_k)_{ii}}}f_i-\frac{1}{\sqrt{(D_k)_{jj}}}f_j\right|^2\right]+\eta\cdot\sum_{i=0}^{n}\left|f_i-y_i\right|^2 \quad (2)$$

where u_k ($1\leq k\leq N_{tu}$) and η ($0<\eta\leq 1$) are the trade-off between the smoothness constrains. $0<u_k$, $\eta<1$ and $\sum_{k=1}^{N_{tu}}u_k+\eta=1$.

Based on the optimization framework, the optimal ranking function F^* can be achieved when Q(F) is minimized. In practice, the following iterative form shown in the formula (3) is more commonly used to get the ranking function, in which $F^{(0)}$ is set to Y and we have $F^{(*)}=\lim_{t\to\infty}F^{(t)}$.

$$F^{(t+1)}=\sum_{k=1}^{N_{tu}}u_kS_kF^{(t)}+(1-\sum_{k=1}^{N_{tu}}u_k)Y \quad (3)$$

Through the above ranking process, the ranking scores, which denote the user-biased informativeness of sentences, can be obtained. Finally, those sentences highly overlapping with other informative sentences are penalized to remove redundancy (Wan and Yang, 2007), and the sentences with high overall scores are chosen from document d into the summary.

4 Experiments

4.1 Dataset

Since there is no benchmark dataset available for the task of personalized social summarization, we collected data from Delicious, one of the most popular social tagging websites. Specifically, we extracted a set of web documents, bookmark tags, and the users who bookmarked these documents to serve as the experimental dataset.

Starting with predefined seed tags, we extracted the top bookmarked documents for each tag and extracted the users and tags used to annotate each of the documents. The result is a collection consisting of 204 bookmarked documents and 2186 unique social tags that were used to annotate these documents by 1696 users. To guarantee the genre consistency, all the documents were crawled from news sources such as CNN, BBC, New York Times, etc.

4.2 Evaluation methods

In this paper, both manual evaluation method and automatic evaluation method are adopted. For each document in the dataset, we randomly select one to five users as the intended users from all the users who annotated the document with multiple social tags.

4.2.1 Manual evaluation

First, we must admit that it would be better to use personalized reference summaries for evaluation. However, it would be quite difficult to get personalized summaries from the actual users of Delicious. The alternative way is to get the external judgments from several judges and take average of their ratings so that we know that multiple people would consider that this summary is relevant or tailored for the intended user to a certain extent. How to develop a better test collection for personalized summarization from the perspective of social context is an important future direction of our research.

In this study, three evaluators are requested to express their judgments over all automatically generated summaries based on both the content they deem to be important for the target document and how "personal" each one is according to the interests of the intended user. We provide each evaluator the intended user’s background knowledge collected by calling the official Delicious.com API and parsing its RSS feeds. The provided information includes all the open document bookmarks of intended users and all the tags they used to annotate the documents including the target document to be summarized. Evaluators can also access the content of the corresponding document by clicking the URL in each bookmark.

In the evaluation process, evaluators are instructed to give an overall score to each summary. The overall score reflects the comprehensive quality of a summary including not only the evaluation for the general content of the generated summary but also the degree of compliance with the intended user’s personalized interests and foci.

All the judgment scores are rated in a 5-point scale, where “1” for “very poor”, “2” for “poor”, “3” for “barely acceptable”, “4” for “good”, and “5” for “very good”. Evaluators are allowed to judge at any scores between 1 and 5, e.g. 3.5.

4.2.2 Automatic evaluation

Considering that manual evaluation is generally time consuming and labour-intensive, we also adopt automatic evaluation strategy.

For each intended user of the target document, we randomly divide the social tags he assigned to the document into two approximately equal parts: a training set and a test set. The former is used to generate personalized social summary on the document for the user, and the latter is used to evaluate the generated summary based on the recall against the tags in the test set, making sure to remove the tags occurring in the training set from the test set.

The idea of this kind of evaluation strategy is to look for overlaps between the generated personalized social summary and those unseen tags used by the intended user on the given document, since the tags in the test set correspond to an alternative, but previously unseen, point of interests for the intended user with respect to the target document.

The automatic evaluation experiments were conducted in the cross validation procedure, and the average recall score was recorded. Intuitively, the higher the average recall score is, the more the generated summaries are in line with the interests of the intended users.

4.3 Baselines

In the experiments, we compare our proposed approach with several baseline methods. For fair comparison, we conduct the same preprocessing for all the methods including sentence segmentation, word stemming, and redundancy removing.

Random: It extracts sentences randomly from each document.

OTS: It is an open source summarizer integrating shallow NLP techniques with statistical word frequency analysis for sentence scoring (Nadav, 2003).

MEAD: It ranks sentences according to the combination of features including centroid value, positional value, and first-sentence overlap (Radev et al., 2000).

LexRank: It first constructs a sentence affinity graph based on the Cosine similarity between sentences in a document, and then extracts a few informative sentences based on eigenvector centrality (Erkan and Radev, 2004).

DcontextLexRank: It is an extension of the original LexRank method by firstly ranking sentences on the document context which the target document belongs to, and then extracting sentences with highest ranking scores from the target document.

PSocialSum: It is our proposed approach using expanded social context to capture the intended user's interests, enrich the target document's content, and collaboratively summarize the target document in a personalized way.

4.4 Overall comparison results

4.4.1 Parameter settings

The parameters m and n, i.e., the number of expanded topic-related documents in the document context $D_d^{(c)}$ and the number of expanded like-minded users in the user context $U_u^{(c)}$, are set as the number of elements in the corresponding committee of the cluster that document d or user u belongs to most likely.

In the multi-manifold ranking process of our approach, parameters u_k and η are the smoothness constraint and fitting constraint respectively, which control the trade-off between the impact from S_k (i.e., both the relationships among all the sentences in the document context and the relationship between each sentence and the k-th interest aspect of user u) and the impact from Y (i.e. the prior vector set for the k-th interest aspect and all the remaining sentences). In the following experiments, the regularization parameter η for the fitting constraint is fixed at 0.01, the same as in (Wan, 2009), and u_k is set to the normalized Cosine similarity between the corresponding vectors of p_k and $D_d^{(c)}$.

4.4.2 Experimental results

In the experiments, for each document, we generate multiple different personalized summaries for each of the intended users by our approach. For comparison purpose, each document in the dataset is also summarized using all the baseline methods described in Section 4.3.

First, we conducted the manual evaluation and the average overall scores of multiple evaluators on all the generated summaries are listed in Table 1.

Method	Average Overall Score
Random	1.2
OTS	2.1
MEAD	2.2
LexRank	2.3
DcontextLexRank	2.4
PSocialSum	3.5

TABLE 1 –The average overall scores of multiple evaluators.

From Table 1, it can be found that Random has the worst summarization performance.

LexRank and DcontextLexRank perform better than those of MEAD and OTS. This is mainly because both LexRank and DcontextLexRank make use of the inter-relationship between sentences to rank them globally, while MEAD and OTS only depend on the combination of some local features.

DcontextLexRank outperforms LexRank in our experiments, which indicates the use of appropriate document context for sentence ranking is an improvement over the use of single document alone which lacks the support of external clues from the similar documents.

Note that all these baseline methods generate the summary based on either the given document itself or the document context, regardless of the intended user's interests. Our proposed approach shows significantly better performance on evaluators' ratings. And the rating difference between PSocialSum and other baselines is significant at the 95% statistical confidence level in all cases. This indicates that consideration of user's interests is critical for generating a better personalized summary, and the improvement achieved is mainly attributed to the personalization aspect as well as informative content.

We also find that the evaluator judgments on MEAD, LexRank, and DcontextLexRank are of little significant difference at the 95% confidence interval, which illustrates that the general summaries generated by these comparable baselines can convey the important information of a document, and different evaluators may have some agreement on the quality of its content, although all of these methods do not consider the intended user's interest at all.

Next, we conducted the automatic evaluation by computing the average recall scores against the tags in the corresponding test set for all the resulting summaries. The process is repeated across multiple different random splits of training and test set. The average recall scores are reported in Table 2.

Method	**Average Recall Score**
Random	0.194
OTS	0.282
MEAD	0.287
LexRank	0.292
DcontextLexRank	0.294
PSocialSum	0.338

TABLE 2 – The average recall scores.

From Table 2, we see that the summarization performance of PSocialSum is consistently better than those of other baselines. Such results also demonstrate that by leveraging part of the social tagging information of the intended users, we can generate better summaries which are more in accordance with the latent interests of them, compared to other summarizers which generate the static summaries ignoring the social contextual information.

4.5 Impact of parameters

In this section, to investigate how the size m and n of the expanded topic-related documents and the expanded like-minded users influence the performance of PSocialSum, we conduct the following experiments with different values.

Considering that m and n in this study are dynamically related to the predefined percentage of objects which have the highest membership values for the cluster from all the objects of the same type, in the experiment, we set the predefined percentage value related with m and n ranging from 10% to 80% with step length 10%, indicating the corresponding percentage of documents or users are selected for the expanded document set or user set.

Figure 2 shows the average recall scores against the tags in the test set for PSocialSum with different percentage values.

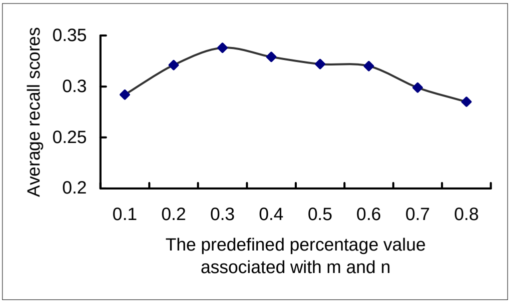

FIGURE 2 –The average recall scores of PSocialSum
vs. the predefined percentage value related with m and n.

From Figure 2, it can be seen that when the percentage value increases from 10% to 30%, the recall increases gradually, and reaches the global maximum when it is set to 30%. When we adjust the percentage value from 30% to 80%, the recall starts to decay. The result demonstrates that appropriate document context and user context are beneficial for improving personalized summarization performance, yet a large size of the expanded context may deteriorate the performance because it may include a lot of irrelevant information even noise.

Conclusion

In this paper, we present a study of personalized social summarization, and propose a novel unsupervised approach. The approach makes use of expanded social context to capture the intended user's interests, enrich the target document's content, and collaboratively summarize the target document in a personalized context-aware way. Preliminary experimental results demonstrate the effectiveness of the proposed approach.

In practice, the dimensions and variability of users, documents, and tags from most social network websites may be quite high, so in future work, we plan to combine link structure association analysis and feature selection to effectively deal with high-dimensional online tripartite clustering dynamically. And more social contextual information such as social relationships among users will also be investigated. For simplicity, this method represents each object with a vector of two sets of features, and this kind of representation would inevitably result in information loss to a certain extent. Therefore, we plan to try better alternatives such as hypergraph or tensor model, and make effort to improve the existing work on content or social network-based user interest modeling. Furthermore, it would be more convincing to resort to crowdsourcing technique to evaluate the proposed approach by a large number of real users on the social tagging websites and on larger-scale social data set.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61133012, 61173062, 61070082, 61070083, 61070243), the major program of the National Social Science Foundation of China (No. 11&ZD189), and the Post-70s Scholars Academic Development Program of Wuhan University.

References

Boydell, O. and Barry, S. (2007). From social bookmarking to social summarization: an experiment in community-based summary generation. In *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI 2007)*, pages 42-51, ACM, New York, NY.

Conroy, J.M. and Oleary, D.P. (2001). Text summarization via hidden markov models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 406-407, ACM, New York, NY.

Díaz, A. and Gervás, P. (2007). User-model based personalized summarization. *Information Processing and Management*, 43(6):1715-1734.

Erkan, G. and Radev, D.R. (2004). LexPageRank: prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Hu, M.S., Sun, A.X., and Lim, E.P. (2008). Comments-oriented document summarization: understanding documents with users' feedback. In *Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 291-298, ACM, New York, NY.

Hu, P., Sun, C., Wu, L.F., Ji, D.H., and Teng, C. (2011). Social summarization via automatically discovered social context. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 483-490.

Li, X., Guo, L., and Zhao, Y.H. (2008). Tag-based social interest discovery. In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pages 675-684, ACM, New York, NY.

Lin, C.Y. and Eduard, H. (2000). The automated acquisition of topic signatures for text summarization. In *Proceedings of the 17th Conference on Computational Linguistics (COLING 2000)*, pages 495-501, Association for Computational Linguistics, Stroudsburg, PA.

Lin, C.Y. and Eduard, H. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, pages 71-78, Association for Computational Linguistics, Stroudsburg, PA.

Lu, C.M., Chen, X., and Park, E. K. (2009). Exploit the tripartite network of social tagging for web clustering. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 1545-1548, ACM, New York, NY.

Luhn, H. P. (1969). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159-165.

Mei, Q.Z. and Zhai, C.X. (2008). Generating impact-based summaries for scientific literature.

In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2008)*, pages 816–824.

Mihalcea, R. and Tarau, P. (2004). TextRank: bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Nadav, R. (2003). The open text summarizer. http://libots.sourceforge.net/.

Nomoto, T. and Matsumoto, Y. (2001). A new approach to unsupervised text summarization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 26-34, ACM, New York, NY.

Qazvinian, V. and Radev, D.R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 555-564, Association for Computational Linguistics, Stroudsburg, PA.

Qu, Y. and Chen, Q.X. (2009). Collaborative summarization: when collaborative filtering meets document summarization. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*, pages 474-483.

Sun, J.T., Shen, D., Zeng,H.J., Yang, Q., Lu, Y.C., and Chen, Z. (2005). Web-page summarization using clickthrough data. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, pages 194-201, ACM, New York, NY.

Sun, P. (2008). Personalized summarization agent using non-negative matrix factorization. In *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, pages 1034-1038.

Wan, X.J. and Yang, J.W. (2007). Single document summarization with document expansion. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI 2007)*, pages 931-936.

Wan, X.J. (2009). Topic analysis for topic-focused multi-document summarization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 1609-1612, ACM, New York, NY.

Yan, R., Nie, J.Y., and Li, X.M. (2011). Summarize what you are interested in: an optimization framework for interactive personalized summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1342–1351.

Yang, Z., Cai, K.K., Tang, J., Zhang, L., Su, Z., and Li, J.Z. (2011). Social context summarization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*, pages 255-264, ACM, New York, NY.

You, O.Y., Li, W.J., Li, S.J., and Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing and Management*, 47(2):227-237.

Zha, H.Y. (2002). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 113-

120, ACM, New York, NY.

Zhou, C., Ma, H., Lyu, M.R., and King, I. (2010). UserRec: a user recommendation framework in social tagging systems. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*, pages 1486-1491.

Zhu, J.Y., Wang, C., He, X.F., Bu, J.J., Chen, C., Shang, S.J., Qu, M.C., and Lu, G. (2009). Tag-oriented document summarization. In *Proceedings of the 18th International Conference on World Wide Web (WWW 2009)*, pages 1195-1196, ACM, New York, NY.