Iterative Feedback Based Manifold-Ranking for Update Summary

He Ruifang, Qin Bing, Liu Ting, Liu Yang, and Li Sheng

***Abstract*—The update summary as defined for the DUC2007 new task aims to capture evolving information of a single topic over time. It delivers *focused* information to a user who has already read a set of older documents covering the same topic. This paper presents a novel manifold-ranking frame based on iterative feedback mechanism to this summary task. The topic set is extended by using the summarization of previous timeslices and the first sentences of documents in current timeslice. Iterative feedback mechanism is applied to model the dynamically evolving characteristic and represent the relay propagation of information in temporally evolving data. Modified manifold-ranking process also can naturally make use of both the relationships among all the sentences in the documents and relationships between the topic and the sentences. The ranking score for each sentence obtained in the manifold-ranking process denotes the importance of sentence biased towards topic, and then the greedy algorithm is employed to rerank the sentences for removing the redundant information. The summary is produced by choosing the sentences with high ranking score. Experiments on dataset of DUC2007 update task demonstrate the encouraging performance of the proposed approach.**

***Index Terms*—Temporal multi-document summarization, update summary, iterative feedback based manifold-ranking.**

I. Introduction

MULTI-DOCUMENT summarization is the process of automatically producing a summary delivering the main information content from a set of documents about an explicit or implicit topic, which has drawn much attention in recent years and exhibits the practicability in document management and search systems. For example, a number of news services, such as Google[1], NewsBlaster[2], and Sina News[3], have been developed to group news articles into news topics, and then produce a short summary for each news topic so as to facilitate users to browse the results and improve users' search experience. News portals usually provide concise headline news describing hot news topic each day and they also produce weekly news review to save user's time and improve service quality.

Temporal multi-document summarization (TMDS) is the natural extension of multi-document summarization, which captures evolving information of a single topic over time. It is assumed that a user has access to a stream of news stories that are on the same topic, but that the stream flows rapidly enough that no one has the time to look at every story. In this situation, a person would prefer to read the update information at a certain time interval under the assumption that the user has already read a number of previous documents. The update summary as defined for the DUC2007 new task just faces this goal, which is a kind of TMDS. For the DUC2007 update task, 100-word summaries has to be generated for three consecutive document subsets sorted by their publication dates, tracking the new development of a single topic through time.

The key problem of summarization is how to identify important content and remove redundant content. The common problem for summarization is that the information in different documents inevitably overlaps with each other, and therefore effective summarization methods are needed to contrast their similarities and differences. However, the above application scenarios, where the objects to be summarized face to some special topics and evolve with time, raise new challenges to traditional summarization algorithms. The first challenge for update summary task is that the information in the summary must be biased to the given topic, and the second is that the information in summary must contain the evolving content. So we need to take into account effectively this topic-biased and temporally evolving characteristics during the summarization process. Thus a good update summary must include information as much as possible, keeping information as novel as possible, and moreover, the information must be biased to the given topic.

In [23], an extractive approach based on manifold-ranking of sentences to topic-focused multi-document summarization by using the underlying manifold structure in data points is proposed without modeling the temporally evolving characteristic. Inspired by this, for the DUC2007 update task, we propose a new manifold-ranking frame based on iterative feedback mechanism, which has the temporally adaptive characteristic. We assume that the data points evolving over time have the long and narrow manifold structure. However,

Manuscript received May 10, 2008. Manuscript accepted for publication June 20, 2008.

He Ruifang, Qin Bing, Liu Ting, Liu Yang and Li Sheng are all with the Information Retrieval Lab, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 15001, China (phone: +86-451-86413683-801; fax: +86-451-86413683-812; e-mail: rfhe@ir.hit.edu.cn).

[1] http://news.google.com
[2] http://www1.cs.columbia.edu/nlp/projects.cgi\#newsblaster
[3] http://news.sina.com.cn

the common topic for three consecutive document subsets is a static query, which cannot represent the dynamically evolving information. Therefore, we use the iterative feedback mechanism to extend the topic by using the summarization of previous timeslices and the first sentences of documents in current timeslice. We believe this topic extension can represent the relay propagation of information in temporally evolving data and improve the ranking score. The proposed approach employs iterative feedback based manifold-ranking process to compute the ranking score for each sentence that denotes the biased information richness of sentence. Then the sentences highly overlapping with other informative ones are penalized by the greedy algorithm. The summary is produced by choosing the sentences with highest overall scores, which are considered informative, novel and evolving. In this improved manifold-ranking algorithm, the intra-document and inter-document relationships between sentences are differentiated with different weights. Experiments on datasets of DUC2007 update task demonstrate the competitive performance of the proposed approach.

The rest of this paper is organized as follows: Section 2 introduces related work. The details of the proposed approach are described in Section 3. Section 4 presents and discusses the evaluation results. We conclude this paper and discuss future work in Section 5.

II. Related Work

In recent years, a series of workshops and conferences on automatic text summarization (e.g. DUC[4] and NTCIR[5]), special topic sessions in ACL, COLING, and SIGIR have advanced the technology and produced a couple of experimental online systems.

Update summary is a new challenge in the field of summarization. It aims to capture evolving information of a single topic over time, and has the characteristics of the topic-focused and temporal summary. It hopes to extract the new information over time, and also must be biased to a certain topic. Generally speaking, the summarization methods can be either extractive summarization or abstractive summarization. Extractive summarization assigns salience scores to some units (e.g. sentences, paragraphs) of the documents and extracts the sentences with highest scores, while abstractive summarization usually needs sentence compression and reformulation. In this paper, we focus on extractive summarization.

The centroid-based method [20] is one of the most popular extractive summarization methods. The clustering based method [3] is also widely used, including term, sentence and sub-topic clustering. Most recently, the graph-ranking based methods, including TextRank [17] and LexRank [6], have been proposed for document summarization. Similar to PageRank [4] or HITS [11], these methods first build a graph based on the similarity relationships between the sentences in documents and then the importance of a sentence is determined by taking into account the global information on the graph recursively, rather than relying only on the local sentence-specific information. The basic idea underlying the graph-based ranking algorithm is that of ``voting" or ``recommendation". When a sentence links to another one, it is basically casting a vote for the linked sentence. The higher the number of votes that are cast for a sentence, the more important the sentence is. Moreover, the importance of the sentence casting the vote determines how important the vote itself is. The computation of sentence importance is usually based on a recursive form, which can be transformed into the problem of solving the principal eigenvector of the transition matrix.

Most topic-focused document summarization methods incorporate the information of the given topic or query into generic summarizers and extract sentences suiting the user's declared information need [21], [8], [5], [9], [7]. Very recently, Wan *et al.* [23] proposed an approach based on manifold-ranking. Their method tried to make use of relationships among all the sentences in the documents and the relationships between the given topic and the sentences. The ranking score is obtained for each sentence in the manifold-ranking process based on graph to denote the biased information richness of the sentence. Then the greedy algorithm is employed to impose diversity penalty on each sentence. The sentences with high ranking score are then selected as the output summary. More related work can be found on DUC2003 and DUC2005 publications.

Temporal summary originates from text summarization and topic detection and tracking (TDT), and is also related to time line construction techniques. Alan *et al.* [1] firstly put forward the concept of temporal summary inspired by TDT in SIGIR2001. Given a sequence of news reports on certain topic, they extract useful and novel sentences to monitor the changes over time. Usefulness is captured by considering whether a sentence can be generated by a language model created from the sentences seen to date. Novelty is captured by comparing a sentence with prior sentences. They report that it is difficult to combine the two factors successfully. Other researchers exploit distribution of events and extract the hot topics on time line by statistical measures. Swan and Allan [22] employ χ^2 statistics to measure the strength that a term is associated with a specified date, and then extract and group important terms to generate ``topics" defined by TDT. In [12], Chen *et al*. import the aging theory to measure the ``hotness" of a topic by analyzing the temporal characteristic of news report. The aging theory implies that a news event can be considered as a life form that goes through a life cycle of birth, growth, decay, and death, reflecting its popularity over time. Then hot topics are selected according to energy function defined by aging theory. Lim *et al*.[14] anchor documents on time line by the publication dates, and then extract sentences from each document based on surface features. Sentence weight is adjusted by local high frequency words in each time slot and global high frequency words from all topic sentences. They evaluate the system on Korean documents and report that time can help to raise the

[4] http://duc.nist.gov
[5] http://research.nii.ac.jp/ntcir/index-en.html

percentage of model sentences contained in machine generated summaries. Jatowt and Ishizuka [10] investigate the approaches to monitor the trends of dynamic web documents, which mean different versions of the same web documents. They employ a simple regression analysis on word frequency and time to identify whether terms are popular and active. The importance of a term is measured by its slope, intercept and variance. The weight of a sentence is measured by the sum of the weights of the terms inside the sentence. The sentences with highest scores are extracted into a summary. However, they do not report any quantitative evaluation results. In [16], Mani is devoted to temporal information extraction, knowledge representation and reasoning, and try to apply them to multi-document summarization. In [13], Li *et al*. explore whether the temporal distribution information helps to enhance event-based summarization based on corpus of DUC2001.

In DUC2007, the top performing systems of update summary task adopted the extractive methods. LCC's GISTexter [2] used Machine Reading mechanism with textual inference information to create new and coherent information. Textual entailment and textual contradiction are recognized to construct representations of knowledge coded in a text collection. Update summary is produced by comparing the entailment and contradiction of sentences. This method preferably fused the deep linguistic knowledge, however, which is difficult to be reconstructed. IIIT Hyderabad's system [19] estimated a sentence prior by a term clustering approach, which incorporated the query independent score and query dependent score in a linear combination way. Sentence reduction and entity dereferencing is also used in the algorithm. NUS [26] proposed a timestamped graph model motivated by human writing and reading processes, which is used to model the dynamic and evolutionary characteristic of information. It assumed that writers write articles from the first sentence to the last, and readers read articles from the first sentence to the last. These two processes are similar to evolution of citation networks and the web. Though the parameters of this model are very complex, the method is an interesting attempt.

Due to different tasks, the above researches do not uniformly fuse the information in the topic and the documents or just incorporate the temporal characteristics. While iterative feedback based manifold-ranking approach to the DUC2007 new update summary task can naturally and simultaneously take into account topic information and the relay propagation of information in temporally evolving data.

III. Iterative Feedback based Manifold-Ranking Approach

The iterative feedback based manifold-ranking approach for update summary consists of three steps: (1) iterative feedback mechanism is used to extend the topic; (2) manifold-ranking score is computed for each sentence in the iterative feedback based manifold-ranking process; (3) based on the manifold-ranking scores, the diversity penalty is imposed on each sentence. Overall ranking score of each sentence is obtained to measure both the importance degree of the sentence relevant to the sentence collection and topic and the novelty degree of information contained in the sentence with respect to all sentences in the summary. The sentences with high overall ranking scores are chosen for the summary.

A. Basic Definitions

The manifold-ranking method [24], [25] is a universal ranking algorithm and it is initially used to rank data points along their underlying manifold structure. However, this method cannot model the temporally evolving characteristic, say, which is not temporally adaptively. For the DUC2007 update task, we assume that the data points evolving over time have the long and narrow manifold-structure. However, the common topic for three consecutive document subsets is a static query, which cannot represent the dynamically evolving information. Therefore, we apply the iterative feedback mechanism to extend the topic by using the summarization of previous timeslices and the first sentences of documents in the current timeslice.

Iterative Feedback mechanism: Given a set of timeslices $TS = \{timeslice_i \mid 1 \le i \le m\}$ and a topic $T = \{topic_i \mid 1 \le i \le m\}$, every $timeslice_i = \{d_j \mid 1 \le j \le n\}$ consists of documents, every document consists of sentences. Let s_{ij} denotes the first sentence of document d_j in $timeslice_i$, then first sentences of all documents in $timeslice_i$ $s_{first}(i) = \{s_{ij} \mid 1 \le i \le m, 1 \le j \le n\}$. The timeslices are ordered chronologically. Every timeslice corresponds to an update summary. When summarizing, the current $timeslice_i$ just can refer to the previous timeslices from 1 to *i*-1, but cannot refer to the ones from *i*+1 to m. Let $updateSum_i$ denotes the update summary of the current $timeslice_i$, and then $topic_i$ is extended as follows:

$$topic_i = \{PubTopic \cup \bigcup_{k=1}^{i-1} updateSum_k \cup s_{first}(i) \mid 1 \le i \le m\}$$

PubTopic denotes the public topic description of all timeslices.

We assume this topic extension can represent the relay propagation of information in temporally evolving data and help to capture the changes of a single topic over time.

B. Modified Manifold-Ranking Process

Given a query and a set of data points, the task of manifold-ranking is to rank the data points according to their relevance to the query [24]. The key to manifold-ranking is the prior assumption of consistency, which means: (1) nearby points are likely to have the same ranking scores; (2) points on the same structure (typically referred to as a cluster or a manifold) are likely to have the same ranking scores.

In our context, the data points are denoted by the topic description and all the sentences in the documents, where topic description dynamically evolves over time. The iterative feedback based manifold-ranking process in our context can be formalized as follows:

For $timeslice_i$, given a set of data points $X=\{x_1,...,x_t,x_{t+1},...,x_n\}\subset R^m$, the first t data points are the topic description and the rest data points are the sentences in the documents. According to the iterative feedback mechanism, x_1 denotes the *PubTopic*, $x_2 \ldots x_p$ denotes the and $x_{p+1} \ldots x_t$ denotes the $s_{first}(i)$. Note that because the *PubTopic* is usually short in our experiments, we treat it as a pseudo-sentence. Then it can be processed in the same way as other sentences. Let $f:X\to R$ denotes a ranking function which assigns to each point $x_q (1 \le q \le n)$ a ranking value f_q. We can view f as a vector $f=[f_1,..., f_n]^T$. We also define three vectors, $Y_1=[y_1,..., y_n]^T$, in which $y_1 = 1$ because x_1 is the *PubTopic* and $y_q = 0 (2 \le q \le n)$ for all the sentences in the documents; similarly, $Y_2=[y_1,..., y_n]^T$, in which $y_2 \ldots y_p = 1$ because $x_2 \ldots x_p$ denotes $\bigcup_{k=1}^{i-1} updateSum_k$ and $y_q = 0 (q = 1, p+1 \le q \le n)$; $Y_3=[y_1,..., y_n]^T$, in which $y_{p+1} \ldots y_t = 1$ because $x_{p+1} \ldots x_t = 1$ denotes the $s_{first}(i)$ and $y_q = 0 (1 \le q \le p,\ t+1 \le q \le n)$. The iterative feedback based manifold-ranking algorithm goes as follows:

In the first step of the algorithm, a connected network is formed. We remove the stop words in each sentence, and stem the remaining words. 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 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 the sentences containing term t. Then $sim(x_i, x_j)$ is computed according to the normalized inner product of the corresponding term vectors. The network is weighted in the second step and the weight is symmetrically normalized in the third step. The normalization in the third step is necessary to prove the algorithm's convergence. The fourth step is the key step of the algorithm, where all points spread their ranking score to their neighbors via the weighted network. The spread process is repeated until a global stable state is achieved, and we get the ranking score in the fifth step. The parameter α specifies the relative contributions to the ranking scores from neighbors and the initial ranking scores, and the parameter β, γ, η denotes the relative contribution to ranking scores from the *PubTopic*, the update summary in the previous timeslices and the first sentences of all documents in the current timeslice, respectively. Note that self-reinforcement is avoided since the diagonal elements of the affinity matrix are set to zero.

Algorithm 1. Iterative feedback based manifold-ranking

Input: $X = \{x_1,..., x_n\}$

Output: $f = \{f_i^* \mid i = 1...n\}$

1: Compute the pair-wise similarity values between sentences (data points) using the standard Cosine measure. Given two sentences x_i and x_j, the Cosine similarity is denoted as $sim(x_i, x_j)$, computed as the normalized inner product of the corresponding term vectors;

2: Connect any two points with an edge if their similarity value exceeds 0. We define the affinity matrix W by $W_{ij} = sim(x_i, x_j)$ if there is an edge linking x_i and x_j. Note that we let $W_{ii} = 0$ to avoid loops in the graph built in next step;

3: Normalize W by $S = D^{-1}W$ in which D is the diagonal matrix with (i,i)-element equal to the sum of the i-th row of W;

4: Iterate $f(t+1) = \alpha S f(t) + (\beta Y_1 + \gamma Y_2 + \eta Y_3)$ until convergence, where α, β, η are parameters in (0,1);

5: Let f_i^* denote the limit of the sequence $\{f_i(t)\}$. Each sentence x_i gets its ranking score f_i^*;

For the original manifold-ranking, the iterative formula of the fourth step is $f(t+1)=\alpha S f(t)+(1-\alpha)Y$. The theorem in [24] guarantees that the sequence $f(t)$ converges to

$$f^* = (I - \alpha S)^{-1} Y \quad (1)$$

Without loss of the generality, we can extend the vector Y. Since $(I-\alpha S)$ is invertible, we have

$$f^* = (I - \alpha S)^{-1} (\beta Y_1 + \gamma Y_2 + \eta Y_3) \quad (2)$$

For real-world problems, the iteration algorithm is preferable due to high computational efficiency. Usually when the difference between the scores computed at two successive iterations for any point falls below a given threshold (0.0001 in this paper), the iteration algorithm will converge.

Wan *et al.* [23] proposed and proved an intuition that intra-document links and inter-document links have unequal contributions in the manifold-ranking algorithm. Given a link between a sentence pair of x_i and x_j, if x_i and x_j come from the same document, the link is an intra-document link; if x_i and x_j come from different documents, the link is an inter-document link. The links between the topic sentences and any other sentences are all inter-document links. In our context, distinct weights are assigned to the intra-document links and the inter-document links respectively. In the second step of the above algorithm, the affinity matrix W can be decomposed as

$$W = W_{intra} + W_{inter} \quad (3)$$

where W_{intra} W_{intra} is the affinity matrix containing only the intra-document links (the entries of inter-document links are set to 0) and W_{inter} is the affinity matrix containing only the inter-document links (the entries of intra-document links are set to 0). $RankScore(x_i) = f_i^*(i = 1,...,n)$

We differentiate the intra-document links and inter-document links as follows:

$$W' = \lambda_1 W_{intra} + \lambda_2 W_{inter} \quad (4)$$

We let $\lambda_1, \lambda_2 \in [0,1]$ in the experiments. If $\lambda_1 \le \lambda_2$, the inter-document links are more important than the intra-document links and vice versa. Note that if $\lambda_1 = \lambda_2 = 1$, then Equation(4) reduces to Equation(3). In the iterative feedback based manifold-ranking algorithm, W' is normalized into S' in the third step and the fourth step uses the following iteration form: $f(t+1) = \alpha S' f(t) + (\beta Y_1 + \gamma Y_2 + \eta Y_3)$. The iteration process is shown in Algorithm 2:

Algorithm 2. Power method for computing the stable state of iterative feedback based manifold-ranking

Input: Normalized similarity matrix S'
Input: Matrix size *N*, error tolerance ε
Output: Eigenvector *f*

1: $f(0) = \frac{1}{N}$;
2: t=0;
3: **repeat**;
4: $f(t+1) = \alpha S'^T f(t) + (\beta Y_1 + \gamma Y_2 + \eta Y_3)$
5: $t = t + 1$;
6: $\delta = \| f(t+1) - f(t) \|$;
7: **until** $\delta < \varepsilon$;
8: return $f(t+1)$;

C. Redundancy Removing in Sentence Selection

Based on the normalized original affinity matrix, we apply the greedy algorithm to impose the diversity penalty and compute the final overall ranking scores, representing the importance and relevance to topic and the information novelty of the sentences. For each $timeslice_i$, the algorithm is shown in Algorithm 3:

The algorithm is based on the idea that the overall ranking score of less informative sentences overlapping with the sentences in update summary is decreased. In the second step, where $\omega > 0$ is the penalty degree factor. The larger ω is, the greater penalty is imposed to the overall ranking score. If $\omega = 0$, no diversity penalty is imposed at all. The sentence with highest ranking score is chosen to produce the summary until satisfying the summary length limit.

Algorithm 3. Redundancy removing

Input: Initialize Summary sentences set $A = \phi$, $B = \{x_i \mid i = 1,...,n\}$
Input: $RankScore(x_i) = f_i^*(i = 1,...,n)$, each sentence's overall ranking score is its manifold-ranking score
Output: A

1: Sort the sentences in *B* by their current overall ranking scores in descending order;
2: Suppose x_i is the highest ranked sentence, i.e. the first sentence in the ranked list. Move sentence x_i from *B* to *A*, and then the diversity penalty is imposed to the overall ranking score of each sentence linked with $x_i \in B$ as follows: for each sentence $x_j \in B$,
$RankScore(x_i) = RankScore(x_j) - \omega * S_{ji} f_i^*$
3: Go to step 2 and iterate until $B = \phi$ or exceed the summary length limit;

IV. Experiments

A. Data Set

The dataset of the DUC2007 update summary task is used in our experiments. The update summary task is the first evaluation about TMDS. This task includes a gold standard dataset consisting of document cluster and reference summaries. Ten documents clusters are selected from the 45 clusters of the main task for preparation of the update summary task, and each cluster has 25 documents. Each of these ten clusters is divided into three smaller clusters, A, B, C, where the time stamps on all the documents in each set are ordered such that time(A)<time(B)<time(C). There are approximately 10 documents in A, 8 in B, and 7 in C. The three smaller clusters have the same query as the original larger cluster. The goal of the update summary task is to create short (100-word) multi-document summaries for each smaller clusters under the assumption that the reader has already read a number of previous documents.

B. Evaluation Metric

In order to evaluate the performance and the stability of the proposed approach, we used two kinds of evaluation metrics.

ROUGE [15] is used as the evaluation metric, which has been widely adopted by DUC for automatic summarization evaluation. It measured summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE toolkit reported separate scores for 1, 2, 3 and 4-gram, and also for longest common subsequence co-occurrences and so on. Among these different scores, unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most. The evaluation results of DUC2007 update summary just gave the ROUGE-2 and ROUGE-SU4 scores. Accordingly, we also showed corresponding ROUGE metrics in the experimental results at a

confidence level of 95%, which were computed by running ROUGE-1.5.5[6] with stemming but no removal of stopwords. The input file implemented jackknifing so that scores of systems and humans could be compared.

Pyramid method [18] is also used to evaluate our proposed approach, which is the latest evaluation metric. It incorporates the idea that no single best model summary for a collection of documents exists. The analysis of summary content is based on Summarization Content Units (SCUs). The reference summary is annotated as the set of the SCUs. If the SCUs is contained in more reference summary, it will have the higher weight. After the annotation procedure is completed, the final SCUs can be partitioned in a pyramid. The partition is based on the weight of the SCUs; each tier contains all and only the SCUs with the same weight.

C. Experimental Results and Analysis

ROUGE Metric. We designed seven baselines in addition to the lead baseline (RUNID=35) and the CLASSYO4 baseline (RUNID=58) employed in the update task of DUC2007. We also compared our system with top five systems with highest ROUGE scores, chosen from the performing systems on update task of DUC2007. The comparison results are showed in Table I.

TABLE I
SYSTEM COMPARISON AND RANK ON UPDATE TASK OF DUC 2007
(RECALL SCORE)

System	ROUGE-2	Rank	ROUGE-SU4	Rank
40	0.11189	1	0.14306	1
IFM-ranking	0.09963	2	0.13176	5
55	0.09851	3	0.13509	3
45	0.09622	4	0.13245	4
IFM-ranking-γ	0.09404	5	0.12705	8
IFM-ranking-ω	0.09389	6	0.12985	7
47	0.09387	7	0.13052	6
44	0.0937	8	0.13607	2
IFM-ranking-$\lambda_1:\lambda_2$	0.09206	9	0.12638	9
IFM-ranking-β	0.09019	10	0.12402	10
IFM-ranking-$\gamma-\eta$	0.0872	11	0.12342	11
IFM-ranking-η	0.08503	12	0.1231	12
CLASSY04(58)}	0.08501	13	0.12247	13
IFM-ranking-α	0.07852	14	0.11523	14
Lead Baseline(35)}	0.04543	15	0.08247	15

The Lead Baseline returns all the leading sentences (up to 100 words) of the most recent document. CLASSY04 Baseline ignores the topic narrative, but which had the highest mean SEE coverage score in Task 2 of DUC2004, a multi-document summarization task. The system uses the CLASSY04 HMM[7] terms as observables and the pivoted QR method for redundancy removal. The sentences are chosen only from the most recent collection of documents. For example, the summary for D0703A-B selects sentences only from the 8 articles in this cluster; however, it uses D0703A-A in the computation of signature terms. Likewise, the summary for D0703A-C selects sentences from only the 7 documents in this cluster and only uses D0703A-A and D0703A-B in the computation of signature terms. S40, S55, S45, S47 and S44 are the system IDs of the top performing systems, whose details are described in DUC publications.

IFM-ranking (Iterative Feedback based Manifold-ranking) is our system, which adopts the proposed approach described in Section 3. IFM-ranking-α, IFM-ranking-β, IFM-ranking-γ, IFM-ranking-η, IFM-ranking-ω, IFM-ranking-($\lambda_1:\lambda_2$) and IFM-ranking-$\gamma-\eta$ are seven other baselines. IFM-ranking-α ignores spreading the data points' ranking score to their nearby neighbors via the weighted network. IFM-ranking-β, IFM-ranking-γ, IFM-ranking-η ignores the common topic, the update summary of previous timeslices and first sentences of all document in current timeslice when extending the topic, respectively. IFM-ranking-$\gamma-\eta$ ignores the iterative feedback mechanism, which just considers the common topic in manifold-ranking process. IFM-ranking-$\lambda_1:\lambda_2$ doesn't differentiate the link between the sentences, say $\lambda_1:\lambda_2=1$. IFM-ranking-ω just computes the ranking score of each sentence without the step of imposing diversity penalty. These baselines are all simplified versions of IFM-ranking.

We conduct experiments to focus on the following research questions, which are related to 7 IFM-ranking parameters α, β, γ, η, λ_1, λ_2, ω.

Q1: Is the modified manifold-ranking process useful?
Q2: Is the iterative feedback mechanism effective?
Q3: Does the update summary in previous timeslice or the first sentences of documents in current timeslice help to extend the information richness of topic?
Q4: How does the intra-document or inter-document link affect the performance?
Q5: Is redundancy removing necessary?

The parameters of the IFM-ranking are set as follows: α=0.8, β=0.7, γ=0.3, η=0.4, λ_1=0.3, λ_2=1, ω=8.5.

Seen from Table I, our system ranks 2th and 5th on ROUGE-2 and ROUGE-SU4, respectively, and outperforms all baseline systems.

In comparison with IFM-ranking, ROUGE-2 and ROUGE-SU4 scores of IFM-ranking-α decrease by 0.02111 and 0.01653. Therefore, modified manifold-ranking process affects the update task and parameter α is very important.

It is shown in IFM-ranking-β, IFM-ranking-γ and IFM-ranking-η that the topic description helps to improve the performance, and both the update summary in previous timeslice and the first sentences of documents in current timeslice are beneficial to extend the information richness of topic. At the same time, parameter η brings the highest contribution on performance, β takes the second place, and γ takes third place. It also shows that the first sentence can availably generalize the topic in news field.

[6] http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html
[7] http://duc.nist.gov/pubs/2004papers/ida.conroy.ps

ROUGE-2 and ROUGE-SU4 scores of IFM-ranking-$\gamma-\eta$ decrease by 0.01143 and 0.00834 in comparison with IFM-ranking. This result verifies that iterative feedback mechanism is effective, which models the dynamically evolving characteristic, and represents the relay propagation of information in temporally evolving data.

If IFM-ranking-$\lambda_1:\lambda_2$ doesn't differentiate the links between the sentences (where $\lambda_1:\lambda_2=1$), its ROUGE scores will slightly decrease by 0.00757 and 0.00538 than that of IFM-ranking, respectively. Thus intra/inter-document link differentiation affects the update task.

Without the step of imposing diversity penalty, ROUGE scores of IFM-ranking-ω will decrease by 0.00574 and 0.00191, respectively. Therefore, redundancy removing is necessary.

Comparing with the performing system (RUNID=40) [2] with the highest ROUGE scores respectively on the sub dataset A, B, C of DUC2007 update task, it is shown in table II and table III that our ROUGE scores on A and B are lower than that of the performing system 40. However, ROUGE scores on C are both higher than that of ones by 0.009514 and 0.00555. The performing system 40 adopted much linguistic knowledge and discourse understanding techniques. Knowledge base and coreference resolution are used to evaluate whether a particular extracted commitment is a textual entailment or textual contradiction. However, we just used the shallow sentence-level feature. This further validates that our proposed approach is effective in capturing the update information.

TABLE II
ROUGE-2 RECALL SCORES FOR THREE SUBSETS A, B, C ON UPDATE TASK OF DUC2007

System	A	B	C
40	0.125132	0.105644	0.104285
IFM-ranking	0.0983582	0.086997	0.113799

TABLE III
ROUGE-SU4 RECALL SCORES FOR THREE SUBSETS A,B,C ON UPDATE TASK OF DUC2007

System	A	B	C
40	0.155344	0.134188	0.139419
IFM-ranking	0.130028	0.120542	0.144969

Pyramid Metric. Altogether, there are in total 30 standard pyramids created by annotators. Figure 1 shows the average score, maximum score and our system's score for each pyramid set. IFM-ranking outperforms the average scores in 22 out of 30 sets. Note that for dataset C, the proposed IMF-ranking approach performs better than average performance in 7 out of 10 sets, which shows that iterative feedback mechanism is effective. The average scores over all pyramid sets are show in Figure 2, the best system has the average score of 0.3403, whereas our system obtains 0.29855 on average, which is ranked 4th among all 24 systems. This further shows our approach is stable.

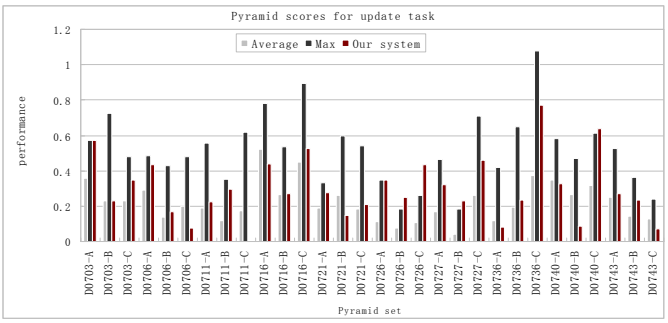

Fig. 1. Pyramid scores for update task.

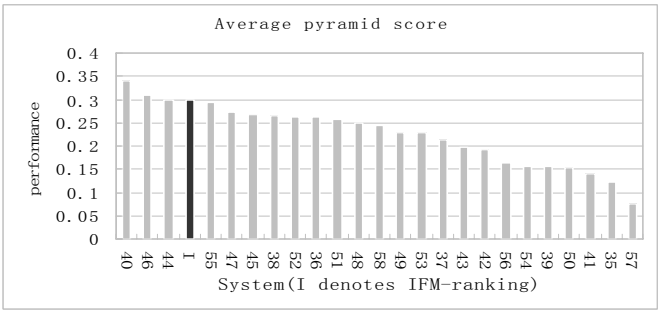

Fig. 2. Average pyramid scores for update task.

Since the update summary task is firstly evaluated in 2007, and we have no other training corpus, thus we cannot directly compare with the top performing system. However, we just use the shallow sentence-level features to achieve the encouraging performance; it will have some instruction on the future participation.

The experiment results suggested that the encouraging performance achieved by IFM-ranking benefits from the following factors: 1) Modified manifold-ranking process; 2) Iterative feedback mechanism; 3) Intra/Inter-document link differentiation; 4) Diversity penalty imposition.

D. Parameter Tuning

As the parameter space is too large to test all possible IFM-ranking algorithms, we adopt the greedy strategy to find the proper parameters value based on ROUGE metric, however, which are impossible optimal. Figures from 3 to 8 show the process of parameter tuning.

When we tune a parameter, the other parameters are set to be the optimal values selected by greedy strategy. Figure 3 demonstrates the influence of the manifold weight α in the proposed approach on performance when β =0.7, γ =0.3, η =0.4, λ_1 =0.3, λ_2 =1, ω =8.5.

Figure 4, Figure 5 and Figure 6 demonstrate the influence of the common topic (β), the update summary in previous timeslices (γ), and the first sentences of documents in current timeslice (η), respectively. From these three figures, it could be observed that both ignoring and excessively depending on the topic description would deteriorate the performance.

Figure 7 demonstrates the influence of the intra/inter-document relationship differentiating weight $\lambda_1:\lambda_2$. It could be observed that the performance curve in field

($\lambda_1 : \lambda_2 < 0.9$) is averagely higher than that in field (λ_1=1 and $\lambda_2 < 0.9$). It shows that inter-document relationships are more important than intra-document relationships for the update task.

Figure 8 demonstrates the influence of the penalty factor ω . It shows that imposing diversity penalty is necessary.

V. Conclusions and Future Work

This paper proposes the iterative feedback based manifold-ranking for update task of DUC2007. Feedback mechanism is used to model the dynamically evolving characteristic, which reveals the relay propagation of information in temporally evolving data. The proposed approach also makes full use of the relationships among sentences and relationships between the topic and the sentences.

However, our approach just used the shallow sentence-level feature, and adopted the greedy strategy to estimate the parameter values, which may be not optimal. In the future, we will mine the deeper level features including temporal event and semantic information, and also explore the parameter optimization algorithm.

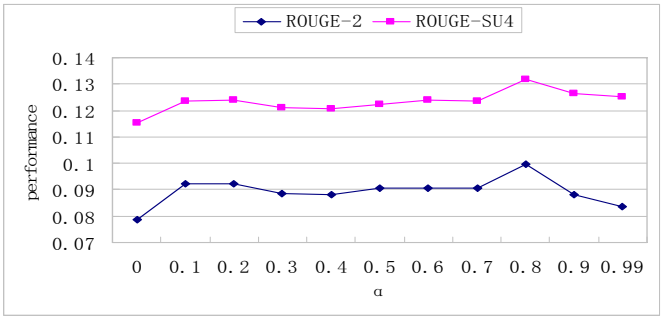

Fig. 3. α vs ROUGE recall scores.

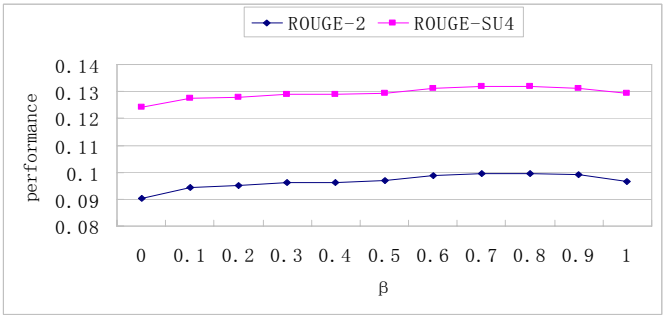

Fig. 4. β vs ROUGE recall scores.

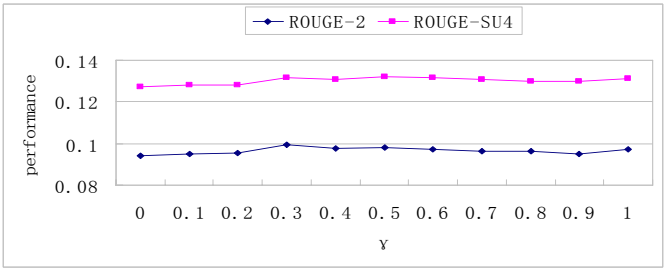

Fig. 5. γ vs ROUGE recall scores.

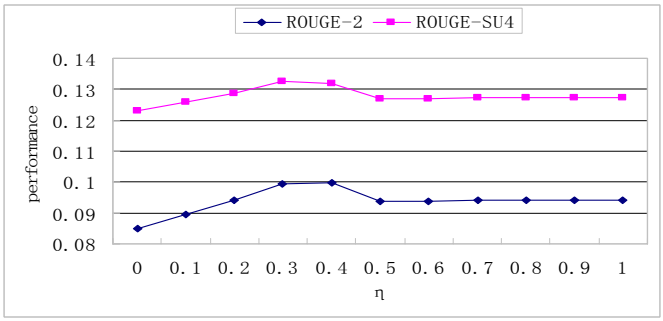

Fig. 6. η vs ROUGE recall scores.

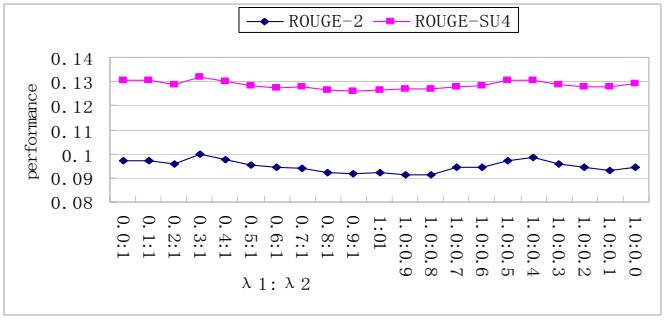

Fig. 7. $\lambda_1 : \lambda_2$ vs ROUGE recall scores.

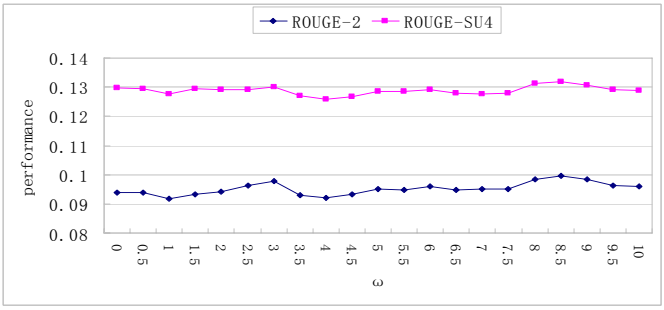

Fig. 8. ω vs ROUGE recall scores.

References

[1] Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18, 2001. K. R. Andrew Hickl and F. Lacatusu. LCC's GISTexter at DUC 2007: Machine Reading for Update Summarization. Proceedings of the DUC2007.

[2] K. R. Andrew Hickl and F. Lacatusu. LCC's GISTexter at DUC 2007: Machine Reading for Update Summarization. *Proceedings of the DUC2007*.

[3] Q. Bing, L. Ting, C. Shanglin, and L. Sheng. Sentences Optimum Selection for Multi-Document Summarization. Journal of Computer Research and Development, 43(6):1129–1134, 2006.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):107–117, 1998.W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[5] J. Conroy, J. Schlesinger, and J. Stewart. CLASSY query based multi-document summarization. Proceedings of the 2005 Document Understanding Workshop, Boston, 2005.

[6] G. Erkan and D. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22:457–479, 2004.

[7] A. Farzindar, F. Rozon, and G. Lapalme. CATS a topic oriented multi-document summarization system at DUC 2005. Proceedings of the 2005 Document Understanding Workshop.

[8] J. Ge, X. Huang, and L. Wu. Approaches to event-focused summarization based on named entities and query words. Proceedings of the 2003 Document Understanding Workshop.

[9] E. Hovy, C. Lin, and L. Zhou. A BE-based multi-document summarizer with query interpretation. Proceedings of the DUC2005.

[10] A. Jatowt and M. Ishizuka. Temporal Web Page Summarization. 5th International Conference On Web Information Systems Engineering, Brisbane, Australia, November 22-24, 2004.

[11] J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632,1999.

[12] C. Kuan-Yu, L. Luesukprasert, and T. Seng-cho. Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. IEEE Transactions on Knowledge and Data Engineering 19, 8 (Aug. 2007), pages1016–1025, 2007.

[13] M. L. Q. W. K. Li, W.J.and Wu. Integrating temporal distribution information into event-based summarization. International Journal of Computer Processing of Oriental Languages, 19:201–222, 2006.

[14] J. Lim, I. Kang, J. J.Bae, and J. Lee. Sentence extraction using time features in multi-document summarization. In Proceedings of the Asia Information Retrieval Symposium 2004, pages 82–93.

[15] C. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of the Workshop on Text Summarization Branches Out, pages 25–26, 2004.

[16] I. Mani. Recent Developments in Temporal Information Extraction (Draft). Nicolov, N., and Mitkov, R. Proceedings of RANLP, 3, 2004.

[17] R. Mihalcea and P. Tarau. TextRank: Bringing Order into Texts. In In Proceedings of Empirical Methods in Natural Language Processing 2004.

[18] A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Trans. Speech Lang. Process., 4(2):4, 2007.

[19] R. K. Prasad Pingali and V. Varma. IIIT Hyderabad at DUC 2007. Proceedings of the DUC2007.

[20] D. Radev, H. Jing, M. Sty´s, and D. Tam. Centroid based summarization of multiple documents. Information Processing and Management, 40(6):919–938, 2004.

[21] H. Saggion, K. Bontcheva, and H. Cunningham. Robust Generic and Query based Summarization. 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2003.

[22] R. Swan and D. Jensen. Constructing Topic-Specific Timelines with Statistical Models of Word Usage. Proceedings of the 6th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 73–80, 2000.

[23] X. Wan, J. Yang, and J. Xiao. Manifold-ranking based topic-focused multi-document summarization. In IJCAI, pages 2903–2908, 2007.

[24] D. Zhou, O. Bousquet, T. Lai, J. Weston, and B. Scholkopf. Learning with Local and Global Consistency. In Proceedings of NIPS2003, 2003.

[25] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf. Ranking on Data Manifolds. In Proceedings of NIPS2003, 2003.

[26] M.-Y. K. W. S. L. L. Q. Ziheng Lin, Tat-Seng Chua and S. Ye. NUS at DUC 2007: Using Evolutionary Models of Text. Proceedings of the DUC2007.