Evolutionary Document Summarization for Disaster Management*

Dingding Wang Li Zheng Tao Li Yi Deng
School of Computer Science
Florida International University
Miami, FL 33199
{dwang003,lzhen001,taoli,deng}@cs.fiu.edu

ABSTRACT

In this poster, we develop an evolutionary document summarization system for discovering the changes and differences in each phase of a disaster evolution. Given a collection of document streams describing an event, our system generates a short summary delivering the main development theme of the event by extracting the most representative and discriminative sentences at each phase. Experimental results on the collection of press releases for Hurricane Wilma in 2005 demonstrate the efficacy of our proposal.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Clustering

General Terms: Algorithms, Experimentation, Performance

Keywords: Evolutionary Summarization

1. INTRODUCTION

Natural calamities such as hurricanes are chaotic events that pose immense threat to businesses, human lives and properties and can inflict huge economic damages. For those disastrous events, numerous news and reports are generated as the events evolve through time. One problem that arises from this fact is that of the information overflow. Imagine, for example, that one wants to keep track of the development/story-line of a hurricane. The vast amount of news/reports makes it difficult to follow the rate at which they are being produced. In addition, it is also very challenging to extract the changes of the event at different phases for disaster management.

In this paper, we develop a novel evolutionary summarization system to summarize the changes/differences of different phases as the event evolves along time. Specifically, given a collection of document streams describing an event, we generate a short summary delivering the main development theme of the event by extracting the most discriminative sentences at each phase. This problem is related to the traditional document summarization problem since both of them extract sentences from documents to form a summary. However, traditional document summarization aims to cover the majority of information among document collections, while our goal is to identify the changes and differences over time. Our work is also different from the temporal summarization described in [1]. Allan et al. [1] addressed temporal summarization of news stories by extracting a single-sentence from each event within a news topic. While their summary is temporal and the summary sentences contain useful and novel information, their work did not explicitly model the changes/difference in different phases and thus did not deal with evolutionary summarization.

2. PROBLEM DESCRIPTION

Given a sequence of documents $D = \{d_1, d_2, \ldots, d_n\}(|D| = n)$ associated with a class indicator c which represent the time period the documents belong to. The set of sentences contained in the documents is $S = \{s_1, s_2, \ldots, s_m\}$ $(|S| = m)$. The problem of evolutionary summarization is to find a subset of sentences, $V \subset S$, to accurately discriminate the documents over different time periods, i.e. to predict the document class, given that the cardinality of V is k $(k < m)$.

3. SYSTEM OVERVIEW

Figure 1 demonstrates the framework of our evolutionary summarization system. The input of the system is a sequence of documents with labels which indicate the time period each document belongs to. First of all, the documents are preprocessed by removing formatting characters and stop words. Then these documents are trunked into sentences, and both document-term matrix and sentence-term matrix are created by calculating the term frequency of each term for each document. The document-sentence similarity matrix is constructed using cosine similarity, and both the document-sentence matrix and the document label vector are input into our sentence selection engine. The details of the sentence selection method are proposed in Section 4. Finally, an evolutionary summary is generated using the selected sentences which represent the evolution of the documents.

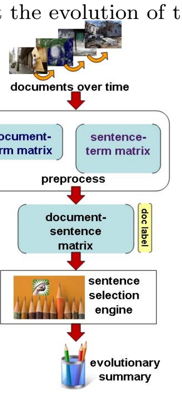


Figure 1: Overview of our system

4. SENTENCE SELECTION ENGINE

A straight-forward solution for the sentence selection problem described in Section 2 is to select a set of sentences with the highest relevance to the target class, which is called the max-relevance method [4]. Given s_i which represents the i-th sentence, and the class label c, their mutual information is defined in terms of their frequencies of appearances $p(s_i)$, $p(c)$, and $p(s_i, c)$ as

$$I(s_i, c) = \iint p(s_i, c) \ln \frac{p(s_i, c)}{p(s_i)p(c)} ds_i dc. \tag{1}$$

The max-relevance method selects the top k sentences in the descent order of $I(s_i, c)$, i.e. the best k individual sentence features correlated to the class labels.

*The work is partially supported by NSF grants IIS-0546280 and HRD-0833093.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Although we can choose the top individual sentences using the max-relevance algorithm, it has been recognized that "the k best sentences are not the best k sentences" since the correlations among those top sentence features may also be high [2]. In order to remove the redundancy among sentences, a min-redundancy and max-relevance (mRMR) framework is proposed in [4]. In mRMR, the mutual information between each pair of sentences is taken into consideration. Suppose set S represents the set of sentences and we already have V_{k-1}, the feature set with k-1 sentences, then the task is to select the k-th sentence from the set $\{S - V_{k-1}\}$. In the following formula, we see that minimizing the redundancy and maximizing the relevance can be achieved concordantly [4].

$$\max_{s_j \in S - V_{k-1}} [I(s_j; c) - \frac{1}{k-1} \sum_{s_i \in G_{k-1}} I(s_j; s_i)] \tag{2}$$

The computational complexity of this incremental sentence selection method is $O(|V| \times |S|) = O(km)$.

5. EXPERIMENTS

5.1 Real World Data

The dataset we use is the collection of press releases obtained from Miami-Dade County Department of Emergency Management and Homeland Security during Hurricane Wilma from Oct. 19, 2005 to Nov. 4, 2005. The collection contains approximately 1,700 documents, which have been categorized into 4 phases based on the status of the hurricane: (1) preparation before hurricane Wilma, (2) damage during the growth of Wilma, (3) reduce of the hurricane, and (4) the recovery after the hurricane.

5.2 Implemented Systems

In order to evaluate the effectiveness of our proposed evolutionary summarization system (called MRMREvoSum), in the experiments we implement the following baselines and compare them with our system: (1) ***Centroid based summarization system (Centroid)***: ranks sentences by computing their centroid values computed as the average cosine similarity between the sentences and the rest of the sentences [5]; (2) ***Graph based summarization system (Graph)***: selects sentences by voting from their neighbors in the sentence graph [3]; (3) ***Max-relevance based evolutionary summarization system (MREvoSum)***: selects the most important sentences using the max-relevance algorithm [4].

Phase1: preparation before Hurricane Wilma
Time Period: from 4:30pm, Oct.19, 2005 to 2:45pm Oct. 23, 2005
- A **Hurricane Watch** means that hurricane conditions are possible in the watch area within the next 36 hours. - Residents advised to be in the **ready phase for hurricane Wilma**.
Phase2: damage of Hurricane Wilma
Time Period: from 3:45pm Oct.23, 2005 to 8:15pm Oct.25, 2005
- There is **heavy tree damage** throughout the County, particularly in the north and central areas. - **MetroRail will not operate** today due to **storm damage**.
Phase3: reduce of Hurricane Wilma
Time Period: from 11:00pm Oct.25, 2005 to 5:00pm Oct.29, 2005
- **County services are resuming** throughout the county. - Residential garbage pick-up has **resumed**, roadway conditions **permitting**.
Phase4: recovery after Hurricane Wilma
Time Period: from 5:00pm Oct.29, 2005 to 11:30am Nov.4, 2005
- Homeowners who experienced roof damage due to Hurricane Wilma may be eligible for a **FEMA program** that provides temporary roof covering until permanent repairs can be made. - The **Help Us Help You program** is designed to connect those in need with these services during this critical **post-storm recovery** period.

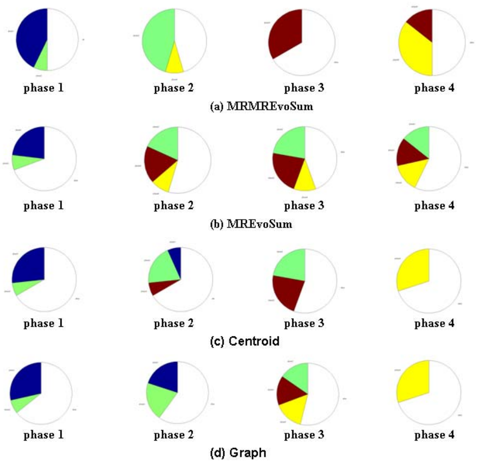
Table 1: Evolutionary Summary of Hurricane Wilma releases

5.3 An Illustrative Case Study

Table 1 demonstrates a case study of the two-sentence evolutionary summary generated by our proposed system. The words and phrases representing different phases of the hurricane growth are highlighted in Table 1. From the results, we observe that (1) the selected sentences concretely reflect the status of the hurricane; (2) the summary summarizes the advisory and notification during different time periods; (3) the two sentences extracted for each phase are informative and minimally redundant.

5.4 Term and Phrase Distribution

We ask domain experts to categorize the key words and phrases in the documents into five classes which represent each of the four phases of the hurricane growth and also a class of general terms. In this set of experiments, we calculate the term and phrase distribution of the selected sentences for each phase. Figure 2 illustrates the results visually. From the results, we can see that (1) the term distribution changes significantly over different phases of the hurricane growth, which indicates that our generated evolutionary summary can capture the hurricane evolution; (2) traditional summarization methods such as Centroid and Graph can not clearly capture the term change with different time periods. This is because the traditional summarization methods usually extract sentences delivering general information among the documents; (3) our MRMREvoSum method also outperforms MREvoSum method due to the redundancy minimum criteria used in our method.



(a) MRMREvoSum

(b) MREvoSum

(c) Centroid

(d) Graph

Figure 2: Term and phrase distribution in the summary of each phase using different summarization methods. Remark: each color represents a term category; blue:phase1; green:phase2; red:phase3; yellow:phase4; white:other.

6. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In *Proceedings of SIGIR*, 2001.

[2] T. Cover. The best two independent measurements are not the two best. *IEEE Trans. Systems, Man, and Cybernetics*, 1974.

[3] G. Erkan and D. Radev. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*, 2004.

[4] F. L. Hanchuan Peng and C. Ding. feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[5] D. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, pages 919–938, 2004.