2013 IEEE 13th International Conference on Data Mining

A Novel Relational Learning-to-Rank Approach for Topic-Focused Multi-Document Summarization

Yadong Zhu Yanyan Lan Jiafeng Guo Pan Du Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
{*zhuyadong, dupan*}*@software.ict.ac.cn,* {*lanyanyan, guojiafeng, cxq*}*@ict.ac.cn*

***Abstract*—Topic-focused multi-document summarization aims to produce a summary over a set of documents and conveys the most important aspects of a given topic. Most existing extractive methods view the task as a multi-criteria ranking problem over sentences, where relevance, salience and diversity are three typical requirements. However, diversity is a challenging problem as it involves modeling the relationship between sentences during ranking, where traditional methods usually tackle it in a heuristic or implicit way. In this paper, we propose a novel relational learning-to-rank approach (R-LTR) to solve this problem. Relational learning-to-rank is a new learning framework which further incorporates relationships into traditional learning-to-rank in an elegant way. Specifically, the ranking function is defined as the combination of content-based score of individual sentence, and relation-based score between the current sentence and those already selected. On this basis, we propose to learn the ranking function by minimizing the likelihood loss based on Plackett-Luce model, which can naturally model the sequential ranking procedure of candidate sentences. Stochastic gradient descent is then employed to conduct the learning process, and the summary is predicted by the greedy selection procedure based on the learned ranking function. Finally, we conduct extensive experiments on benchmark data sets TAC2008 and TAC2009. Experimental results show that our approach can significantly outperform the state-of-the-art methods from both quantitative and qualitative aspects.**

I. Introduction

Due to the explosive growth of information on the Web, there is a great need to provide improved techniques for information presentation and exploration. Automatic summarization of a single document, or a set of related documents, has become an important means for people to consume large scale online textual information. One specific task of document summarization is Topic-focused Multi-Document Summarization (TMDS), which aims to create a short summary over a set of documents, and conveys the most important aspects of the given topic. In this paper, we focus on the scenario of extractive summarization, where the summary is produced by extracting sentences from the original documents.

A good summarization of TMDS task is generally supposed to meet the following typical demands: *relevance*, *salience* and *diversity* [1], [2]. Relevance requires the summary to provide related information with respect to the given topic. Salience means that the summary must neglect those trivial contents and retain the most important pieces of information. Diversity requires that the summary should be with less redundant information, and cover different aspects of the given topic. Among the three criteria, relevance and salience are usually captured based on the content of individual sentences with respect to the topic, which can be computed independent of the extraction procedure. While diversity involves modeling the relationship between the current sentence and those already selected [2], thus is coupled with the extraction procedure and become a major challenging problem.

In the literature, most previous work takes the topic-focused multi-document summarization as a multi-criteria ranking problem and proposes different methods to solve this problem. For example, the authors of [3], [4], [5] take summarization as a metric-based sentence ranking problem, where the diversity is described by some heuristically predefined metrics such as marginal relevance [4], distortion measures [3] and the combined weights of the terms covered [5]. Some other methods such as DivRank [6] and MRSP [2] leverage the sentence graph for summarization. In these methods, pairwise relationships are directly defined between the sentence and its neighbors, and the diversity is actually implicitly captured through some random walk process in the graph. However, neither heuristic nor implicit ways can well reflect the demand of diversity. Therefore, topic-focused multi-document summarization remains an challenging problem, especially from the criteria of diversity.

In this paper, we propose a novel relational learning-to-rank approach to solve this problem. Firstly, the topic-focused multi-document summarization is also formalized as a sentence ranking problem. Secondly, the inter-relationships among candidate sentences are considered in the ranking process, besides the relevance and salience information of sentences as used in traditional learning-to-rank [7].

Specifically, the ranking function is defined as the combination of content-based score and relation-based score. The content-based score (for relevance and salience) only depends on the features of an individual sentence, while the relation-based score (for diversity) depends on the relationships between the current sentence and those already selected. In this paper, we describe three different ways to represent the relation-based score. The loss function is then

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.38

defined as the likelihood of the ground-truth summary based on Plackett-Luce model [8], which can naturally model the sequential ranking procedure of candidate sentences. On this basis, stochastic gradient descent is employed to conduct the learning process, and the summary is predicted by the greedy selection procedure based on the learned ranking function.

For evaluation, we conduct extensive experiments based on the benchmark data sets of TAC2008 and TAC2009. The experimental results show that: (1) our approach is more effective than all the state-of-the-art summarization methods with ROUGE-1, ROUGE-2 and ROUGE-4 as evaluation measures; (2) our approach can cover more important facts with respect to reference summaries as compared with traditional methods. In summary, our approach can significantly outperform traditional methods in both quantitative and qualitative aspects.

The rest of the paper is organized as follows. We first review some related work in Section II. Then we describe our proposed relational learning-to-rank approach in section III, including the definitions of ranking function and loss function, learning and prediction procedures. Section IV presents the experimental results and Section V concludes the paper.

II. Related work

We will briefly review some related extractive summarization methods, especially those that considers relevance and diversity requirements. Most extractive methods treat topic-focused multi-document summarization as a ranking problem, and try to assign scores to candidate sentences and extract the sentence with high scores.

Some approaches take summarization as a metric-based sentence ranking problem [4], [3], [9], [5], and predefine different metrics to conduct the ranking process. For example, Carbonell and Godlstein [4] employ a linear combination of relevance and diversity as the metric called "marginal relevance", and iteratively select sentences with the largest "marginal relevance". The method in [3] proposes to define a proper distortion measure, and tackles the multi-document summarization via minimum distortion. Davis et al. use LSA [10] to produce term weights, and then try to maximize "the combined weights of the terms covered" [5]. These methods deal with the relevance and diversity issues heuristically by using a pre-defined fixed criterion, and do not well capture the relationships among sentences.

Many other approaches are mainly graph-based which leverage the sentence graph for summarization. For example, Erkan and Radev proposes LexRank [11], which is a variant of PageRank [12] for generic text summarization. Topic-sensitive LexRank [13] has been applied to the task of query-focused summarization. Zha in [14] proposes a mutual reinforcement principle for sentence extraction using the idea of HITS [15]. Wan et al. apply a manifold-ranking algorithm to topic-focused summarization [16]. DivRank [6] leverages a vertex-reinforced random walk to achieve relevance and diversity for summarization. Absorbing random work [17] is also applied to achieve a diverse ranking for summarization. Du et al. [2] introduce sink points into manifold ranking for summarization to capture relevance and diversity. The work in [1] proposes a supervised lazy random walk approach, which learns the strengths of edges and self-loops and then conducts the lazy random walk on such a weighted sentence graph. Overall, in these graph-based methods, the pairwise relationships are directly defined between the sentences and its neighbors, and the diversity is implicitly captured through some random walk process in the graph. This implicit ways also can not well reflect the demand of diversity. Additionally, these methods are usually with high time complexity and low computing efficiency.

Some previous work also attempts to view the summarization task as a machine-learning problem, and trains accurate models for extractive summary. For example, the Support Vector Machine (SVM) [18], [19] based methods focus on constructing a decision boundary between summary sentences and non-summary sentences. Several traditional learning-to-rank methods such as ranking SVM, support vector regression and gradient boosted decision trees are also applied to the summary task [20], [21]. These methods rely on the assumption that sentences are independent from each other and ignore the relationships among sentences. Some other methods such as Hidden Markov Models (HMM) based [22] and Conditional Random Field (CRF) based [23], have been proposed by relaxing the independence assumption and try to modeling the relations between sentences, but the demand of diversity is still not well captured.

The method proposed in [24] is closely related to our work for it explicitly considers relevance and diversity requirements, by the formulation of a supervised sentence ranking problem, based on Structural SVM techniques. Whereas, it focuses on generic summarization rather than topic-focused scenario, and is different from our research problem. In our work, we propose a novel relational learning-to-rank approach to formulate the TMDS task as a sentence ranking problem, which explicitly captures the diversity, besides the relevance and salience information of sentences.

III. Our Approach

As described in Section 2, many traditional methods formalize the topic-focused multi-document summarization as a ranking problem. In this paper, we follow this methodology and utilize learning-to-rank technique to solve this problem. In traditional learning-to-rank (LTR for short) [7], a ranking function is defined on the content of each individual sentence and learned toward some loss function. However, in the task of TMDS, the overall ranking of sentences for a given topic, should depend not only on the individual ranked sentences, but also on how they related to each other. Therefore,

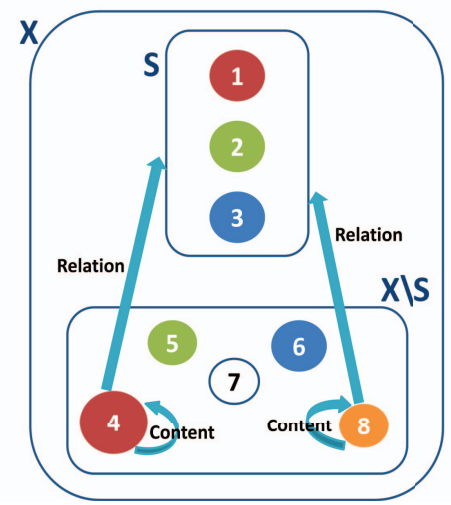

Figure 1. An illustration of the sequential ranking for candidate sentences.

in this paper, we introduce a novel relational learning-to-rank approach to solve the topic-focused multi-document summarization problem. The difference between LTR and R-LTR is that the latter considers both content of individual sentences and relations between sentences when defining the ranking function

First, we introduce the framework of relational learning-to-rank. Let $X = \{x_1, \cdots, x_n\}^T$ be a $n \times d$ matrix representing d dimensional feature vectors of n candidate sentences; each row corresponds to one sentence and each column corresponds to one feature. Let $R \in \mathcal{R}^{n \times n \times l}$ denote a 3-way tensor representing relationships between the n sentences, where R^{ijk} stands for the k-th feature of relation between sentences x_i and x_j. Let y be a ground-truth summary of X. Supposing that $f(X, R)$ is a ranking function, and the goal of relational learning-to-rank is to output the best ranking function from a function space $\mathcal{F}$.

In training, the labeled data about N topics are given as $(X_1, R_1, y_1), (X_2, R_2, y_2), \cdots, (X_N, R_N, y_N)$, where X_i denote feature vectors of n_i sentences. A loss function L is defined, and the learning process is conducted by minimizing the total loss with respect to the given training data.

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} L(f(X_i, R_i), y_i). \quad (1)$$

In prediction, given features X_t and relations R_t for n_t sentences, we output y_t based on the learned ranking function $\hat{f}(X_t, R_t)$. We will define specific ranking function and loss function for the TMDS task in the following subsections.

A. Definition of Ranking Function

Before we define the ranking function, we first consider how human beings extract a summary. People usually extract a summary in a sequential manner, considering relevance, salience and diversity for each sentence based on what he/she has already extracted in the previous steps. Therefore, the TMDS task can be treated as a sequential ranking process, where each sentence is ranked according to its relevance/salience to the topic and the relation between all the sentences ranked before it.

The intuitive idea is illustrated in Figure 1, all the balls represent candidate sentences of a given topic, and different colors represent different subtopics (or aspects). The solid ball is relevant to the topic, and the hollow ball is irrelevant to the topic, and larger size means more relevant. X denotes all the candidate sentences. S denotes previously selected sentences, and $X \backslash S$ denotes the remanent sentences. When ranking sentences in $X \backslash S$ given the already ranked results S, both content-based relevance/salience, and diversity relations between this sentence and the previously selected sentences in S should be considered. Noting that larger size of the ball means the sentence is more relevant to the topic, and different colors represent different subtopics (or aspects) of the given topic. Therefore, the sentence 8 may be more preferred than sentence 4 given S, since it is relevant to the topic, and also provides different aspects additionally comparing with the selected set S.

Based on this ranking process, we give the precise definition of ranking function as follows. Given a topic q, we assume that a set of sentences have been selected, denoted as S, the ranking function on the rest sentences $X \backslash S$ is then defined as the combination of the content-based score (for relevance and salience) and the relation-based score (for diversity) between the current sentence and those previously selected, as shown below:

$$f_S(x_i, R) = \omega_r^T x_i + \omega_d^T h_S(R^i), \forall x_i \in X \backslash S, \quad (2)$$

where R^i stands for the matrix of relationships between sentence x_i and other sentences, with each R^{ij} stands for the relationship between x_i and x_j, represented by the feature vector of $(R^{ij1}, \cdots, R^{ijl})$, $h_S(R^i)$ stands for the relational function on $R^{ij}, x_j \in S$, ω_r^T and ω_d^T stands for the corresponding content and relation weight vector. When $S = \phi$, $f_S(x_i, R)$ is directly represented as $\omega_r^T x_i$. From the above definition, we can see that if we do not consider diversity relations, our ranking function reduces to $f(X)$, which is the traditional ranking function in learning-to-rank.

In order to accomplish the definition of new ranking function, we first need to define three key components: relational function $h_S(R^i)$, relation-based feature vector R^{ij} and the content-based feature vector x_i. We introduce their definitions one by one.

1) Relational Function $h_S(R^i)$: Please note that the relational function $h_S(R^i)$ represents the diversity relationship between the current sentence x_i and the previously selected sentences in S, and each $R^{ij}, x_j \in S$ stands for the diversity relationship between sentence x_i and x_j. If we treat diversity

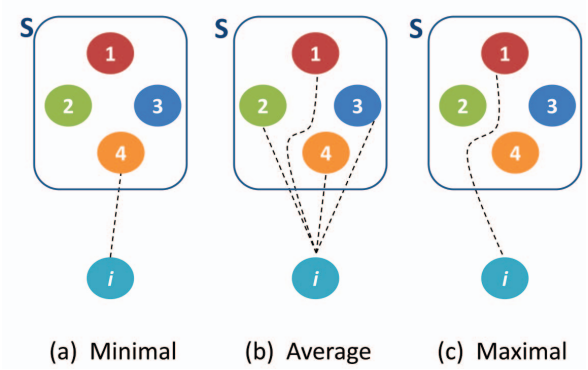

Figure 2. Different ways to measure the distance of x_i to S

as distance, $h_S(R^i)$ can be viewed as the distance of x_i to the set S. According to different definitions of the distance between an item and a set of items, $h_S(R^i)$ can be defined as the following three forms (illustrated in Figure 2).

Minimal Distance. The distance between a sentence x_i and a set S is defined as the minimal distance of all the sentence pairs $(x_i, x_j), x_j \in S$. Therefore, $h_S(R^i)$ is defined as follows:

$$h_S(R^i) = \min_{x_j \in S} R^{ij}.$$

Average Distance. The distance between a sentence x_i and a set S is defined as the average distance of all the sentence pairs $(x_i, x_j), x_j \in S$. Therefore, $h_S(R^i)$ is defined as follows:

$$h_S(R^i) = \frac{1}{|S|} \sum_{x_j \in S} R^{ij}.$$

Maximal Distance. The distance between a sentence x_i and a set S is defined as the maximal distance of all the sentence pairs $(x_i, x_j), x_j \in S$. Therefore, $h_S(R^i)$ is defined as follows:

$$h_S(R^i) = \max_{x_j \in S} R^{ij}.$$

2) Relation-based Feature Vector R^{ij}: How to define discriminative relation-based features that can well capture diversity relation is critical for the success of R-LTR in TMDS. In this work, we provides several representative features for the learning process.

Cosine Diversity. The cosine diversity between two sentences is calculated based on their weighted term vector representations, and define the feature as follows.

$$R^{ij1} = 1 - \frac{\mathbf{s}_i \cdot \mathbf{s}_j}{\|\mathbf{s}_i\| \|\mathbf{s}_j\|}$$

where $\mathbf{s}_i$, $\mathbf{s}_j$ are the weighted term vectors of sentences based on $tf*isf$, and tf denotes the term frequencies, isf denotes inverse sentence frequencies.

Jaccard Diversity. The Jaccard diversity between two sentences measures the ratio of overlapped terms, and is defined as follows.

$$R^{ij2} = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$$

where S_i, S_j are the term vectors of sentences.

Subtopic Diversity. Different sentences may associate with different aspects of the given topic. We use Probabilistic Latent Semantic Analysis (PLSA) [25] to model implicit subtopics distribution of candidate sentences. Then we can define a kind of subtopic diversity feature based on the KL distance, as follows.

$$R^{ij3} = \sum_{z_i \in Z} P(z_i|S_i) log \frac{P(z_i|S_i)}{P(z_i|S_j)}$$

$$P(z_i|S_i) = \frac{1}{|S_i|} \sum_{w_j \in S_i} P(z_i|S_i, w_j)$$

where $P(z_i|S_i, w_j)$ is calculated and saved in the E-step of the EM procedure.

Document-Level Co-occurrence. The document-level co-occurrence is a binary feature, which indicates whether the two sentences co-occur in a same document. It is set to be 0 if they are within one document, 1 otherwise, and denoted as R^{ij4}.

Based on these diversity features, we can obtain the relation-based feature vector $R^{ij} = (R^{ij1}, R^{ij2}, R^{ij3}, R^{ij4})$. Please note that here we only list some representative diversity features used in our work, and our model is flexible to incorporate other useful diversity features for capturing the diversity relation.

3) Content-based Feature Vector x_i: Content-based features are used to capture the relevance and salience of each individual sentence. For content-based features, we view each sentence as a "short" document, and then employ some standard relevance and salience features used in LTR research [26], as summarized in Table I. They include four topic-dependent and two topic-independent features, and T-S means that the feature is dependent on both the topic and the sentence, and S means that the feature only depends on the sentence.

Weighting features. The typical weighting models include TF-IDF, BM25 and language models. For language model, we use query-likelihood language model with Dirichlet prior.

Table I
CONTENT-BASED FEATURE FOR LEARNING.

Category	Feature Description	Total
T-S	TF-IDF	1
T-S	BM25	1
T-S	QL.DIR	1
T-S	MRF	2
T	Length	1
T	Pos	1

Term Dependency features. We also employ the classic term dependency features such as MRF [27], to enhance capturing of relevance. Additionally, the MRF has two types of values: ordered phrase and unordered phrase, so the total feature number is 2.

Length. It is the number of words in each sentence after removing the stopping words. We prefer sentences with its length located in a certain range. Several length levels are introduced to denote the sentence length, and a feature vector can be generated with each component denoting a length level.

Pos. It is a binary feature which indicates whether the sentence is the first sentence of a document. If the sentence appears at the beginning of a document, it is set to be 1, otherwise 0.

B. Definition of Loss Function

In order to accommodate the sequential ranking process as described in the definition of ranking function, we propose to model the generation of the summary in a sequential way, and define the loss function as the likelihood loss of the generation probability.

Intuitively, the probability of a extracted summary can be viewed as a process to iteratively select the top ranked sentences from the remaining sentences. The precise definition is given as follows.

$$\begin{aligned} P(y|X) &= P(x_{y(1)}, x_{y(2)}, \cdots, x_{y(n)}|X) \\ &= P(x_{y(1)}|X)P(x_{y(2)}|X\backslash S_1)\cdots P(x_{y(n-1)}|X\backslash S_{n-2}) \end{aligned} \tag{3}$$

where $y(i)$ stands for the index of sentence which is ranked in position i in the summary y, $S_i = \{x_{y(1)}, \cdots, x_{y(i)}\}, i = 1, \cdots, n$, $P(x_{y(1)}|X)$ stands for the probability that $x_{y(1)}$ is ranked first among the sentences in X, and $P(x_{y(j)}|X\backslash S_{j-1})$ stands for the probability that sentence $x_{y(j)}$ is ranked first among the sentences in $X\backslash S_{j-1}$.

On the basis of the generation probability of a summary, we define the loss function as the likelihood loss:

$$L(f(X, R), y) = -\log P(y|X). \tag{4}$$

The above sequential definition approach can be well captured by the Plackett-Luce Model [8]. Specifically, we propose to define $P(x_{y(1)}|X)$ and $P(x_{y(j)}|X\backslash S_{j-1})$ as follows.

$$P(x_{y(1)}|X) = \frac{\exp\{f(x_{y(1)})\}}{\sum_{k=1}^{n}\exp\{f(x_{y(k)})\}}, \tag{5}$$

$$P(x_{y(j)}|X\backslash S_{j-1}) = \frac{\exp\{f_{S_{j-1}}(x_{y(j)}, R)\}}{\sum_{k=j}^{n}\exp\{f_{S_{j-1}}(x_{y(k)}, R)\}}. \tag{6}$$

Incorporating Eq.(5) and Eq.(6) into Eq.(3), the probability of a ranking list is formulated as follows.

$$P(y|X) = \prod_{j=1}^{n}\frac{\exp\{f_{S_{j-1}}(x_{y(j)}, R)\}}{\sum_{k=j}^{n}\exp\{f_{S_{j-1}}(x_{y(k)}, R)\}}, \tag{7}$$

where $S_0 = \phi$, $f_\phi(x, R) = \omega_r^T x$.

Algorithm 1 Optimization Algorithm

Input: training data $\{(X_i, R_i, y_i)\}_{i=1}^N$,
parameter: learning rate η, tolerance rate ϵ
Output: model vector: ω_r, ω_d

```
Initialize parameter value ω_r, ω_d
repeat
    Shuffle the training data
    for i = 1, ..., N do
        Compute gradient Δω_r^(i) and Δω_d^(i)
        Update model: ω_r = ω_r − η × Δω_r^(i),
                      ω_d = ω_d − η × Δω_d^(i)
    end for
    Calculate likelihood loss on the training set
until the change of likelihood loss is below ε
```

C. Learning and Prediction

Based on the definitions of ranking function and loss function, we present the learning and prediction process in this subsection. Specifically, we first introduce the optimization procedure, and then we show how to make predictions based on the learned ranking function.

1) Optimization: Given the training data $\{(X_i, R_i, y_i)\}_{i=1}^N$, the total loss is represented as follows.

$$-\sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\left\{\frac{\exp\{\omega_r^T x_{y(j)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(j)})\}}{\sum_{k=j}^{n_i}\exp\{\omega_r^T x_{y(k)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(k)})\}}\right\}$$

We employ Stochastic Gradient Descent (SGD) to conduct minimization as shown in Algorithm 1, and the gradient at summary X_i is computed as follows.

$$\begin{aligned}\Delta\omega_r^{(i)} = \sum_{j=1}^{n_i}\Bigg\{&\frac{\sum_{k=j}^{n_i} x_{y(k)}^{(i)}\exp\{\omega_r^T x_{y(k)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(k)})\}}{\sum_{k=j}^{n_i}\exp\{\omega_r^T x_{y(k)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(k)})\}} \\ &- \frac{x_{y(j)}^{(i)}\exp\{\omega_r^T x_{y(j)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(j)})\}}{\exp\{\omega_r^T x_{y(j)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(j)})\}}\Bigg\},\end{aligned}$$

$$\begin{aligned}\Delta\omega_d^{(i)} = \sum_{j=1}^{n_i}\Bigg\{&\frac{\sum_{k=j}^{n_i} h_{S_{j-1}^{(i)}}(R_i^{y(k)})\exp\{\omega_r^T x_{y(k)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(k)})\}}{\sum_{k=j}^{n_i}\exp\{\omega_r^T x_{y(k)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(k)})\}} \\ &- \frac{h_{S_{j-1}^{(i)}}(R_i^{y(j)})\exp\{\omega_r^T x_{y(j)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(j)})\}}{\exp\{\omega_r^T x_{y(j)}^{(i)} + \omega_d^T h_{S_{j-1}^{(i)}}(R_i^{y(j)})\}}\Bigg\}.\end{aligned}$$

2) Prediction: In order to accommodate the sequential ranking process as described in the definition of ranking function, we propose a greedy selection procedure for prediction, as described in Algorithm 2. Specifically, in the first step, the most relevant and salient sentence with maximal content-based score will be selected and ranked first. If the top k items have been selected, then the sentence in position $k+1$ should be with maximum f_{S_k}. At last, a proper number of sentences are selected to generate a final summary.

Algorithm 2 Summary Generation via Greedy Selection

Input: $X_t, R_t, \omega_r, \omega_d$
Output: y_t

```
Initialize S_0 ← ∅, y_t = (1, ··· , n_t)
for k = 1, ..., K do
    best_sen ← argmax_{x∈X_t} f_{S_{k−1}}(x, R_t)
    S_k ← S_{k−1} ∪ best_sen
    y_t(k) ← the index of best_sen
end for
return y_t = (y_t(1), ··· , y_t(K))
```

IV. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed approach in topic focused multi-document summarization. Firstly, we give some introductions on the data sets, evaluation measures, baseline methods and experimental setup. Secondly, we present our experimental results from both quantitative and qualitative aspects.

A. Datasets and Evaluation Metric

Table II
The statistical information of set A in TAC2008 and TAC2009

	TAC2008	TAC2009
#docs	480	440
#topics	48	44
#average sentences	22.1	21.6
#average words	22.5	22.3
data collection	AQUAINT-2	AQUAINT-2
summary length	100 words	100 words

TMDS has been one of main task in Text Analysis Conference (TAC), which is hold by NIST[1] for several years. In our experiments, we use the benchmark datasets of TAC2008 and TAC2009, which provide 48 topics and 44 topics respectively. Each topic was associated with 20 relevant documents from the AQUAINT-2 collection of news articles. The documents were further equally divided into two sets: set A and set B. All the documents in set A chronologically preceded the documents in set B. The detailed statistical information of set A is present in Table II. In TMDS task, a 100-word summary is required to be generated for each set of documents. The summary of set A should be a topic-focused multi-document summary, and that of B should be a update summary. Here we focus on the performance comparison of topic-focused multi-document summarization (the summary of set A), and do not show the summarization results of set B.

In our work, we use the ROUGE-1 (unigram-based), ROUGE-2 (bigram-based) and ROUGE-SU4 (an extended version of ROUGE-2) recall metrics [28] as the evaluation measures, which were shown to have a high correlation with human judgements. They were also used as official automatic evaluation metrics for TAC2008 and TAC2009. The evaluation results are obtained with version 1.5.5 of ROUGE with the default settings used for TAC2008.

ROUGE evaluates summary quality by counting the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary and the ideal summaries created by humans. The n-gram recall metric, ROUGE-N, is computed as follows.

$$ROUGE\text{-}N = \frac{\sum_{S \in Refs} \sum_{gram_n \in S} Cnt_{match}(gram_n)}{\sum_{S \in Refs} \sum_{gram_n \in S} Cnt(gram_n)}$$

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

B. Baseline Methods

We compare our proposed R-LTR method with the state-of-the-art approaches as described in Section 2, including Leading Sentence Selection (LEAD), Maximal Marginal Relevance (MMR), Personalized PageRank (PPR), Manifold Ranking (MR), DivRank (DR), GrassHopper (GH), Supervised Lazy random walk (SuperLazy) and Support Vector Machine (SVM).

LEAD. It is a simple baseline method that selects the leading sentences of documents to compose the summary. It can be viewed as a lower bound of extractive approach for TMDS task.

MMR. It employs a linear combination of relevance and diversity as the metric called "marginal relevance" [4]. MMR will iteratively select sentences with the largest "marginal relevance".

PPR. PPR is a widely used random walk based ranking approach [29], which is appropriate for identifying relevant and salient sentences for summary.

MR. MR is a graph-based ranking approach using the graph regularization [30]. It is capable of giving higher ranks to the salient sentences which are close to the topic on the manifold.

DR. DR is also a graph-based ranking approach which uses a vertex-reinforced random walk [6]. It focuses on achieving diversity in sentence ranking for summary.

GH. GH conducts on absorbing random walk over the graph [17]. It aims to capture diversity in ranking by iteratively select top ranked nodes and set it as absorbing state.

SuperLazy. SuperLazy is a supervised approach based on sentence graph [1]. It learns the strength of edges and self-

[1]http://www.nist.gov

Table III
THE SUMMARIZATION PERFORMANCE BASED ON 4-FOLD CROSS VALIDATION ON TAC2009

Method	ROUGE-1	ROUGE-2	ROUGE-SU4
LEAD	0.30192	0.06311	0.09869
MMR	0.34291 (+13.58%)	0.07915 (+25.42%)	0.11138 (+12.86%)
MR	0.35499 (+17.58%)	0.08612 (+36.46%)	0.12072 (+22.32%)
PPR	0.36163 (+19.78%)	0.08490 (+34.53%)	0.12497 (+26.63%)
DR	0.36571 (+21.13%)	0.08283 (+31.25%)	0.12306 (+24.69%)
GH	0.36802 (+21.89%)	0.08927 (+41.45%)	0.12469 (+26.34%)
SVM	0.35889 (+18.87%)	0.09103 (+44.24%)	0.12794 (+29.64%)
SuperLazy	0.38210 (+26.56%)	0.10857 (+72.03%)	0.14245 (+44.34%)
R-LTR$_{min}$	**0.40874** (+35.38%)	**0.12964** (+105.42%)	**0.15873** (+60.83%)
R-LTR$_{avg}$	0.40135 (+32.93%)	0.12761 (+102.20%)	0.15461 (+56.66%)
R-LTR$_{max}$	0.40627 (+34.56%)	0.12548 (+98.83%)	0.15394 (+55.98%)

Table IV
THE SUMMARIZATION PERFORMANCE ON TAC2008

Method	ROUGE-1	ROUGE-2	ROUGE-SU4
LEAD	0.28889	0.05871	0.09300
MMR	0.33666 (+16.53%)	0.07533 (+28.31%)	0.11383 (+22.40%)
MR	0.36312 (+25.69%)	0.09402 (+60.14%)	0.12910 (+38.81%)
PPR	0.36045 (+24.77%)	0.08889 (+51.41%)	0.12708 (+36.65%)
DR	0.36548 (+26.51%)	0.09077 (+54.61%)	0.12897 (+38.68%)
GH	0.36549 (+26.52%)	0.09039 (+53.96%)	0.13017 (+39.97%)
SVM	0.36092 (+24.93%)	0.09432 (+60.65%)	0.12814 (+37.78%)
SuperLazy	0.36739 (+27.17%)	0.09645 (+64.28%)	0.13070 (+40.54%)
R-LTR$_{min}$	**0.38927** (+34.75%)	**0.11576** (+97.17%)	**0.14765** (+58.76%)
R-LTR$_{avg}$	0.38654 (+33.80%)	0.11043 (+88.09%)	0.14386 (+54.69%)
R-LTR$_{max}$	0.38576 (+33.53%)	0.10924 (+86.07%)	0.14205 (+52.74%)

loops, and then conducts the lazy random walk on such a weighted sentence graph.

SVM. SVM is widely used as a binary classifier [18]. It is a supervised approach that generally used to distinguish summary sentences from non-summary sentences. Additional steps are taken to remove redundant sentences.

According to the different ways in defining the diversity function $h_S(R^i)$ in section III, our R-LTR methods will have three variants, denoted as R-LTR$_{min}$, R-LTR$_{avg}$, and R-LTR$_{max}$, respectively.

C. Experimental Setup

In our experiments, we view each candidate sentence as a "short" document, and then use Indri toolkit (version 5.2)[2] as the retrieval platform for indexing and basic feature extraction. For data preprocessing, we apply Porter stemmer and stopwords removing for topics and sentences.

For training, we leverage the human extracted summaries as ground truth. NIST has provided a set of manual summarization results on TAC2009, which is extracted by a team of five human "sentence-extractors" from the University of Montreal. This kind of summary can be viewed as an approximate upper bound on what can be achieved with a purely extractive summarizer. Therefore, we used such human extracted summaries as ground truth for our approach and two other supervised baseline methods such as SuperLazy and SVM.

For the topic set on TAC2009, we use a 4-fold cross validation for training and testing. The learning rate η parameter in Algorithm 1 is chosen from 10^{-7} to 10^{-1}, and the best learning rate is chosen based on the performance of training. For testing, we permute the folds until all the folds have been chosen for the test set, and the final performance is reported as the average over all test data. For all the baseline methods, we set their parameters (if exist) to be the values as they can achieve their best performances on ROUGE-2 evaluation metric.

D. Quantitative Evaluation

1) Evaluation Results on TAC2009: We first conduct performance comparison based on the dataset TAC2009. The 4-fold cross-validation results are demonstrated in Table III. The number in the parentheses are the relative improvements compared with the baseline method LEAD. Boldface indicates the highest scores among all runs.

From the results we can see that, our three methods all outperform metric-based sentence ranking method, i.e. MMR, in terms of all the ROUGE metrics consistently. Specifically, the relative improvement over the ROUGE-1,

[2]http://lemurproject.org/indri

ROUGE-2 and ROUGE-SU4 scores of our best method R-LTR$_{min}$ are 19.20%, 63.79% and 42.51%, respectively. It indicates that our approach can better capture the relevance, salience and diversity by using rich features in a supervised way, as compared with the heuristically predefined way in MMR. Meanwhile, our R-LTR approaches also outperform the graph-based methods. For example, when compared with the best performed graph-based approach SuperLazy, the relative improvements over the ROUGE-1, ROUGE-2 and ROUGE-SU4 scores of our best method R-LTR$_{min}$ are 6.97%, 19.41% and 11.43%, respectively. It shows that better summary can be produced by explicitly model the content and relation based features, as compared with the implicit way in graph-based methods.

Furthermore, our approach is also better than SVM. The relative improvement over the ROUGE-1, ROUGE-2 and ROUGE-SU4 scores of our best method R-LTR$_{min}$ are 13.89%, 42.41% and 24.07%, respectively. This demonstrates that by modeling the sequential generation procedure of extracted summary in a ranking perspective, we can learn a much better summarization model than that in a classification view. We further conduct statistical tests on the results, which indicates that all these improvements are statistically significant ($p\text{-}value < 0.01$).

It can also be seen that the R-LTR$_{min}$ obtains better performance than the other two variants of our R-LTR approach. It indicates that when defining the diversity function $h_S(R^i)$, the minimal distance would be a better choice, yet the performance differences among them are very small.

2) Evaluation Results on TAC2008: We conduct experiments on dataset TAC2008 to further evaluate the effectiveness of our model. Since there is no human extracted summarization results as ground truth on dataset TAC2008, we train our model based on dataset TAC2009 and use dataset TAC2008 as the test set. In this way, we can further evaluate the generalization ability of our approach. The evaluation results are shown in Table IV. The evaluation results again demonstrate that our methods can achieve much better performance than both the state-of-the-art supervised and unsupervised approaches in terms of all the ROUGE metrics. This is consistent with the results we obtained on dataset TAC2009. The results also indicate that our approach has strong generalization ability on different datasets.

E. Qualitative Evaluation

Besides the quantitative evaluation in above subsection, we also conduct qualitative comparison based on the content of summary generated by our approaches and baseline methods to get a more intuitive understanding. We randomly choose a topic on dataset TAC2009 as example. The topic and its description is, "glendale–describe the glendale train crash, the cause of the crash, and the arrest and trial of the man accused of causing it." We show four manual summaries provided by NIST as reference summaries, and the summaries generated by our approach and three representative baselines in Table V.

We manually annotated the important facts covered by the reference summaries with circled numbers and identified in total 11 topic-related facts in the references. We then manually annotated these facts in the summaries generated by our approach and baseline methods. Due to the page limitation, here we list the statistics of the fact coverage of R-LTR$_{min}$ and three baseline methods (i.e. LEAD, GH, and SuperLazy) in Table VI. The *Facts* column shows the major related words, the *Weight* column shows the times that the corresponding fact has appeared in the four reference summaries, and the rest of columns record the fact coverage status of different approaches.

The results in Table V show that our R-LTR approach covers the most facts as compared to other baselines. Specifically, our R-LTR approach captures all the four facts with the highest weight (i.e., weight = 4), and 3 out of 4 important facts (weight = 3), which can strongly prove the effectiveness of our approach. Among baseline methods, SuperLazy and GH have shown comparable performance. GH method also captures all the facts with the highest weight, but its total number of facts covered is less than our approach. The LEAD method performs worst, and only covers one fact with the highest weight.

V. Conclusion

In this paper, we propose a novel relational learning-to-rank approach for the task of topic-focused multi-document summarization. Our approach models the diversity relationship among sentences, besides the content information of sentences.

Firstly, we define the ranking function as the combination of content-based score and relation-based score between the current sentences and those previously selected. Secondly, the loss function is defined as the likelihood of the ground-truth summary based on Plackett-Luce model, which can naturally model the sequential ranking procedure of candidate sentences. On this basis, stochastic gradient descent is employed to conduct the training process, and the summary is generated by a greedy selection process based on the learned ranking function. The experimental results on the benchmark data sets TAC2008 and TAC2009 demonstrate that our approach can significantly outperform the state-of-the-art methods from both quantitative and qualitative aspects.

For the future work, it would be interesting to adapt the learning method to the update summarization task, and also test our approach on other types of dataset beyond the traditional documents, e.g., the prevalent short texts in social media.

Acknowledgment

This research work was funded by the 973 Program of China under Grants No. 2012CB316303 and No.

Table V
QUALITATIVE COMPARISON ON A EXAMPLE TOPIC IN TAC2009

Topic	Describe the ***Glendale train crash***, the ***cause of the crash***, and the ***arrest and trial*** of the man accused of ***causing*** it.
RefSum I	①*Juan Alvarez left his car on a railroad track in Glendale CA*. It ②*set off a collision that derailed two trains*, ③*killed at least 10 people and injured nearly 200*. Witness identified Alvarez and ④*he was detained*. He was distraught and ⑤*put on suicide watch*. Glendale police believed ⑥*he intended to commit suicide but changed his mind*. ⑦*He was taken into custody and charged with at least 10 counts of homicide*. Investigator felt Alvarez wanted to kill himself because his ⑧*estranged wife denied him visits with his children*. He was held on ⑦*suspicion of murder* in Los Angeles without bail.
RefSum II	⑤*A man intend on suicide* ①*left his car on a railroad track* in Clendale, California, where it ②*set off a collision that derailed two commuter trains* and ③*killed ten people and injured nearly 200*. Police held Juan Manuel Alvarez on ⑦*ten counts of murder* and indicated that he ⑥*had intended suicide but changed his mind and fled his vehicle* before the train struck. Alvarez, who ⑨*had self-inflicted superficial wounds*, remained at the scene and was cooperating with police. ⑩*He has a criminal record involving drugs* and his wife obtained a ⑧*restraining order against him*.
RefSum III	Juan Manuel Alvarez ①*drove his SUV onto Glendale train tracks* in ⑤*a suicide attempt*. When his SUV got stuck between the track, ⑥*he got out and left*. A southbound Metrolink commuter train ②*crashed into* the immovable SUV and climed over it. The front cab car rammed into an adjacent stationary Union Pacific work train and toppled it. The remaining cars jackknifed and derailed a northbound Metrolink commuter train. The accident and subsequent fire ③*killed 11 and injured 200*. Alvarez watched the crash then attempted suicide again by ⑨*slashing his wrists and stabbing himself*. He ④*was arrested* and ⑦*charged with homicide*.
RefSum IV	②*A train crash in Glendale*, California, on 26 January ③*killed 11 people and injured close to 200*, many critically. A southbound commuter train, pushed from behind by a locomotive, hit an SUV lodged on the tracks, then hit a freight train and a northbound commuter train. ④*Police arrested Juan Manuel Alvarez,* the driver of SUV, who had ⑥*jumped from the car just before the impact*. At first, ⑨*self-inflicted wounds on his wrists and chest* led investigators to think that ⑤*it was an attempted suicide*. It was determined later that he ⑪*got the wounds after he fled the scene*.
LEAD	Juan Miguel Alvarez, ⑦*charged with murder* with special circumstances in the deaths of 11 Metrolink passengers, ⑨*slashed and stabbed himself* after seeing the horrific train crash he caused, sources close to the investigation said Thursday. Alvarez, 25, despondent over the ⑧*breakup of his marriage had planned to kill himself* when he drove his green Jeep Grand Cherokee in the path of a Metrolink train, officials said, but ⑥*bolted from the vehicle at the last moment*. As he watched, southbound Train No.
GH	A man ⑤*intend on committing suicide* ①*left his car on a railroad track in Glendale* near downtown Los Angeles Wednesday, where it ②*set off a collision that derailed two commuter trains*, ③*killing at least 10 people and injuring nearly 200*, authorities said. Several passengers said they saw Alvarez ⑥*jump from his SUV* just before the southbound train plowed into the vehicle, Adams said. Metrolink train 100, a four-car train being pushed by a locomotive, was southbound from Moorpark to Los Angeles and approaching the Glendale station at 6:02 morning when its front car hit Alvarez's Jeep.
SuperLazy	A man ⑤*intend on committing suicide* ①*left his car on a railroad track in Glendale* near downtown Los Angeles Wednesday, where it ②*set off a collision that derailed two commuter trains*, ③*killing at least 10 people and injuring nearly 200*, authorities said. Alvarez is accused of leaving a Jeep Cherokee on the tracks, causing the derailment of one train, which then crashed into another. Juan Miguel Alvarez, ⑦*charged with murder* with special circumstances in the deaths of 11 Metrolink passengers, ⑨*slashed and stabbed himself* after seeing the horrific train crash he caused, sources close to the investigation said Thursday.
R-LTR$_{min}$	A man ⑤*intend on committing suicide* ①*left his car on a railroad track in Glendale* near downtown Los Angeles Wednesday, where it ②*set off a collision that derailed two commuter trains*, ③*killing at least 10 people and injuring nearly 200*, authorities said. The driver of the SUV, identified as Juan Manuel Alvarez, 25, of Compton, Calif., ④*was taken into custody*, and police said ⑦*he would be charged with homicide.* He apparently ⑥*changed his mind about killing himself*, ①*abandoned the vehicle* and watched the crash, the Glendale police chief, Randy Adams, said.

Table VI
COMPARISON ON FACT COVERAGE OVER DIFFERENT APPROACHES

	Facts	Weight	**R-LTR**$_{min}$	SuperLazy	GH	LEAD
②	It set off a collision that derailed two trains	4	√	√	√	-
③	It killed at least 10 people and injured nearly 200	4	√	√	√	-
⑤	He put on suicide watch	4	√	√	√	-
⑥	He intended to commit suicide, but change his mind	4	√	-	√	√
①	He left his car on a railroad track	3	√ √	√	√	-
④	He was arrested	3	√	-	-	-
⑦	He was charged at least 10 counts of homicide	3	√	√	-	√
⑨	He has self-inflicted superficial wounds	3	-	√	-	√
⑧	He has a estranged wife	2	-	-	-	√
⑩	He has criminal records involving drugs	1	-	-	-	-
⑪	He got the wounds after he fled the scene	1	-	-	-	-

2013CB329602, National Natural Science Foundation of China under Grant No. 61232010, No. 61003166, and No. 61203298, and National Key Technology R&D Program under Grants No. 2012BAH46B04, and No. 2012BAH39B02.

REFERENCES

[1] P. Du, J. Guo, and X. Cheng, "Supervised lazy random walk for topic-focused multi-document summarization," in *Proceedings of the 11th ICDM*, ser. ICDM '11, 2011, pp. 1026–1031.

[2] X.-Q. Cheng, P. Du, J. Guo, X. Zhu, and Y. Chen, "Ranking on data manifold with sink points," *IEEE Trans. on Knowl. and Data Eng.*, vol. 25, no. 1, pp. 177–191, Jan. 2013.

[3] T. Ma and X. Wan, "Multi-document summarization using minimum distortion," in *Proceedings of the 10th ICDM*, ser. ICDM '10, 2010, pp. 354–363.

[4] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st ACM SIGIR*, 1998, pp. 335–336.

[5] S. T. Davis, J. M. Conroy, and J. D. Schlesinger, "Occams – an optimal combinatorial covering algorithm for multi-document summarization," *2012 IEEE 12th ICDM Workshops*, vol. 0, pp. 454–463, 2012.

[6] Q. Mei, J. Guo, and D. Radev, "Divrank: the interplay of prestige and diversity in information networks," in *Proceedings of the 16th ACM SIGKDD*, ser. KDD '10, 2010, pp. 1009–1018.

[7] T.-Y. Liu, *Learning to Rank for Information Retrieval*. Springer, 2011.

[8] J. I. Marden, *Analyzing and Modeling Rank Data*. Chapman and Hall, 1995.

[9] X. Li, Y.-D. Shen, L. Du, and C.-Y. Xiong, "Exploiting novelty, coverage and balance for topic-focused multi-document summarization," in *Proceedings of the 19th ACM CIKM*, ser. CIKM '10, 2010, pp. 1765–1768.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, vol. 41, no. 6, pp. 391–407, 1990.

[11] G. Erkan and D. R. Radev, "Lexrank: graph-based lexical centrality as salience in text summarization," *J. Artif. Int. Res.*, vol. 22, no. 1, pp. 457–479, Dec. 2004.

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the 7th WWW*, ser. WWW7, 1998, pp. 107–117.

[13] J. Otterbacher, G. Erkan, and D. R. Radev, "Using random walks for question-focused sentence retrieval," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05, 2005, pp. 915–922.

[14] H. Zha, "Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering," in *Proceedings of the 25th ACM SIGIR*, ser. SIGIR '02, 2002, pp. 113–120.

[15] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999.

[16] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topic-focused multi-document summarization," in *Proceedings of the 20th IJCAI*, ser. IJCAI'07, 2007, pp. 2903–2908.

[17] X. Zhu, A. B. Goldberg, J. Van, and G. D. Andrzejewski, "Improving diversity in ranking using absorbing random walks," in *Physics Laboratory C University of Washington*, 2007, pp. 97–104.

[18] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[19] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[20] D. Metzler and T. Kanungo, "Machine learned sentence selection strategies for query-biased summarization," in *Proceedings of SIGIR Learning to Rank Workshop*, 2008.

[21] Y. Ouyang, W. Li, S. Li, and Q. Lu, "Applying regression models to query-focused multi-document summarization," *Inf. Process. Manage.*, vol. 47, no. 2, pp. 227–237, Mar. 2011.

[22] J. M. Conroy and D. P. O'leary, "Text summarization via hidden markov models," in *Proceedings of the 24th ACM SIGIR*, ser. SIGIR '01, 2001, pp. 406–407.

[23] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in *Proceedings of the 20th IJCAI*, ser. IJCAI'07, 2007, pp. 2862–2867.

[24] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu, "Enhancing diversity, coverage and balance for summarization through structure learning," in *Proceedings of the 18th WWW*, ser. WWW '09, 2009, pp. 71–80.

[25] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd ACM SIGIR*, ser. SIGIR '99, 1999, pp. 50–57.

[26] T. Qin, T.-Y. Liu, J. Xu, and H. Li, "Letor: A benchmark collection for research on learning to rank for information retrieval," *Inf. Retr.*, pp. 346–374, 2010.

[27] D. Metzler and W. B. Croft, "A markov random field model for term dependencies," in *Proc. of the 28th ACM SIGIR*, 2005, pp. 472–479.

[28] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, July 2004, pp. 74–81.

[29] T. H. Haveliwala, "Topic-sensitive pagerank," in *Proceedings of the 11th WWW*, ser. WWW '02, 2002, pp. 517–526.

[30] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.