Correlating Summarization of Multi-source News with K-Way Graph Bi-clustering

Ya Zhang
School of Information Sciences & Technology
The Pennsylvania State University
University Park, PA 16802

Xiang Ji
NEC Laboratories America
Cupertino, CA 95014

Chao-Hsien Chu
School of Information Sciences & Technology
The Pennsylvania State University
University Park, PA 16802

Hongyuan Zha
Department of Computer Science & Engineering
The Pennsylvania State University
University Park, PA 16802

ABSTRACT

With the emergence of enormous amount of online news, it is desirable to construct text mining methods that can extract, compare and highlight similarities of them. In this paper, we explore the research issue and methodology of correlated summarization for a pair of news articles. The algorithm aligns the (sub)topics of the two news articles and summarizes their correlation by sentence extraction. A pair of news articles are modelled with a weighted bipartite graph. A mutual reinforcement principle is applied to identify a dense subgraph of the weighted bipartite graph. Sentences corresponding to the subgraph are correlated well in textual content and convey the dominant shared topic of the pair of news articles. As a further enhancement for lengthy articles, a k-way bi-clustering algorithm can first be used to partition the bipartite graph into several clusters, each containing sentences from the two news reports. These clusters correspond to shared subtopics, and the above mutual reinforcement principle can then be applied to extract topic sentences within each subtopic group.

Keywords

bipartite graph, biclustering, news summarization, correlated summarization

1. INTRODUCTION

The fast growth of the World Wide Web has been accompanied by the explosion of online information. Recent studies [14] have showed that there are more than hundreds of millions of web pages and the number is still increasing. How to make the overwhelming amount of information accessible to users is an important task. A lot of research efforts have recently been made on web mining [2], text mining [6], information extraction [3], and information retrieval [8].

Online news represents an important part of online information. Considering the scenario of browsing news online, hundreds or even thousands of news articles may be found on almost any topic/event. They may report the same event or the development of events over time. There is, with no doubt, a high degree of redundancy in information provided by the set of news articles. Some of them may convey a view of the corresponding reporters. This raises an inevitable problem: how can we provide users an quick overview of the complete story? This problem becomes even more crucial with the increase of popularity in accessing the online news with mobile handheld devices such as cell phones and PDAs. Because of the small screen size of a typical handheld devices, presenting the whole lists of news in the user-friendly manner is a nontrivial task. In addition, the low bandwidth wireless channels used for data transfers make another hurdle in data transport layer during user examines the retrieved sets. How to present useful information to handheld users, while keeping the length down to fit into the small screen of handhold devices, is a challenge task. In these cases, it is desirable to design an automatically-generated comprehensive summarization of the contents in a non-redundant way.

In this paper, we tackle the problem of automatic summarization of multi-resource news from multiple sources in a correlated manner. High degree of content redundancy is resulted from correlated contents across news articles. Discovering and utilizing the correlation among them represents an important research issue. However, identifying where the redundancies occur is by no means a trivial task as the relation between sentences in arbitrary documents is difficult to characterize [22]. The correlation between a pair of news articles takes various forms. They may present the same event, or they may describe the same or related events from different points of view. For example, news reports about the same events provided by different web sites, which may include different comments of the reporters on the issue. The proposed algorithm intends to automatically correlate the textual contents of these news reports and highlight their similarities and shared features by extracting sentences that convey shared (sub)topics. This provides readers first step towards advanced summarization as well as helps them understand the multi-source news and reduce the redundancy in information. It also meets the application demand of finding related news articles based on users' query [5; 11] with even finer granularity.

The essential idea of our approach is to apply a mutual reinforcement principle on a pair of news articles to simultaneously extract two sets of sentences. Each set is from one article, and the two sets of sentences together convey non-

trivial shared topics. These extracted sentences are ranked based on their importance in embodying the shared dominant topics, and produce a shallow correlating summarization of the pair of news articles. By selecting different numbers of sentences based on their ranking, we can vary the size of summarization text. In the case that the pair of articles are long and with several shared subtopics, a k-way bi-clustering algorithm is first employed to co-cluster sentences into k clusters each containing sentences from the pair of articles. Each of these sentence clusters corresponds to a shared subtopic, and within each cluster, the mutual reinforcement algorithm can then be used to extract topic sentences.

The rest of the paper is organized as follows. Section 2 reviews some related work. Section 3 introduces the bipartite graph model for a pair of news articles. In Section 4, we present the mutual reinforcement principle to extract sentences that embody the dominant shared topics from the pair of news articles simultaneously. The criteria for verifying the existence of the dominant shared topics is presented as well. Section 5 discusses k-way bi-clustering algorithm that co-clusters sentences in a pair of articles based on shared subtopics. This algorithm is applied to a pair of long articles with more than one shared subtopics. In Section 6, we present detailed experimental results on a set of news articles, and demonstrate that our bi-clustering algorithm and mutual reinforcement algorithm are effective in correlating summarization. We conclude the paper in Section 7 and discuss possible future work.

2. RELATED WORK

With the huge amount of information available online, the Web mining research becomes a converging area from several research communities, such as database, information retrieval (IR), machine learning, and natural language processing (NLP). Basically, Web mining can be classified into three types: Web content mining, Web structure mining, and Web usage mining [13]. But because the Web is huge, diverse, dynamic, unstructured, and redundant, the traditional IR and NLP approaches may not be sufficiently to perform the tasks.

Automated text summarization is an important field of study that covers both content and usability of the Web. Given any text document, automatic text summarization attempts to identify and abstract important information in text and present it to users in a condensed form and in a manner sensitive to the users' or applications' needs [17]. Applications range from snippet generation by search engine to tailoring of the content to suit specific displays such as PDA. A comprehensive survey on text summarization has been provided by Marcu in [18]. There are generally two types of tasks in text summarization: single-document summarization [20] and multiple-document summarization [15], according to the number of documents for summarization. In many instances, summarizing a collection of documents is a more difficult task than summarizing an individual document because the former needs to address additional challenges such as the inconsistence in contents [18], difference in subtopics, and the form of the abstraction.

Many efforts have been performed on clustering and summarizing multiple documents [7; 22; 24]. These approaches exploit meaningful relations among documents based on the analysis of text cohesion and the context in which the comparison is desired. Providing summarization for web pages is a nature application of multi-document summarization techniques. Systems [19; 23] were developed to perform web pages clustering and generic summarization based on relevant results from a search engine. Centroid words are used to find the most salient themes in a cluster of documents. Techniques particularly targeted on news articles have also been proposed in [4].

Multi-document summarization techniques partially address the challenges in summarizing multi-source news. The summary generated by these approaches, however, is usually consist of the most salient information gleaned from all the documents. They may leave out some important information which is enclosed in some but not all the documents. Techniques to generate non-redundant summarization [1; 9] alleviate the problem, but they fail to differentiate the shared (sub)topics from the un-shared ones. To provide a comprehensive summarization of the shared and un-shared (sub)topics, we proposed to generate a correlated summarization of the documents. Our method is different from the above approaches by presenting similarities and differences in content of news articles, rather than a generic salient theme for the collection of news. The correlated summarization provides more explicit contextual information to users than most existing text summarization approaches and directly help users navigate through text and quickly obtain desired information. There have been very limited research in correlated summarization. Mani et al. [16] address the problem of summarizing the similarities and differences of documents by analyzing the relations of meaningful text units. An algorithm developed for summarizing customer reviews [12] takes a correlated approach. The comments are extracted from a set of reviews on a same product and classified as positive or negative opinions. The final summarization is generated from these comments.

3. BIPARTITE GRAPH MODEL

For our purpose of correlated summarization, each news article is viewed a consecutive sequence of sentences. First, as the preprocessing, text is tokenized, generic stop words are eliminated, and words are stemmed [21]. The text in each news article is split into sentences and each sentence is represented with a vector. Each entry in a vector is the occurrence of the corresponding word in the corresponding sentence. An article can further be represented with a sentence-word count matrix, whose rows corresponding to sentences in the text, and whose columns corresponding to words in the text after stop word deletion and stemming.

The correlation between two news articles may be represented with a bipartite graph (see Figure 1), which includes two components: two types of nodes and edges connecting nodes of different types. The weights on the edges are nonnegative. In the context of summarizing a pair of news articles, the sentences are modelled as the nodes in the bipartite graph. Nodes corresponding to an article are of the same type, and they are of different types if they are from different articles. The edges connect two nodes corresponding to sentences from different articles. The similarity between the two sentences is expressed as the edge weights of the bipartite graph. Suppose there are m sentences in article A and n sentences in article B. We denote the two

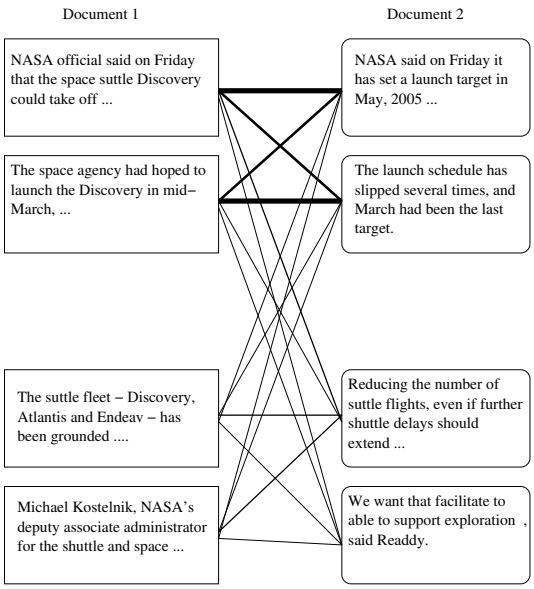

Figure 1: A bipartite graph representation of a correlated documents. The width of the edges correspond to the edge weights.

types of nodes $A = \{a_1, a_2, ..., a_m\}$ and $B = \{b_1, b_2, ..., b_n\}$, where a_i $(i = 1, 2, ..., m)$ and b_j $(j = 1, 2, ..., n)$ are vector representation of sentences. The edge weights $W = [w_{ij}]$ $(i = 1, 2, ..., m,\ j = 1, 2, ..., n)$ between node a_i and b_j are computed as the pairwise similarities between the sentence a_i in A and the sentence b_j in B. The weighted bipartite graph of the news articles is then denoted as $G(A, B, W)$, where W is a m-by-n matrix.

4. THE MUTUAL REINFORCEMENT PRINCIPLE

Suppose the two news articles in question is presented in a bipartite graph $G(A, B, W)$. The dominant shared topic of the article A and the article B is embodied by two subsets of sentences, X and Y, while $X \subseteq A$ and $Y \subseteq B$. Suppose news articles A and B are related in topic. Sentences in X then tend to correlate well with sentences in Y, and they together embody the dominant shared topic of the pair of news articles. In the bipartite graph model, the nodes corresponding to X and Y form a dense subgraph. We use the mutual reinforcement principle to extract this dense subgraph from the bipartite graph.

We associate each sentence a_i in A a saliency score $u(a_i)$ to indicate its importance in the delivery of the topic. High saliency score implies the sentences is the key to the topic. Similarly, the saliency score is assigned to each sentence b_i in B, denoted as $v(b_i)$. Those sentences that embody the dominant shared topic of the two articles should have high saliency scores. It is easy to see that the following observation are valid.

> A sentence in A are topic sentences if it is highly related to many topic sentences in B, while a sentence in B are topic related if it is highly related to many topic sentences A.

In another words,

> the saliency score of a sentence in news article A is determined by the saliency scores of the sentences in B which are correlated to the sentence in A, and the saliency score of a sentence in article B is determined by the saliency scores of the sentences in A which are correlated to the sentences in B.

Mathematically, the above statement is rendered as

$$u(a_i) \propto \sum_{v(b_j)\sim u(a_i)} w_{ij} v(b_j),$$
$$v(b_j) \propto \sum_{u(a_i)\sim u(a_i)} w_{ij} u(a_i),$$

where $a_i \sim b_j$ represents that there is an edge between vertices a_i and b_j, the symbol $\propto$ stands for "proportional to". Now we collect the saliency scores for sentences into two vectors u and v, respectively, the above equation can then be written in the following matrix format

$$u = \frac{1}{\sigma} W v, \quad v = \frac{1}{\sigma} W^T u,$$

where W is the weight matrix of the bipartite graph of the article in question, W^T stands for the matrix transpose of W, and $1/\sigma$ is a proportionality constant. u and v are the left and right singular vectors of W corresponding to the singular value σ. If we choose σ to be the largest singular value of W, then both u and v have *nonnegative* components. The corresponding component values of u and v are then the sentences saliency scores of articles A and B, respectively. Sentences with high saliency scores are selected from the sentence set A and B. These sentences and cross edges among them construct a dense subgraph of the original weighted bipartite graph, which are well correlated in textual content.

A direct question raised is to determine the number of sentences included in the dense subgraph. We first reorder the sentences in A and B according to their corresponding saliency scores to obtain a permuted weight matrix $\hat{W}$. For $\hat{W} = [\hat{W}_{s,t}]_{s,t=1}^2$, we compute the quantity of

$$J_{ij} = h(\hat{W}_{11})/n(\hat{W}_{11}) + h(\hat{W}_{22})/n(\hat{W}_{22})$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. $\hat{W}_{11}$ is the i-by-j submatrix at the upper left of $\hat{W}$, $\hat{W}_{22}$ is the $(m-i)$-by-$(n-j)$ submatrix at lower right, and $h(\cdot)$ is the sum of all the elements of a matrix and $n(\cdot)$ is the square root of the product of the row and column dimensions of a matrix. We then choose first i^* sentences in A and first j^* sentences in B such that

$$J_{i^* j^*} = \max\{\ J_{ij} \mid i = 1, \ldots, m, j = 1, \ldots, n\}.$$

The i^* sentences from A and the j^* sentences from B are considered as sentences in articles A and B that most closely correlate with each other. A post-evaluation step is performed to verify the existence of shared topics in the pair of articles in question. Only when the average cross-similarity density of the sub-matrix $\hat{W}_{11}$ is greater than a certain threshold, we say that there is shared topic between the pair of articles and the extracted i^* sentences and j^* sentences efficiently embody the dominant shared topic. If article A and B present unrelated event or object, then the average cross-similarity density of extracted sub-graph is below the threshold. This sentence selection criteria avoids local maximum solution and extremely unbalanced bipartition of the graph.

The implementation of the mutual reinforcement principle contains left and right singular vectors computing. The singular vectors can be obtained by Lanczos bidiagonalization process [25], which is an iterative process each involving the two matrix-vector multiplications. The computation cost is proportional to the number of nonzero elements of the object matrix W, $nnz(W)$. The major computational cost in the implementation of the mutual reinforcement principle comes from computing the largest pair of singular vectors and maximizing J_{ij}. The cost of selecting the maximum value of J_{ij} is proportional to the number of entries in the matrix, which is greater than or equal to the $nnz(W)$. Therefore, the total cost of mutual reinforcement principle for topic sentences selection is proportional to the number of entries in the matrix.

5. K-WAY GRAPH BI-CLUSTERING

The above approach usually extracts a dominant topic that is shared by the pair of news articles. However, the two articles may be very long and contain several shared subtopics besides the dominant shared topic. Each shared subtopic may be presented at various positions in a long article as well. To extract these less dominant shared topics, a k-way bi-clustering algorithm is applied to the weighted bipartite graph introduced above before the mutual reinforcement principle is used for shared topic extraction. The k-way bi-clustering algorithm will divide the bipartite graph into k subgraphs. Each subgraph contains sentences from both of the two articles and corresponds to one of shared subtopics if available. Within each subgraph, we then apply the mutual reinforcement principle to extract topic sentences.

Given the bipartite graph $G(A, B, W)$ representing the correlated news articles A and B, the k-way bi-clustering algorithm partition the graph into

$$A = A_1 \cup A_2 \cup \cdots \cup A_k,$$

and

$$B = B_1 \cup B_2 \cup \cdots \cup B_k,$$

where A_i and B_i form a dense sub-graph. We define vectors Ia_i of length m and Ib_i of length n as the component indicators of A_i in A and B_i in B, respectively. The elements of Ia_i and Ib_i are 1 for those corresponding to the sentences in A_i and B_i, respectively. And the rest elements of Ia_i and Ib_i are 0. That is,

$$Ia_{ij} = \begin{cases} 1 & \text{if } a_j \in A_i \\ 0 & \text{otherwise} \end{cases},$$

and

$$Ib_{ij} = \begin{cases} 1 & \text{if } b_j \in A_i \\ 0 & \text{otherwise} \end{cases}.$$

A pair of sentences a and b is considered matched with respect to the partition $P(A, B)$ if $a \in A_i$ and $b \in B_i$ for some $1 \le i \le k$. Hence, sentences in the same bicluster are about the same sub-topic. Intuitively, the desired partition should have the following property: the similarities between sentences in A_i and sentences in B_i are as high as possible, and the similarities between sentences in A_i and sentences in B_j $(i \neq j)$ are as less as possible. This would give rise to partitions with closely similar sentences concentrated between all A_i and B_i pairs. In our context of extracting shared subtopics, this strategy leads to the desired tendency of discovering subtopic bi-clusters $A_1 \cup B_1$, $A_2 \cup B_2$, ..., $A_k \cup B_k$. Now we introduce the Normalized Cut algorithm we used for the graph bi-partition[10].

The k-way Normalized Cut to find a partition $P(A, B)$ aims at minimizing the objective function

$$Ncut(G(A, B, W)) = \frac{w(V_1, V_1^c)}{w(V_1, V)} + \frac{w(V_2, V_2^c)}{w(V_2, V)} + \cdots + \frac{w(V_k, V_k^c)}{w(V_k, V)} \quad (1)$$

where $V_i = A_i \cup B_i$ $(i = 1, 2, ..., k)$, $V = A \cup B$, and $w(V_i, V_j)$ is the summation of weights between vertices in subgraph V_i and vertices in subgraph V_j.

$$w(V_r, V_r^c) = \sum_{j \neq r} Ia_r^T W Ib_j + \sum_{i \neq r} Ia_i^T W Ib_r$$

$$w(V_r, V) = \sum_{j=1}^{k} Ia_r^T W Ib_j + \sum_{i=1}^{k} Ia_i^T W Ib_r$$

Let

$$Z = \begin{pmatrix} \hat{Ia}_1 & \dots & \hat{Ia}_k \\ \hat{Ib}_1 & \dots & \hat{Ib}_k \end{pmatrix}.$$

Z is column orthogonal. The bipartite Normalized Cut problem can then be simplified to

$$min_{P(A,B)}(Ncut(G(A, B, W)))$$

$$= min_Z \left(k - trace \left(Z^T \begin{pmatrix} 0 & \hat{W} \\ \hat{W}^T & 0 \end{pmatrix} Z \right) \right).$$

The relax problem of the minimizing problem can be solved by setting columns of Z to be any orthonormal basis for the subspace spanned by the columns of the matrix $\begin{pmatrix} U \\ V \end{pmatrix}$, where $U = (u_1, u_2, \ldots, u_k)$ and $V = (v_1, v_2, \ldots, v_k)$ are the left and right singular vectors corresponding to the k largest singular value of $\hat{W}$, respectively.

Given the weight matrix of the edges, the Ncut algorithm partitions the graph according to the following steps:

1. Compute D_A and D_B with $We = D_A e$ and $W^T e = D_B e$, respectively. Calculate the scaled weight matrix $\hat{W} = D_A^{-1/2} W D_B^{-1/2}$.
2. Compute the left and right singular vectors, U and V, of $\hat{W}$ correspond to the k largest singular value.
3. Calculate $U = D_A^{-1/2} \hat{U}$ and $V = D_B^{-1/2} \hat{V}$ respectively, and then permute the matrix W based on U and V.
4. Partition the rows of U and V into k submatrices by minimizing the Ncut score.

6. EXPERIMENTAL RESULTS

We downloaded 20 pairs of news articles from Google News[1]. Each pair of the news articles are about the same topic according to Google News. These news are generally fall into the categories of IT news, business new and world news. We label them *it* as IT news, *buz* as business news, and *wld* as world news. The overflow of our experiments is summarized as follows:

[1] http://news.google.com

Article Pair	Average Similarity of Extracted Sentences	Precision	Recall
$it11$ & $it12$	3.250	0.833	0.714
$it21$ & $it22$	1.337	0.667	0.800
$it31$ & $it32$	2.717	0.750	0.857
$it41$ & $it42$	1.450	0.667	0.571
$buz11$ & $uz12$	1.234	0.714	0.833
$buz21$ & $buz22$	0.774	0.667	0.667
$buz31$ & $buz32$	0.712	0.625	0.714
$wdl11$ & $wld12$	1.328	0.571	0.667
$wdl21$ & $wld22$	1.257	0.667	0.750
$wld31$ & $wld32$	1.112	0.80	0.667
$buz11$ & $it12$	0.554	N/A	N/A
$buz11$ & $wld12$	0.272	N/A	N/A
$wdl11$ & $it22$	0.318	N/A	N/A

Table 1: Dominant shared topic extraction from pairs of news articles

1. Text pre-processing: we take a pair of concatenated news articles from the corpus, parse them into sentences, then delete words appearing in a stop word list. Porter's stemming [21] is applied to the words.

2. Calculate sentence-sentence similarities across the pair of news articles. In all the experiments, cosine similarity is used as the measurement of sentence similarity. The matrix resulted is the weight matrix in the bipartite graph model.

3. Use k-way biclustering to partition sentences in the news articles into groups, which correspond to subtopics.

4. For each subtopic group, use mutual reinforcement principle to extract the key sentences.

To our best knowledge, there is no standard measures defined for evaluating correlated summarization methods. Some researchers use human-generated summarization or extraction to evaluate their algorithms. But most manual sentence extraction and correlation results tend to differ significantly, especially for long articles with several subtopics. Hence, the performance evaluation for correlating summarization of a pair of news articles becomes a very challenging task. We try to provide some quantitative and qualitative evaluation on our algorithms based on our experimental results with the news articles, each generated through combining two news articles of different subjects. We report the experimental results in two sections. The first section demonstrates the mutual reinforcement principle for extracting the dominant shared topic of a pair of news articles, and the second section demonstrates the bi-clustering algorithm for sentences co-clustering based on subtopics sharing.

6.1 The Mutual Reinforcement Principle to Extract Topic Sentences

We extract the shared topic of each pair of news articles with mutual reinforcement principle. Sentences that embody the dominant shared topic were also manually chosen for evaluation purposed. We then use two standard measure metrics of Information retrieval, *precision* and *recall* to evaluate our method. The automatically extracted sentences for each news article is seen as the retrieved sentences, and the manually extracted sentences for each news article is seen as the relevant sentences. The precision measure is defined as the fraction of the retrieved sentences that are relevant, and the recall measure is the fraction of the relevant sentences that are retrieved. Higher values of precision and recall usually indicate more accurate algorithms for key sentence extraction. Table 1 lists our experimental results for key sentence extraction. For the purpose of comparison, we also tried to find the dominant shared topic between two un-correlated news articles. The average cross-similarities between extracted sentences are computed to determine an empirical cutoff for correlated topics. The threshold was determined to be 0.7. The score for the pairs of news articles where there is no shared topic, are listed in the last three row of Table 1.

As demonstration for mutual reinforcement principle to extract dominant shared topic of a pair of news articles, we present the experimental results of two pairs of news articles: $it1$ and $it2$, and $wld1$ and $wld2$. The mutual reinforcement principle is applied to each pair of news articles, and sentences with high saliency scores are extracted from each pair of news articles. Two sentences with top two saliency scores are marked in bold font in Table 2 for each pair of articles. In the first example, both news articles are about Apple introducing iPod U2 special edition. The mutual reinforcement principle has successfully extracted the key sentences from each articles. The average cross-similarity among the extracted sentences is 2.845, which is greater than our empirical threshold 0.7. This indicates mathematically that there is a shared topic between the pair of articles and the extracted sentences are efficient in embodying the shared topic. Similar case happens in the second example.

6.2 Bi-clustering to Find Shared Subtopics

We now present the results of our experiments on bi-clustering algorithm for sentences co-clustering to reveal shared subtopics. Due to the lack of labelled data, we generated our own news collection where each article is a concatenation of two news articles. We use the concatenated news articles to simulate news articles of multiple shared subtopics. We then apply the k-way bi-cluster algorithm to the long news articles to group sentences with shared topics. Sentences from the same pair of news articles should be put into a bicluster.

*it*1	*it*2
1 **Apple today introduced the iPod U2 Special Edition as part of a partnership between Apple, U2 and Universal Music Group (UMG) to create innovative new products together for the new digital music era**. 2 The new U2 iPod holds up to 5,000 songs and features a gorgeous black enclosure with a red Click Wheel and custom engraving of U2 band member signatures. 3 U2 and Apple have a special relationship where they can start to redefine the music business, said Jimmy Iovine, Chairman of UMG's Interscope Geffen A&M Records. 4 The iPod along with iTunes is the most complete thought that we've seen in music in a very long time." 5 "We want our audience to have a more intimate online relationship with the band, and Apple can help us do that," said U2 lead singer Bono. 6 "With iPod and iTunes, Apple has created a crossroads of art, commerce and technology which feels good for both musicians and fans." 7 **"iPod and iTunes look like the future to me and it's good for everybody involved in music," said U2 guitarist, The Edge.**	1 **Apple Computer has unveiled two new versions of its hugely successful iPod: the iPod Photo and the U2 iPod**. 2 Apple also has expanded its iTunes Music Store to nine more European markets, evidence that the company intends to maintain its global hegemony in the portable-music-player and digital-download markets. 3 The announcements came Oct. 26 at a media event that featured a keynote by Apple CEO Steve Jobs and a live performance by U2's Bono and the Edge, who played songs from the band's forthcoming Island album, "How to Dismantle an Atomic Bomb." 4 The high-end 60GB iPod Photo ($599) offers some innovations. 5 In addition to its music capacity, it has another 20GB of memory – making it the largest-capacity iPod – and a high-resolution color screen to display album art and other digital images. 6 It also comes in a 40GB version ($499). 7 The limited-edition, 20GB black U2 iPod ($149), features custom engraving of the band members' signatures, plus coupon discounts on a "digital boxed set" of the band's catalog and rare tracks available exclusively through iTunes. 8 Apple's European Union iTunes Music Store rolled out in Austria, Belgium, Finland, Greece, Italy, Luxembourg, the Netherlands, Portugal and Spain. 9 It features more than 700,000 songs from all four major labels and more than 100 independents. 10 The latest territories join iTunes stores in the United Kingdom, France, Germany and the United States. 11 The computer giant will launch iTunes in Canada next month. 12 **Apple claims that iPods represent 65% of portable-music-player sales and that iTunes represents 72% of all digital downloads.**
*wld*1	*wld*2
1 **The timing of Osama bin Laden's late-breaking appearance on America's Friday evening newscasts seemed, at first blush, to leave little time for reflection before Tuesday's election.** 2 Republicans, reacting privately, tended to believe the visceral reminder of the war on terror - President Bush's strongest issue - would work to their advantage. 3 Democrats, also speaking privately, worried that this jolt had the potential to tip some undecided voters the president's way, in a race where relatively few votes in key states could determine the outcome. 4 **If nothing else, the bin Laden tape abruptly changed the conversation, away from an issue that had put the Bush administration on the defensive all last week.** 5 Iraq's missing explosives - to one that put Bush back in comfortable territory and speaking in firm absolutes: "I will never relent in defending our country, whatever it takes,"' Bush said in a weekend campaign appearance.	1 **OSAMA bin Laden appears to have boosted George Bush's US election campaign.** 2 Dubya last night held a six-point lead over challenger John Kerry in the first polls since the al-Qaeda chief's video threat. 3 Fifty per cent of likely voters said they backed the Republican president as the Democrat's support slipped to 44 per cent. 4 They were tied in the same poll a week ago and Kerry held a one per cent lead in a survey on Friday. 5 If bin Laden planned to humiliate Bush before tomorrow's election, it backfired. 6 Whenever the subject of the campaign has turned to terrorism, it has benefitted Bush,' said Newsweek. 7 In every poll, voters have said they trust Bush more than Kerry to handle terrorism and homeland security. 8 **Bin Laden's tape was so well-timed for Bush some suggested the Republicans were behind it.** 9 Former TV news anchorman Walter Cronkite joked he was inclined to think the political manager at the White House probably set up bin Laden to this thing. 10 Bush ally Senator John McCain said: 'It's very helpful to the president. 11 I'm not sure if it was intentional or not, but I think it does have an effect. 12 Bush did not mention the tape specifically during campaign appearances over the weekend.

Table 2: Extracted key sentences that embody the dominant shared topics of news article pairs

	*news*1	*news*2
subtopic 1	1 **China will take more measures to protect the country's deteriorating environment to contribute to the sustainable development of both China and the world, Chinese Vice Premier Zeng Peiyan said here Sunday.** 2 "China's economic and social development is facing increasingly heavy pressure from environment and nature resources, marked by pollution and environmental degradation" Zeng said. 3 He pledged that the country will strive to build an environment-friendly society by further strengthening its legal framework, relying on scientific advancement and strongly promoting recycling economy. 4 The vice premier outlined four major tasks in combating environment degradation for the country, including adjusting economic structures, curbing pollution in rivers and lakes, shutting down heavily-polluted enterprises and enhancing international cooperation. 5 Zeng made the remarks at the closing ceremony of the annual meeting of the China Council for International Cooperation on the Environment and Development, the government's international advisory body. 6 The council consists of 50 senior government officials and experts from China and other countries and representatives of international organizations, including the United Nations Environment Program. 7 **During the three-day meeting, members of the council focused their discussions on sustainable agricultural and rural development and drafted a report of recommendations to the Chinese government.** 12 *Tikrit is about 80 miles north of Baghdad and is the hometown of former Iraqi leader Saddam Hussein.*	1 **Vice-Premier Zeng Peiyan has called on experts from around the world to help the Chinese Government develop the nation's environment protection, reported China Daily on Monday.** 2 Zeng said he wanted the experts to pay close attention to China's 11th Five-Year Plan (2006-10). 3 "The role of the council should be given full play," he told delegates yesterday attending the third meeting of the Third Phase of the China Council for International Co-operation on Environment and Development. 4 **The council, established by the State Council of China in 1992, is a high-level non-governmental advisory body that aims to strengthen co-operation and exchanges between China and the international community on environmental and development issues.** 5 Zeng is currently the council's chairman. 6 The focus of this year's meeting, which opened on Friday, was sustainable agriculture and rural development. 7 Recommendations proposed by the council on sustainable agriculture and rural development are very helpful to the Chinese Government, Zeng said. 8 "Such suggestions include creating a new rural development framework in China, which involves an increase of investment in education and science and technology in rural areas," said Xie Zhenhua, minister of the State Environmental Protection Administration and also a vice-chairman of the council. 13 *Dr al-Juburi said he did not know the cause of the blast.* 14 *A police official confirmed the incident and said the explosion might have been caused by a projectile.* 16 *Tikrit, about 80 miles north of Baghdad, is the hometown of former leader Saddam Hussein.*
subtopic 2	8 **Authorities said at least 15 people are dead after an explosion hit a hotel in the northern Iraqi city of Tikrit.** 9 Police and hospital officials said eight others were seriously wounded in Sunday's attack at the Sunubar Hotel. 10 **All the victims are said to be Iraqis – including two police officers.** 11 It's not clear what caused the blast, although one police official said it might've been caused by a projectile.	9 **An explosion in a hotel in the northern Iraqi town of Tikrit has killed 15 Iraqis, police and hospital officials said, bringing the day's total death toll across the country to 25.** 10 Dr Hassan al-Juburi, director of the Tikrit Teaching Hospital, said the blast tore through the Sunubar Hotel at 8:00pm (1700 GMT) on Sunday. 11 **Eight others were seriously wounded in the explosion, including two policemen.** 12 All the victims were Iraqis, he said. 15 He had no further details.

Table 3: Biclustering of the pair of news articles and extracting key sentences that embody the shared subtopics.

Article Pair	Accuracy for #1	Accuracy for #2
$it11it21$ & $it12it22$	0.857	0.917
$it21it41$ & $it42it22$	0.864	0.840
$buz11buz21$ & $buz12buz22$	0.871	0.790
$buz21buz31$ & $buz32buz22$	0.786	0.923
$wld11wld21$ & $wld12wld22$	0.917	0.813
$wld21wld31$ & $wld32wld22$	0.871	0.889
$it41wld11$ & $it41it11$	0.864	0.870
$it11buz21$ & $wld22buz11$	0.840	0.889
$it11it31it41$ & $it12it32it42$	0.917	0.923
$buz11buz21buz31$ & $buz22buz32buz12$	0.740	0.857
$wld11wld31wld21$ & $wld22wld12cwld32$	0.815	0.917
$it11wld31wld41$ & $it42it12wld32$	0.703	0.716
$it21it11buz21$ & $it12it22wld32$	0.729	0.851

Table 4: Bi-clustering pairs of new articles

An example of biclustering the concatenated news articles $news1$ ($wld11$ and $wld21$) and $news2$ ($wld12$ and $wld22$) is presented in Table 3. The sentences in the wrong bicluster are in italic, and the key sentences extracted with mutual reinforcement principle are labelled in bold font.

Quantitative results of some experiments are listed in Table 4. Each row of the table corresponds to one experiment. Some news articles are produced by concatenating news articles from the same fields, others are produced by concatenating news articles from different fields. The number of subtopics within a long news article are related to the number of news articles being concatenated together. Higher accuracy represents that higher percentage of sentences in a article are correctly clustered. The results in Table 4 demonstrate that the bi-clustering algorithm is effective in co-clustering sentences based on shared subtopics. When varying the number of shared subtopics of a pair of news articles, the performance of the algorithm is stable. However, when the ratio of the number of shared subtopics to the number of subtopics in a article decreases, the accuracy tends to deteriorate. In experiments listed in the last two rows of Table 3, there are three subtopics in each long news article. Only two of them are shared between the pair of news articles. The accuracies in the experiments are lower than those in other groups of experiments.

7. CONCLUSIONS AND FUTURE WORK

The text in a web page usually addresses a coherent topic. However, a web page with long text could address several subtopics and each subtopic is usually made of consecutive sentences. Thus, it is necessary to segment sentences into topical groups before studying the topical correlation of multiple web pages. There has been many text segmentation algorithms available.

In this paper, we propose a new procedure and algorithm to automatically summarize correlated information from online news articles. Our algorithm contains the mutual reinforcement principle and the bi-clustering method. The mutual reinforcement principle extracts sentences, which convey the dominant shared topic, from a pair of news articles simultaneously. The bi-clustering algorithm, which co-clusters the sentences in a pair of articles, is employed to find the shared subtopics of them. We test our algorithm with news articles in different fields. The experimental results suggest that our algorithms are effective in extracting dominant shared topic and/or subtopics of a pair of news articles.

This study has two major contributions: (1) we bring up the research issue of correlated summarization for news articles, and (2) we present a new algorithm to align the (sub)topics of a pair of news articles and summarize their correlation in content. The proposed method can automatically extract, correlate, and summarize information to prevent users from information overload. It also allows user to use small-screen mobile handheld devices to read news.

The proposed algorithms could be improved to handle more than two news articles simultaneously. Another research direction boosted by NIST, known as Topic Detection and Tracking[2], is to discover and thread together topically related material in streams of data [26]. Our algorithm may also be applied to generating a completed story line from a set of news articles about the same event over time. The method is applicable to correlated summarize multilingual articles, considering the growing volume of multilingual documents online.

8. ACKNOWLEDGMENT

The authors want to thank the editors and the anonymous referees for many constructive comments and suggestions.

9. REFERENCES

[1] J. G. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking or reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336, 1998.

[2] R. Cooley, J. Srivastava, and B. Mobasher. Web mining: Information and pattern discovery on the world wide web. pages 558–567, 1997.

[3] J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, January 1996.

[4] S. B.-G. D. Radev, Z. Zhang, and R. S. Raghavan. Interactive, domain-independent identification and summarization of topically related news articles. In *5th European Conference on Research and Advanced Technology for Digital Libraries*, pages 225–238, Darmstadt, Germany, 2001.

[2]http://www.nist.gov/speech/tests/tdt/

[5] J. Dean and M. Henzinger. Finding related pages in the world wide web. In A. Mendelzon, editor, *Proceedings of the 8th International World Wide Web Conference (WWW-8)*, pages 389–401, Toronto, Canada, 1999.

[6] M. Dixon. An overview of document mining technology. *http://citeseer.ist.psu.edu/dixon97overview.html*, 1997.

[7] L. Ertoz, M. Steinbach, and V. Kumar. Finding topics in collections of documents: A shared nearest neighbor approach. In *Text Mine '01, Workshop on Text Mining, First SIAM International Conference on Data Mining*, Chicago, IL, 2001.

[8] P. Gawrysiak. Using data mining methodology for text retrieval. In *Proceedings of International Information Science and Education Conference*, Gdansk, Poland, 1999.

[9] J. Goldstein, V. Mittal, J. Carbonell, and K. M. Multi-document summarization by sentence extraction. In *Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization*, pages 40–48, Seattle, WA, May 2000.

[10] M. Gu, H. Zha, C. Ding, X. He, and H. Simon. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering. Technical Report CSE-01-007, Department of Computer Science and Engineering, the Pennsylvania State University, 2001.

[11] T. H. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In D. Lassner, D. D. Roure, and A. Iyengar, editors, *Proceedings of 11th International World Wide Web Conference (WWW-11)*, Hawai, USA, May 2002. ACM Press.

[12] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, Seattle, Washington, USA, August 2004.

[13] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2(1):1–15, July 2000.

[14] S. Lawrence and C. Giles. Accessibility of information on the web. *Nature*, 400:107–109, 1999.

[15] I. Mani and E. Bloedorn. Multi-document summarization by graph search and matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pages 622–628, Providence, RI, 1997.

[16] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2):35–67, 1999.

[17] I. Mani and M. Maybury. *Advances in automatic text summarization*. MIT Press, 1999.

[18] D. Marcu. Automatic abstracting. *Encyclopedia of Library and Information Science*, pages 245–256, 2003.

[19] J. L. Neto, A. D. Santos, C. A. A. Kaestner, and A. A. Freitas. Document clustering and text summarization. In *4th International Conference on Practical Applications of Knowledge Discovery and Data Ming*, London, 2000.

[20] C. D. Paice. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, 26(1):171–186, 1990.

[21] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[22] D. Radev. A common theory of information fusion from multiple sources. step one: cross-document structure. In *Proceedings of the 1st ACL SIGDIAL workshop on Discourse and Dialogue*, pages 74–83, Hong Kong, August 2000.

[23] D. R. Radev and W. Fan. Automatic summarization of search engine hit lists. In *Proceedings of ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, Hong Kong, October 2000.

[24] D. R. Radev and K. R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, 1999.

[25] H. D. Simon and H. Zha. Low rank matrix approximation using the lanczos bidiagonalization process. *SIAM Journal of Scientific Computing*, 21:2257–2274, 2000.

[26] C. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 1487–1494, 2000.