Comparative Document Summarization via Discriminative Sentence Selection

DINGDING WANG, Florida International University
SHENGHUO ZHU, NEC Laboratories America, Inc.
TAO LI, Florida International University
YIHONG GONG, NEC Laboratories America, Inc.

12

Given a collection of document groups, a natural question is to identify the differences among them. Although traditional document summarization techniques can summarize the content of the document groups one by one, there exists a great necessity to generate a summary of the differences among the document groups. In this article, we study a novel problem, that of summarizing the differences between document groups. A discriminative sentence selection method is proposed to extract the most discriminative sentences which represent the specific characteristics of each document group. Experiments and case studies on real-world data sets demonstrate the effectiveness of our proposed method.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Comparative document summarization, discriminative sentence selection

ACM Reference Format:
Wang, D., Zhu, S., Li, T., and Gong, Y. 2012. Comparative document summarization via discriminative sentence selection. ACM Trans. Knowl. Discov. Data. 6, 3, Article 12 (October 2012), 18 pages.
DOI = 10.1145/2362383.2362386 http://doi.acm.org/10.1145/2362383.2362386

1. INTRODUCTION

Currently, most existing research efforts on document summarization focus on generating a compressed summary delivering the major (or query-relevant) information of the original documents. However, in many applications, when facing a set of document groups sharing similar topics, people are interested to know the differences in these document groups. Thus instead of a generic summary, a summary describing major differences among the given documents is needed to facilitate the comparison of these document groups. For example, there are many recent news articles reporting President Obama's inaugural speech—however, different reports may have different focuses (e.g., some focus on his plan to restore economic growth, some focus on the politics, and there are even some articles that mainly discuss his dress during the inauguration). The news summaries created by traditional summarization methods would all report

This work was partially supported by NSF grants DBI-0850203, HRD-0833093, and DMS-0915110 and DHS grants 2009-ST-062-000016 and 2010-ST-062-000039.
Authors' addresses: D. Wang and T. Li, Department of Computer Science, Florida International University, 11200 SW 8th ST, Miami, FL 33199; email: {dwang003, taoli}@cs.fiu.edu; S. Zhu and Y. Gong, NEC Laboratories America, Inc. 10080 N. Wolfe Rd, SW 3-350, Cupertino, CA 95014; email: {zsh, ygong}@sv.nec-labs.com.

DOI 10.1145/2362383.2362386 http://doi.acm.org/10.1145/2362383.2362386

that President Obama was inaugurated and gave an inauguration speech; however, the different points of view in these articles are also of great interest. Another example is comparing different blog communities and finding the changes in community evolution, for example, the blogs in a blog community discussing the changes in hurricane Katrina from the preparation before the hurricane to the recovery after the hurricane. Thus it is necessary to summarize the changes/differences in different phases of the event.

To the best of our knowledge, the problem of summarizing the distinctness of documents has not been well-defined and studied. Thus in this article, we study the novel problem referred to as *comparative extractive document summarization* (CDS) to summarize the differences between comparable document groups. Specifically, given a collection of document groups, the CDS problem is to generate a short summary delivering the differences of these documents by extracting the most discriminative sentences in each document group. This problem is related to the traditional document summarization problem, since both of them try to extract sentences from documents to form a summary. However, traditional document summarization aims to cover the consensus of information among document collections [Gong and Liu 2001; Wang et al. 2008a; Mani 2001; Baeza-Yates and Ribeiro-Neto 1999], while our goal is to find the differences among different document groups.

A straightforward solution to the CDS problem is to sequentially select sentences from the documents by a greedy approach which minimizes the remaining uncertainty (entropy) of the documents after extracting sentences one by one based on the empirical distribution estimation. However, empirical distribution faces a data sparseness problem. In this article we propose a discriminative sentence selection approach based on a multivariate normal generative model to extract sentences that best describe the unique characteristics of each document group. Given a collection of document groups (clusters), we decompose these documents into sentences and calculate sentence-document and sentence-sentence similarities using cosine similarity. Since each document is labeled to indicate which cluster it belongs to, we select sentences one by one to minimize the average variance of all the cluster targets under the distribution estimated using a multivariate normal generative model. Evaluation on various text data demonstrates the effectiveness and the discriminative ability of the summaries generated by our method.

In summary, our main contributions are (1) we define a novel comparative extractive document summarization problem (CDS); (2) a discriminative sentence selection approach is proposed to solve the CDS problem efficiently; and (3) experiments and case studies of real-world data demonstrate the effectiveness of our proposed method. A preliminary version of this work was published as a four-page paper [Wang et al. 2009a]. This manuscript provides detailed description, in-depth theoretical analysis, and comprehensive experimental results.

The rest of the article is organized as follows. Section 2 discusses the related work. In Section 3, the problem is stated formally. Section 4 describes our discriminative sentence-selection approach in detail and presents an illustrative example. In Section 5, we conduct comprehensive experiments and case studies to demonstrate the effectiveness of our approach. Section 6 concludes.

2. RELATED WORK

2.1. Document Summarization

The work is related to traditional multidocument summarization. The goal of traditional summarization is to generate a summary delivering the major information expressed in a collection of documents.

In general, there are two types of summarization: extractive summarization and abstractive summarization [Knight and Marcu 2002; Jing and McKeown 2000]. Extractive summarization selects the important sentences from the original documents to form a summary, while abstractive summarization paraphrases the corpus using novel sentences, which usually involves information fusion, sentence compression and reformulation [Knight and Marcu 2002; Jing and McKeown 2000]. Although an abstractive summary could be more concise, it requires deep natural language processing techniques. Thus extractive summaries are more feasible and practical, and in the related work, our discussion focuses on extractive document summarization.

The previous extractive summarization research can date back to Baxendale [1958] and Edmundson [1969]. The sentences are ranked according to their scores, calculated based on features (e.g., position, term frequency, and keywords). Recent summarization methods develop more sophisticated approaches to determine the importance of the sentences. The most widely used recent extractive summarization methods are discussed in detail, as follows.

—*Centroid-based methods.* This type of method ranks sentences by computing their salience using a set of features. For example, MEAD [Radev et al. 2004] is a typical centroid-based algorithm which extracts sentences according to three parameters (i.e., centroid value, positional value, and first-sentence overlap). The centroid value of a sentence is computed as the average cosine similarity between the sentences and the rest of the sentences in the document collection. The positional value is computed as follows: the leading sentence is assigned score 1 and the score decreases by 1/n for each sentence, where n is the number of sentences in these documents. The overlap value is computed as the cosine similarity between a sentence and the first sentence in the same document. Then the three values are linearly combined with equal weights.
—*Graph-based methods*. This type of method constructs a sentence graph, in which each node is a sentence in the document collection, and if the similarity between a pair of sentences is above a threshold or the sentences belong to the same document, there is an edge between the pair of sentences. The sentences are selected to form the summaries by voting from their neighbors. Erkan and Radev [2004] propose an algorithm called LexPageRank to compute the sentence importance based on the concept of eigenvector centrality (prestige) which has been successfully used in Google PageRank. Other graph-based summarization have been proposed [Mihalcea and Tarau 2005; Wan and Yang 2008; Shen and Li 2010].
—*Latent semantic analysis (LSA)*. Gong and Liu [2001] propose a method using latent semantic analysis (LSA) to select highly ranked sentences for summarization. The method first creates a term-sentence matrix, where each column represents the weighted term-frequency vector of a sentence in the set of documents. Then singular value decomposition (SVD) is used on the matrix to derive the latent semantic structure. The sentences with the greatest combined weights across all the important topics are included in the summary.
—*Nonnegative matrix factorization (NMF).* This type of method conducts NMF on the sentence-term matrix to extract sentences with the highest probability in each topic [Wang et al. 2008b]. NMF can also be viewed as a clustering method, which has many nice properties and advantages [Ding et al. 2005; Li and Ding 2006]. Intuitively, this method clusters these sentences and chooses the most representative ones from each cluster to form the summary.
—*Other methods*. Other methods include CRF-based summarization [Shen et al. 2007], sentence-based topic models [Wang et al. 2009b], hidden Markov model-(HMM) based method [Conroy and O'leary 2001], and ensemble methods [Wang and Li 2010; Li

and Ding 2008]. Goldstein et al. [1999] propose a method using maximal marginal relevance (MMR) to select sentences by calculating the cosine similarity between a sentence and the document topic and also the sentence and previously selected sentences. This method aims to select a set of sentences having high relevance to the document topic while keeping redundancy low. Some query-based summarization systems are also proposed. For example, the Language Computer Corporation (LCC) [DUC 2006], a DUC participant, proposes a system combining the question-answering and summarization systems and using k-nearest neighbor clustering based on cosine similarity for the sentence selection. There also exists some important work using information extraction and natural language processing techniques to identify representative sentences [Barzilay et al. 1999; Nenkova et al. 2007].

In summary, our CDS work is different from traditional document summarization, as it aims to identify the differences among different document groups. There are also some novel document summarization problems. For example, Allan et al. [2001] propose a temporal summarization approach to monitor the changes of news over time. This type of work aims to catch the novelty during the evolutionary change of documents.

2.2. Topic and Novelty Detection

Topic detection and tracking (TDT) aims to group news articles based on the topics discussed in them, detect some novel and previously unreported events, and track future events related to the topics. Challenging problems in TDT tasks such as event detection, novelty detection, and topic tracking have been studied widely, and information retrieval techniques (e.g., information extraction, filtering, and document clustering), are often applied to these problems [Allan et al. 1998; Brants et al. 2003; Kumaran and Allan 2004; Makkonen et al. 2004; Yang et al. 1998]. For example, Zhang et al. [2007] propose an event detection system based on news indexing trees and term reweighting schemes. In Zhao et al. [2007], events are detected from a social text stream using content, temporal information, and social dimensions. In Zhang et al. [2002], adaptive information filtering is used to decide the novelty and redundancy of documents. In Li and Croft [2006] a novelty detection approach is propsoed that identifies sentence-level information patterns. An algrirthm is proposed in Fung et al. [2007] to extract and partition documents into events and organize them in a hierarchical structure based on a given query. In Morinaga and Yamanishi [2004], a framework is proposed for tracking dynamics of topic trends using a finite mixture model. The novelty detection usually monitors the changes of topics over time. However, in our comparative summarization, we aim to extract the key differences of the documents, which is a more general task than the novelty detection.

2.3. Comparing Documents

There is a limited number of works focusing on comparing documents. Note that, in a multidocument environment, many related documents are likely to share common theme and only differ in certain components. Thus many methods have been developed to identify and synthesize similarities and differences across documents [Carbonell and Goldstein 1998]. For example, Mani and Bloedorn described a method for finding similarities and dissimilarities between related documents by comparing their corresponding graphs representing salient concepts and their relationships [Mani and Bloedorn 1997, 1999]. Ou et al. [2007] proposed a method focused on extracting and integrating similarities and differences using semantic concepts. Instead of identifying the similarities/differences of related documents, our CDS problem aims to generate a short summary delivering the differences among given document groups.

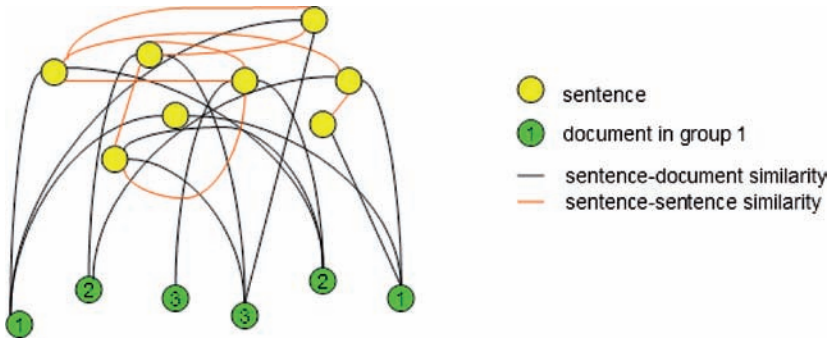

Fig. 1. An example problem.

Some related work has been done in comparing customer reviews [Lerman and McDonald 2009; Paul et al. 2010]. For example, Hu and Liu [2004] use natural language processing techniques to identify opinion words in the reviews and categorize them into positive and negative features. Then opinion sentences are predicted using these features and ranked based on their frequency. Finally, top-ranking sentences are selected to form the summaries straightforwardly. Although the summaries consists of positive/negative sentences, the essence of the work is still based on word-level opinion mining. Another work referred to as comparative text mining (CTM) [Zhai et al. 2004] tries to discover common and specific themes in multiple documents using a generative probabilistic mixture model. The results are listed in a comparison table and keywords are selected to represent the common/specific characteristics of the documents. However, word-level representation has limited interpretative ability and is hard for people to understand naturally. In this article, we analyze sentence features directly by taking into account the sentence-document and sentence-sentence relationships, and the most discriminative sentences are selected to minimize the average variance of the group prediction.

3. PROBLEM FORMULATION

The discriminative sentence selection problem can be described in a formal way, as follows. Suppose we have f sentences of the document collection, denoted by $\{X_i|i \in F\}$, where F is the full sentence index set, having $|F| = f$. We have the group variable, Y, represented by multiple group indicator variables. For example, in Figure 1, there are six documents belonging to three groups (denoted by green nodes). We decompose these documents into eight sentences (denoted by yellow nodes) and a pair of nodes are linked if their cosine similarity is greater than zero. All group variables can be represented by a matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}, \tag{1}$$

where each row represents a group and each column represents a document.

The problem of sentence selection is selecting a subset of sentences, $S \subset F$, to accurately discriminate the documents in different groups, that is, to predict the group identity variable Y, given that the cardinality of S is $m (m < f)$. Let us denote $\{X_i|i \in S\}$ by X_S, for any set S. The prediction capability of Y given X_S can be measured by the entropy of Y given X_S, which is defined as

$$H(Y|X_S) \stackrel{\text{def}}{=} -\mathsf{E}_{p(Y,X_S)}(\ln p(Y|X_S)),$$

Table I. A summary of Notation

$\mathbf{X}$	sentence-document similarity matrix
$\mathbf{Y}$	document group indicator, e.g. Eq. (1)
$\mathbf{K}$	a matrix
$\mathbf{K}_{SR}$	the sub-matrix of $\mathbf{K}$, with row indices S and column indices R
$\mathbf{K}_{sR}$	the sub row vector of $\mathbf{K}$, with row index s and column indices R
$\mathbf{I}_p$	an identity matrix of size $p \times p$
$\mathbf{x}$	a column vector
$\mathbf{x}^\top$	the transposition of vector $\mathbf{x}$
$\mathbf{1}$	a column vector whose elements are all ones
$\|D\|$	the cardinality of set D.
F	the index set in $\mathbf{Z}$ corresponding to the sentences. $\|F\| = f$.
T	the index set in $\mathbf{Z}$ corresponding to the targets (document groups), $\mathbf{Y}$. $\|T\| = t$.
D	the full index set. $\|D\| = d = f + t$.
S	the index set of selected sentences.

where $\mathsf{E}_p(\cdot)$ is the expectation given the distribution p, and p stands for the underlying document distribution, that is, the joint distribution $p(Y, X_s)$. The sentence selection problem using the mutual information criterion is

$$\arg\min_S H(Y|X_S). \tag{2}$$

4. DISCRIMINATIVE SENTENCE SELECTION

A summary of the notation used in this article is shown in Table I.

Selecting an optimal subset of sentences is a combinatorial optimization problem, which is an NP-hard problem. The effective practice is to take a greedy approach, that is, to sequentially select features to achieve a suboptimal solution. Given a selected sentence set, S, the one-step goal of the sentence selection is to select one sentence to minimize the entropy [Zhu et al. 2010]. The one-step objective, named as information gain, is to find i to maximize

$$\mathrm{IG}(i; S) \stackrel{\mathrm{def}}{=} H(Y|X_S) - H(Y; X_{S\cup\{i\}}).$$

The distribution $p(Y, X_S)$ can be estimated by the empirical distribution, that is, measuring the proportion of Y and X_S values in the given data. The empirical distribution faces the data sparseness problem. Thus, we discuss the estimation based on a multivariate normal generative model in our proposed method.

4.1. Multivariate Normal Model

We assume that the joint distribution of $\{X_i\}$ and Y is a multivariate normal distribution,

$$\mathbf{z} = \begin{pmatrix} X_F \\ Y \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix. Let F be the index set of X in $\mathbf{z}$, and T be the index set of Y in $\mathbf{z}$.

We denote the sentence-document similarity matrix by $\widetilde{\mathbf{X}}$, where each row represents a sentence and each column represents a document. For example, the sentence-document similarity matrix can be constructed using the dot product of the sentence-term and term-document matrices which are computed using cosine similarity. We consider multiple target variables. For grouped documents, we denote

the group identity matrix as $\widetilde{\mathbf{Y}}$, where each column represents group identity variables of a document and each row represents a group identity variable.

Given the data, we can estimate the parameters of Eq. (3) by

$$\widehat{\boldsymbol{\mu}} \stackrel{\text{def}}{=} \frac{1}{n}\begin{pmatrix}\widetilde{\mathbf{X}}\\ \widetilde{\mathbf{Y}}\end{pmatrix}\mathbf{1}, \qquad \widehat{\boldsymbol{\Sigma}} \stackrel{\text{def}}{=} \frac{1}{n}\mathbf{Z}\mathbf{Z}^{\top},$$

where n is the number of rows of matrix $\widetilde{\mathbf{X}}$, $\mathbf{1}$ is a column vector of size n, whose elements are all ones, and

$$\mathbf{Z} \stackrel{\text{def}}{=} \begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix}\widetilde{\mathbf{X}}\\ \widetilde{\mathbf{Y}}\end{pmatrix} - \widehat{\boldsymbol{\mu}}\mathbf{1}^{\top}.$$

Next, we consider the sentence-sentence similarity matrix $\mathbf{W}$. For example, the sentence-sentence similarity matrix can be obtained by the product of the standardized sentence-term matrix and its transpose. We use a sentence-sentence similarity matrix to augment the covariance matrix. Also, we add a regularization term to prevent the ill-posed problem of the estimation. Then, we define our covariance matrix by

$$\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}} + \alpha\begin{pmatrix}\mathbf{W} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_t - \frac{1}{t}\mathbf{1}\mathbf{1}^{\top}\end{pmatrix} + \lambda\begin{pmatrix}\mathbf{I}_d & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_t\end{pmatrix}, \tag{4}$$

where $\mathbf{I}$ is the identity matrix of the size of the number of groups, α is a mixture parameter to weigh the importance of the sentence-sentence matrix, and λ is the regularization parameter to increase the robustness. The reason for the lower-right corner of the second term is that we consider that the document group is exclusive.

4.2. Sequential Selection Method

In the multivariate normal model, the sentence selection problem in Eq. (2) becomes

$$\arg\min_{S} \quad \ln\left|\boldsymbol{\Sigma}_{T|S}\right|, \tag{5}$$

where $\boldsymbol{\Sigma}_{T|S} \stackrel{\text{def}}{=} \boldsymbol{\Sigma}_{TT} - \boldsymbol{\Sigma}_{TS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{ST}$, known as the Schur complement, as the determinant of the covariance matrix is known as the *generalized variance*. This criterion is to minimize the generalized variance of the joint distribution of targets.

We have the following property which will be used in the derivation of our algorithm.

PROPERTY 1. *Let D be the full index set of $\mathbf{z}$, $S \subset F$, $i \in F - S$. We have*

$$\boldsymbol{\Sigma}_{T|S\cup\{i\}} = \boldsymbol{\Sigma}_{TT}^{(S)} - \frac{1}{\boldsymbol{\Sigma}_{ii}^{(S)}}\boldsymbol{\Sigma}_{Ti}^{(S)}\boldsymbol{\Sigma}_{iT}^{(S)}, \tag{6}$$

where

$$\boldsymbol{\Sigma}^{(S)} \stackrel{def}{=} \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{DS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SD}. \tag{7}$$

The proof of the above property can be found in Zhu et al. [2010]. We now describe the greedy approach to solve Eq. (5). Let $\mathbf{K} = \boldsymbol{\Sigma}_{D|S}$. By Eq. (6), we have

$$\begin{aligned}\ln\left|\boldsymbol{\Sigma}_{T|S\cup\{i\}}\right| &= \ln\left|\mathbf{K}_{TT} - \frac{1}{K_{ii}}(\mathbf{K}_{Ti}\mathbf{K}_{iT})\right|\\ &= \ln|\mathbf{K}_{TT}| + \ln\left(1 - \frac{\mathbf{K}_{iT}(\mathbf{K}_{TT})^{-1}\mathbf{K}_{Ti}}{\mathbf{K}_{ii}}\right).\end{aligned}$$

Therefore

$$\arg\min_{i} \ln\left|\boldsymbol{\Sigma}_{T|S\cup\{i\}}\right| = \arg\max_{i} \frac{\mathbf{K}_{iT}(\mathbf{K}_{TT})^{-1}\mathbf{K}_{Ti}}{K_{ii}}.$$

We can compute $\mathbf{\Sigma}_{D|S\cup\{i\}}$ from $\mathbf{K} = \mathbf{\Sigma}_{D|S}$ by

$$\mathbf{\Sigma}_{D|S\cup\{i\}} = \mathbf{K} - \frac{1}{K_{ii}}(\mathbf{K}_{Di}\mathbf{K}_{iD}).$$

To sequentially solve Eq. (5), since it is very simple to select just one sentence at one step, we use an easy-to-implement algorithm that iteratively performs the following two steps until the predefined number of sentences have been selected. Algorithm 1 shows the computational procedure. The time complexity of Algorithm 1 is $\mathcal{O}(md^2+mft^2+mt^3)$.

ALGORITHM 1: Discriminative Sentence Selection (DSS)

Input: m: number of selected sentences;
$\mathbf{\Sigma}$: obtained from Eq. (4);
Output: S: selected sentences;
1. $\mathbf{K} = \mathbf{\Sigma}$;
2. $S = \emptyset$;
3. repeat
4. $i = \arg\max_{i\notin S}(\mathbf{K}_{iT}(\mathbf{K}_{TT})^{-1}\mathbf{K}_{Ti})/K_{ii}$;
5. $\mathbf{K} \leftarrow \mathbf{K} - (\mathbf{K}_{Di}\mathbf{K}_{iD})/K_{ii}$;
6. $S \leftarrow S \cup \{i\}$;
7. until $|S| = m$.

The procedure is similar to the sequential algorithm in Yu et al. [2006], which is used to solve transductive active learning problems. The difference is that we compute the minimum determinant, while transductive active learning computes the minimum trace.

4.3. An Illustrative Example

Here we give an example to demonstrate our proposed sentence selection method. In this example, we use the top four largest clusters of documents from the TDT2 corpora. The topics of the four document clusters are as follows: topic 1: Current Conflict with Iraq; topic 2: Monica Lewinsky Case; topic 3: 1998 Winter Olympics; and topic 4 : Asian Economic Crisis. For each topic, one most representative sentence is extracted to describe the topic. Since all these events happened when President Clinton served as president of the United States, some common information about President Clinton is usually extracted using traditional summarization methods. Table II shows the sentences selected by our discriminative sentence selection method, and some keywords and phrases representing for each topic are highlighted.

From Table II, we observe that (1) no sentences describing the general information of all the topics are included; and (2) the selected sentences represent the specific characteristics of each topic, and the generated summaries clearly show the content differences in those topics.

5. EXPERIMENTS

In the experiments, we compare our discriminative sentence selection method with six other typical summarization systems. We examine the effectiveness and the discriminative ability of these methods using two real data sets and various evaluation methods.

5.1. Data Description and Annotation

5.1.1. Data Set. Blog data. The real blog data was collected by NEC in-house blog crawler during 2005 and 2006. Two databases are used by the crawler. The first

Table II. Most Discriminative Sentence Selected by Our Discriminative Sentence Selection Approach for Each Topic

1	In this option, the United States would invade **Iraq**, occupy Baghdad, unseat Saddam, establish a new **Iraqi regime and rid Iraq** of all its weapons of mass destruction and the equipment for rebuilding them.
2	CNN has confirmed independent counsel Ken Starr may ask president Clinton under oath about his **relationship with Monica Lewinsky**.
3	The Games were described Sunday by International Olympic Committee President Juan Antonio Samaranch as having the "best organization" of any **Winter Olympics** in history.
4	The economic miracle has come to an end in the Asian **financial crisis**.

database contains a set of seed blogs, which initially consist of some well-known blogs with a politics focus. For the seed blogs, the crawler continuously aggregates the RSS feeds and their corresponding entries. For each newly crawled entry, its content is analyzed and the hyperlinks embedded in the content are extracted. If an extracted hyperlink points to another entry and that entry belongs to a blog that is not a member of the seed blogs, then that entry and its blog are stored in the second database. The second database is checked regularly to see if any blog in the database meets the criteria to become a new seed blog (the criteria are based on the number of citations and trackbacks from current seed blogs) and if so, that blog is moved to the first database and begins to be crawled continuously [Chi et al. 2007; Ning et al. 2007]. From this data set, we have 407 English blogs with 274,679 entries in 441 days (63 weeks) between July 10th 2005 and September 23rd 2006.

In this experiment, we only use blog entries and remove the comments. Each blog entry only contains a post without comments. Since we involve human effort to create reference summaries for evaluation, we sample 500 entries, which contain 7 communities, as follows:

—*Community 1 War and Terrorist*: discusses the war and conflicts with Iraq;
—*Community 2 Race Issues*: consists of entries on the topic of race and ethnic policy and facts;
—*Community 3 Duke Lacrosse Case*: describes the scandal that started in March 2006 when a black stripper falsely accused three white members of the Duke University's men's lacrosse team of raping her;
—*Community 4 Religion*: mainly focuses on stories about the Christian religion;
—*Community 5 911 Commission*: discusses the national commission on terrorist attacks upon the United States on September 11, 2001;
—*Community 6 China Issues*: contains entries about China's democracy, politics, and economics;
—*Community 7 Hurricane Katrina*: describes the destruction caused by hurricane Katrina in New Orleans in 2005, and discusses the government's behavior in this disaster.

The entries in the seven topics above are evenly distributed. In order to obtain a meaningful subset of the blog entries, we perform topic detection on each community. The entries that focus most on the topics in each community are included in the sample data set.

Cora Data. The Cora data set is provided by McCallum et al. [2000]. This data set consists of the abstracts and references of about 34,000 computer science research

papers. We use 279 abstracts of the papers in information retrieval area and categorize them into four subfields as digital library, extraction, filtering, and retrieval.

5.1.2. Data Annotation. In order to evaluate the quality of the generated summaries by different methods, the blog entries in the Blog data and the abstracts of the papers in the Cora data are manually labeled by three hired human labelers, who are undergraduate and graduate students at the Florida International University. We hired three human labelers instead of one because summarization is a subjective task, and the bias could be reduced by increasing the number of summaries by different annotators. After reading the content of each data set, the human labelers conducted the following two tasks: the first was to create reference summaries to describe the differences between the document clusters, and the length of each summary was no more than 100 words; and the second task was to select all the discriminative sentences in these documents to form a "discriminative sentence set," and there was no limit on the number of sentences. These human-generated annotations were used for our experimental evaluation. In practice, the following instructions were given to the annotators: (1) read the given document clusters, and use a highlighting pen to highlight the all the sentences containing information that will help in understanding the differences among the document clusters; and (2) write a 100-word summary of the highlighted text.

In the evaluation, the machine-generated summaries are also limited to 100 words and compared with all the human-generated summaries, and various statistical metrics are used to examine the performance of different systems.

5.2. Implemented Systems

We implement three types of summarization strategies as baselines and compare the summaries generated by them with the human-created summaries which summarize the major differences of the document groups.

—*Type1*. The first type of strategy generates a short summary for each group of documents in the datasets using the following traditional summarization methods. (1) *Centroid*: a centroid-based method similar to MEAD algorithm proposed in Radev et al. [2004] using centroid value, positional value, and first-sentence overlap as features; (2) *LexPageRank*: a graph-based summarization method recommending sentences by the voting of their neighbors [Erkan and Radev 2004]; and (3) *LSA*: latent semantic analysis on terms by sentences matrix as proposed in Gong and Liu [2001]. For each group of documents in the two data sets, we generate a short summary using each of the three methods. In the experimental results, we denote these methods as Centroid-T1, Lex-T1, and LSA-T1, respectively.

—*Type2*. The second type of strategy uses the above three traditional summarization methods to extract candidate sentences from each group of documents and then selects the sentences from each group that are farthest from the candidates in other groups in terms of cosine similarity. These baselines are executed in an incremental manner. That is, they start from the most representative sentence in the first cluster, and select the sentence in the second cluster which is the furthest from the first one, and then select the sentence from the third cluster which is the furthest from the two selected sentences, and so on. We denote these methods as Centroid-T2, Lex-T2, and LSA-T2 respectively.

—*Type3*. The third type of strategy performs maximal marginal relevance (MMR) [Goldstein et al. 1999] on the two datasets, aims to generate summaries that considers novelty and diversity. MMR usually requires a starting point. Here we use the centroid sentence, which is the sentence to the highest similarity to all the other sentences in each topic, as the starting point, and the sentence selection

is performed for each topic, respectively. We denote this method as MMR in the experiments.

5.3. Evaluation Measures

In the evaluation, we will compare the results by different methods with the human-created summaries using the following evaluation measures.

ROUGE toolkit. To compare with the human summaries, we use ROUGE [Lin and E.Hovy 2003] toolkit (version 1.5.5), which is widely applied by the Document Understanding Conference(DUC) for document summarization performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of reference summaries. Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-SU. ROUGE-N is an n-gram recall computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{ref}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \text{ref}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}, \tag{8}$$

where n is the length of the n-gram, and ref stands for the set of the reference summaries. $\text{Count}_{\text{match}}(\text{gram}_n)$ is the maximum number of n-grams co-occurring in a candidate summary and the reference summaries, and $\text{Count}(\text{gram}_n)$ is the number of n-grams in the reference summaries. ROUGE-L uses the longest common subsequence (LCS) statistics, while ROUGE-W is based on weighted LCS, and ROUGE-SU is based on skip-bigram plus unigram. Each of these evaluation methods in ROUGE can generate three scores (recall, precision, and F-measure). As we have similar conclusions for the three scores, for simplicity, in this article, we only report the average F-measure scores generated by ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-SU to compare our method to other implemented systems. Intuitively, the higher the ROUGE scores, the more similar the two summaries.

Precision. We further examine the sentences selected by different methods with the human-labeled discriminative sentence set, and compute the precision defined as follows:

$$\text{Precision} = \frac{\{\text{selected sentences}\} \cap \{\text{discriminative sentences}\}}{\{\text{selected sentences}\}}.$$

The recall results are not explicitly reported because the entire pool of discriminative sentences is much larger than the number input limit to the systems. However, the recall is implicit in the ROUGE measures.

5.4. Experimental Results

First of all, we compare the comparative summaries generated by different methods with the human-created reference summaries using the ROUGE toolkit. Note that we require that the length of all the summaries is 100-words. In practice, the average length of an English sentence is around 15 words, so we choose 7 sentences as the number of sentences. If the generated summary is over the 100 word limit, the evaluation tool will automatically trunk the summary and keep only 100-words. Table III and Table IV show the ROUGE scores. From the results, we have the following observations. (1) Methods using both Type1 and Type2 strategies perform poorly. For methods in Type1, this is because they may select sentences delivering general information contained in each document group and the redundancy is very high. For methods in Type2, although they try to select sentences highly relevant to each document topic and dissimilar to each other, in many cases general sentences and terms are still hard

Table III. Overall Performance Comparison on Blog Data Using ROUGE

Systems	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
Centroid-T1	0.23856	0.02411	0.20837	0.09343	0.05577
Lex-T1	0.28173	0.02709	0.24551	0.11316	0.06923
LSA-T1	0.21622	0.00853	0.16321	0.08011	0.04127
Centroid-T2	0.27921	0.03315	0.27293	0.11037	0.07255
Lex-T2	0.30323	0.03862	0.29113	0.11920	0.07925
LSA-T2	0.25113	0.02837	0.2921	0.11572	0.07270
MMR	0.34257	0.05461	0.32305	0.11946	0.08378
DSS	**0.57152**	**0.11294**	**0.45763**	**0.16537**	**0.17045**

Table IV. Overall Performance Comparison on Cora Data Using ROUGE

Systems	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	ROUGE-SU
Centroid-T1	0.21922	0.02306	0.19323	0.09040	0.05321
Lex-T1	0.25324	0.02762	0.25951	0.13412	0.06191
LSA-T1	0.21371	0.02319	0.19023	0.10115	0.05287
Centroid-T2	0.24563	0.02433	0.23572	0.11857	0.06011
Lex-T2	0.31420	0.02587	0.27135	0.12933	0.06523
LSA-T2	0.23981	0.02527	0.23623	0.11037	0.06312
MMR	0.31785	0.02590	0.26383	0.11977	0.06503
DSS	**0.53153**	**0.03027**	**0.48649**	**0.25325**	**0.12308**

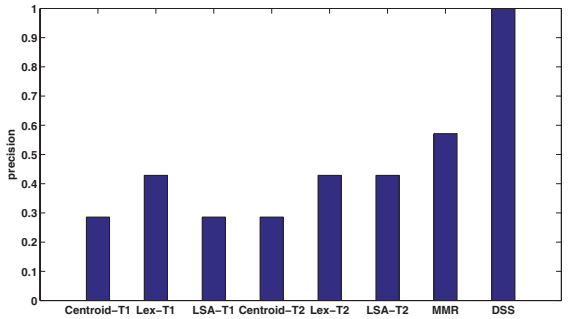

Fig. 2. Comparing the precision of selected discriminative sentences by different methods using Blog data.

to exclude. (2) MMR performs better than methods using Type1 and Type2 strategies because MMR reduces redundancy in the selected sentences, while keeping the potentially relevant information. However, it is not guaranteed that all the topics in the document groups can be covered. (3) Our proposed DSS method outperforms all the other baselines because the sentences selected by DSS can extract the unique features of each topic, and successfully discriminate the document groups.

We also examine the discriminative ability of different methods. We count how many sentences belong to the human-labeled discriminative sentence set and compute the precision scores of the generated summaries by different methods. Figure 2 and Figure 3 show the comparison results. From the results, we observe that (1) for both data sets, over 75% of the sentences extracted by our proposed discriminative sentence selection method belong to the discriminative sentence set; (2) the other implemented baselines can select some discriminative sentences—however, there are still some general sentences which are selected by them.

5.5. A Case Study

A case study is conducted to further examine the results of different summarization methods on Blog data. Each method selects the most representative sentence to

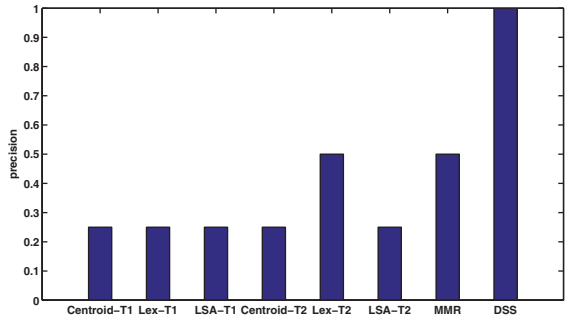

Fig. 3. Comparing the precision of selected discriminative sentences by different methods using Cora data.

Table V. Sentences Selected by Our DSS Approach

	Discriminative Sentence Selection
1	There is no cold war, there is no Saddam. Lebanon has also changed.
2	If hiring rap sheet-free intelligent people means they won't hire a black applicant for another five years.
3	He should drop the case against the lacrosse players but not the sexual assault case itself.
4	Rahman, who is about 41 years old, converted from islam to christianity over 16 years ago.
5	To be totally honest with you, we believed that there may have been a classified annex.
6	I suspect that his position reflects conventional wisdom among the Chinese military establishment.
7	In both the short and long term what those displaced by hurricane Katrina need most is money.

The first column represents the community ID to which the selected sentences belong.

generate a one-sentence summary for each blog community. Table V shows the summaries generated by our DSS approach, and Table VI and Table VII show the sentences selected by methods using Type1 and Type2 strategies and MMR-based summarization, respectively. We also provide an example of a human-generated summary as reference in Table VIII.

From the results, we have the following observations. (1) Some sentences selected by traditional document summarization methods using either Type1 or Type2 strategies are general sentences, which means they may appear in any news articles and do not reflect the specific characteristics of each blog community. For example, the fourth sentence selected by Centroid-T2 is "If you're looking to make relief donations in aggregate try the Red Cross link in the left sidebar." This sentence should describe the religion issues, but we could not obtain any insight of the topic from this sentence. Other general sentences are identified using the "X" in Table VI. (2) Some sentences cover more than one topic, that is, these sentences can be used to describe two or more communities. For example, the first and the fifth sentences selected by Lex-T1 both discuss the terrorists, and neither of them can represent the unique information in the corresponding communities. Similar problems also appear in the results generated by other methods including Lex-T2 (the first and the fifth sentences) and MMR (the third and the sixth sentences).

While looking at the results by our proposed discriminative sentence selection method, each of the sentences represents one community respectively, and the

Table VI. Sentences Selected by Summarization Methods Using Type1 and Type2 Strategies

Centroid-T1		Lex-T1		LSA-T1	
1	For this, he says, the bush team "needs the dying, withering,but still powerful press axis."	1	She has aligned herself now with Michael Moore, who considers those very Iraqi terrorists Minutemen.	1	The Lebanese have also experienced twenty years of syrian occupation and thuggery.
2	I clearly see the sort of depraved thinking and low expectations certain whites have about blacks.	2	If hiring rap sheet-free intelligent people means they wont hire a black applicant for another five years, so be it.	2	The first thing whites do when making these decisions is appeal to the lowest element.
3	There is no current trend of white men raping black women, in other words.	3	I laughed out loud when I read that Dukes president met with "black leaders."	3	If that had been a car full of white boys chasing a black personOh Boy! Ill wrap this up.
4X	He didn't say it was gone; plenty of journalists still heard the call.	4	Afghanistan has already had enough of religious extremism under the Taliban.	4	Purchasing power parity, Pamong other things, a way to countries feel better about them.
5	One person comes to the commission a week prior to the release of the report and says, wait.	5	I discussed this in earlier posts, but it bears repeating: terrorists can change tactics situationally.	5X	This information was published contemporaneously in March 2001 and was in the public domain.
6	Chinese sources seem to raise doubts on the official government version of the story.	6	I have not seen much MSM coverage of this, although it seems like a big deal.	6	If analysts could establish a legitimate reason to investigate a person further, they could keep the corresponding data.
7	While the feds were killing themselves to get poor black people out the city, the white people got ignored.	7X	Several people have asked what, if anything, they could do to help.	7	If you're comfortable with that arrangement I'd like to personally encourage you to donate.
Centroid-T2		Lex-T2		LSA-T2	
1	If the US leads a successful global counter-terror war many of these cadres will turn gray, get fat, and rot	1	For this, he says, the bush team "needs the dying, withering,but still powerful press axis."	1	But Iranian nukes are also risky, and even Russia and China acknowledge that.
2	Research a few predominantly black school districts and let me know what you find.	2	Start with the predominantly black, heavily-funded government school system in our nation's capital.	2X	The system is designed to keep you running in circles so you wont see the real issues.
3	The "duke rape" case has been salaciously splashed all over the news.	3	Anyway, there are some discrepancies about who called 911 and whether the stripper arrived at the party already injured.	3	There is no current trend of white men raping black women, in other words
4X	Now I dont see either statement as useful or wise.	4	Afghanistan has already had enough of religious extremism under the Taliban.	4	Students quickly picked up on this creed, and newsoom culture supported it.
5	This would explain why Atta changed his usual routine and traveled under an assumed identity.	5	Not being on the commission, not being – not working at that level, I had no way of knowing.	5	She added:"We must face the reality that the way we are proceeding now is inherently and in actuality very dangerous".
6	Chinese sources seem to raise doubts on the official government version of the story.	6	China is also moving into the traditional role of institutions like the International Monetary Fund (IMF).	6	A snippet: Even in a country that celebrates free speech, you don't spontaneously vocalize grievances at state events.
7	Billions of dollars in property damage, much of it not covered by insurance.	7	If you're comfortable with that arrangement I'd like to personally encourage you to donate.	7X	But if you listen to the national media, you didn't even know this area existed

The "X" next to the community ID represents that the sentence selected for that community is too general and not discriminative.

Table VII. Sentences Selected by MMR

	MMR
1	Israel may also escalate by striking Syrian intelligence targets throughout the regionCsending the message that supporting proxies can cost the supporter.
2	Flashback: Voter fraud that's legal Flashback: "He thinks illegal aliens can vote?
3	Other college rapes, that is, black-on-white rapes, dont get a fraction of coverage, such as this one or this one.
4	A number of Christian nonprofit groups do humanitarian work in Afghanistan.
5	He's trustworthy and knowledgeable, which so far beats anything the Commission has going for it.
6	He added that China's definition of its territory included warships and aircraft.
7	If you're looking to make relief donations in aggregate try the Red Cross link in the left sidebar.

Table VIII. An Example One-Sentence Human Generated Summary

	A reference one-sentence summary
1	I would prefer everyone to live in a democratic, prosperous community that knows no war or want.
2	I clearly see the sort of depraved thinking and low expectations certain whites have about blacks.
3	Last month two black strippers worked a party given by members of Duke UniversityŠs lacrosse team, and one claims to have been gang-raped and beaten by several white members.
4	Abdul Rahman, 40, was arrested last month, accused of converting to Christianity.
5	Either they lied then, are lying now, or the Commission and their staff have lied.
6	The statement said China was "squeezing" its international diplomatic efforts.
7	After Katrina, we all knew that a huge flow of money would be headed to the Gulf and New Orleans.

specific characteristics of the community are summarized well. In Table V, we highlight some keywords representing the unique features of each topic. We also provide one example of a human-generated summary. Although the content of the sentence sets in Table V and Table VIII are from different points of view on the topics, both of them separate the seven communities well.

6. CONCLUSION

In this article we define and study a novel problem of summarizing the differences of the different document groups to compare their specific characteristics. We propose a novel discriminative sentence selection (DSS) method based on a multivariate normal model. Comprehensive experiments and a case study of real-world data show the effectiveness of our method.

The good results of our DSS method benefit from the algorithm design and the development procedure. In the algorithm design, we use document-sentence representation in which the sentences can be viewed as the features, and the problem of selecting discriminant sentences can be formulated as a sentence-based feature selection problem. In the development procedure, we solve the combinatorial optimization problem based

on the estimation on a multivariate normal generative model, and thus the sequential selection method developed is efficient and effective.

REFERENCES

ALLAN, J., CARBONELL, J., DODDINGTON, G., YAMRON, J., AND YANG, Y. 1998. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. 194–218.

ALLAN, J., GUPTA, R., AND KHANDELWAL, V. 2001. Temporal summaries of new topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. ACM, New York, 10–18.

BAEZA-YATES, R. AND RIBEIRO-NETO, B. 1999. *Modern Information Retrieval*. ACM, New York.

BARZILAY, R., MCKEOWN, K., AND ELHADAD, M. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the ACL*.

BAXENDALE, P. B. 1958. Machine-made index for technical literature: An experiment. *IBM J. Res. Dev. 2*, 354–361.

BRANTS, T., CHEN, F., AND FARAHAT, A. 2003. A system for new event detection. In *Proceedings of the SIGIR'03 Conference*. ACM, New York, 330–337.

CARBONELL, J. AND GOLDSTEIN, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. ACM, New York, 335–336.

CHI, Y., ZHU, S., SONG, X., TATEMURA, J., AND TSENG, B. L. 2007. Structural and temporal analysis of the blogosphere through community factorization. In *Proceedings of the SIGKDD Conference.* ACM, New York.

CONROY, J. M. AND O'LEARY, D. P. 2001. Text summarization via hidden Markov models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. ACM, New York, 406–407.

DING, C., HE, X., AND SIMON, H. 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the SIAM Data Mining Conference*.

DUC. 2006. http://www-nlpir.nist.gov/projects/duc/pubs/.

EDMUNDSON, H. P. 1969. New methods in automatic extracting. *J. ACM 16*, 264–285.

ERKAN AND RADEV, D. R. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the EMNLP*.

FUNG, G. P. C., YU, J. X., LIU, H., AND YU, P. S. 2007. Time-dependent event hierarchy construction. In *Proceedings of the KDD'07 Conference*. ACM, New York, 300–309.

GOLDSTEIN, J., KANTROWITZ, M., MITTAL, V., AND CARBONELL, J. 1999. Summarizing text documents: Sentence selection and evaluation metrics. In *Research and Development in Information Retrieval*, 121–128.

GONG, Y. AND LIU, X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the SIGIR Conference*.

HU, M. AND LIU, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the SIGKDD Conference*.

JING, H. AND MCKEOWN, K. 2000. Cut and paste based text summarization. In *Proceedings of the NAACL Conference.*

KNIGHT, K. AND MARCU, D. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. In *Artificial Intelligence*, 91–107.

KUMARAN, G. AND ALLAN, J. 2004. Text classification and named entities for new event detection. In *Proceedings of SIGIR'04 Conference*. ACM, New York, 297–304.

LERMAN, K. AND MCDONALD, R. 2009. Contrastive summarization: An experiment with consumer reviews. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Companion Volume: Short Papers, 113–116.

LI, T. AND DING, C. 2006. The relationships among various nonnegative matrix factorization methods for clustering. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, Los Alamitos, CA, 362–371.

LI, T. AND DING, C. 2008. Weighted consensus clustering. In *In Proceedings of 2008 SIAM International Conference on Data Mining (SDM)*.

LI, X. AND CROFT, W. B. 2006. Improving novelty detection for general topics using sentence-level information patterns. In *Proceedings of the CIKM'06*. ACM, New York, 238–247.

LIN, C.-Y. AND E. HOVY. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of NLT-NAACL Conference*.

MAKKONEN, J., AHONEN-MYKA, H., AND SALMENKIVI, M. 2004. Simple semantics in topic detection and tracking. *Inf. Retrieval 7*, 347–368.

MANI, I. 2001. *Automatic Summarization*. John Benjamins Co.

MANI, I. AND BLOEDORN, E. 1997. Multi-document summarization by graph search and matching. In *AAAI/IAAI*, 622–628.

MANI, I. AND BLOEDORN, E. 1999. Summarizing similarities and differences among related documents. *Inf. Retrieval 1*, 35–67.

MCCALLUM, A., NIGAM, K., RENNIE, J., AND SEYMORE, K. 2000. Automating the construction of Internet portals with machine learning. *Inf. Retrieval J.* 127–163.

MIHALCEA, R. AND TARAU, P. 2005. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP*.

MORINAGA, S. AND YAMANISHI, K. 2004. Tracking dynamics of topic trends using a finite mixture model. In *Proceedings of KDD'04*. ACM, New York, 811–816.

NENKOVA, A., PASSONNEAU, R. J., AND MCKEOWN, K. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *Trans. Speech Lang. Process. 4,* 2.

NING, H., XU, W., CHI, Y., GONG, Y., AND HUANG, T. S. 2007. Incremental spectral clustering with application to monitoring of evolving blog communities. In *Proceedings of SIAM Data Mining Conference.*

OU, S., KHOO, C., AND GOH, D. 2007. Multi-document summarization focusing on extracting and integrating similarities and differences among documents. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2007)*. 442–446.

PAUL, M.J., ZHAI, C., AND GIRJU, R. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. (EMNLP'10)*. ACL, 66–76.

PETERSEN, K. B. AND PEDERSEN, M. S. 2006. The matrix cookbook. Version 20051003.

RADEV, D., JING, H., STYS, M., AND TAM, D. 2004. Centroid-based summarization of multiple documents. *Inf. Process. Manage.* 919–938.

SHEN, C. AND LI, T. 2010. Multi-document summarization via the minimum dominating set. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. 984–992.

SHEN, D., SUN, J.-T., LI, H., YANG, Q., AND CHEN, Z. 2007. Document summarization using conditional random fields. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)*. 2862–2867.

WAN, X. AND YANG, J. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31 Annual International SIGIR Conference*.

WANG, D. AND LI, T. 2010. Many are better than one: Improving multi-document summarization via weighted consensus. In *Proceedings of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10)*. 809–810.

WANG, D., LI, T., ZHU, S., AND DING, C. 2008a. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. ACM, New York,307–314.

WANG, D., LI, T., ZHU, S., AND DING, C. 2008b. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the SIGIR Conference*.

WANG, D., ZHU, S., LI, T., AND GONG, Y. 2009a. Comparative document summarization via discriminative sentence selection. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. ACM, New York,1963–1966.

WANG, D., ZHU, S., LI, T., AND GONG, Y. 2009b. Multi-document summarization using sentencebased topic models. In *Proceedings of the ACL-IJCNLP Conference*. (Short Paper). 297–300.

YANG, Y., PIERCE, T., AND CARBONELL, J. 1998. A study of retrospective and on-line event detection. In *Proceedings of SIGIR'98 Conference*. ACM, New York, 28–36.

YU, K., BI, J., AND TRESP, V. 2006. Active learning via transductive experimental design. In *Proceedings of the ICML Conference.*

ZHAI, C., VELIVELLI, A., AND YU, B. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the SIGKDD Conference.*

ZHANG, K., ZI, J., AND WU, L.G. 2007. New event detection based on indexing-tree and named entity. In *Proceedings of the SIGIR '07 Conference*. ACM, New York, 215–222.

ZHANG, Y., CALLAN, J., AND MINKA, T. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the SIGIR'02 Conference*. ACM, New York, 81–88.

ZHAO, Q., MITRA, P., AND CHEN, B. 2007. Temporal and information flow-based event detection from social text streams. In *Proceedings of the 22nd National Conference on Artificial Intelligence*. Vol. 2, AAAI Press, 1501–1506.

ZHU, S., WANG, D., YU, K., LI, T., AND GONG, Y. 2010. Feature selection for gene expression using model-based entropy. *IEEE/ACM Trans. Comput. Biol. Bioinform. 7,* 1, 25–36.

Received June 2011; revised April 2012; accepted April 2012