N-33 Topic Extraction and Summarization in News Archive using TF*PDF Algorithm and Sentence Vector Clustering

Khoo Khyou Bun† Hiroshi Dohi† Mitsuru Ishizuka†

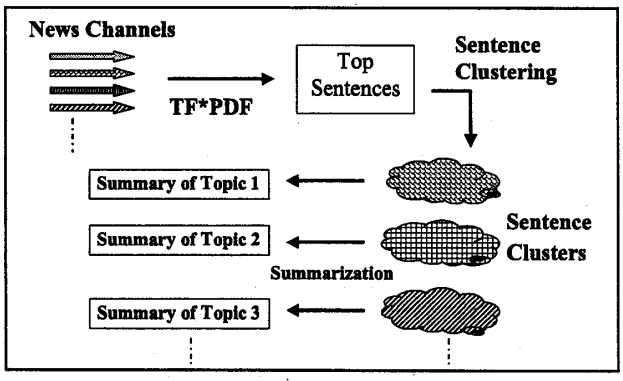
1. Abstract

Busy and no time to digest the news archive .... ? Ever since the Web wide-spreading, the amount of electronically available information online, especially news archive proliferates and threatens to overwhelm human attention. Seeing this, we propose an information system that will extract the main topics in the news archive in a weekly basis. By getting a weekly report, user can know what were the main news events in the past week.

2. Introductions and System Architecture

The goal of this research is to summarize the main topics in weekly news archive automatically for users. Figure 1 illustrates the information flow of the system.

Figure 1: System Information Flow



Every week, this system collects the news articles from multiple news channels. The terms in the archive that try to describe the main topics will be heavily weighted using TF*PDF algorithm [1] (Eqn. 1). This algorithm makes use of the concept that when there is a hot (main) topic on the air, the terms that describe the topic will appear in many documents in many news channels. Then, starts from the sentences with highest average weight, we do sentence clustering. Each resulted sentence cluster represents a main topic. The sentences in each cluster will be arranged chronologically to form a summary of the main topic respectively.

3. TF*PDF Algorithm

$$W_j = \sum_{c=1}^{c=D} \left|\overrightarrow{F_{jc}}\right| \exp\left(\frac{n_{jc}}{N_c}\right) \qquad \rightarrow (Eq.1)$$

$$\left|\overrightarrow{F_{jc}}\right| = \frac{F_{jc}}{\sqrt{\sum_k (F_{kc})^2}} \qquad \rightarrow (Eq.2)$$

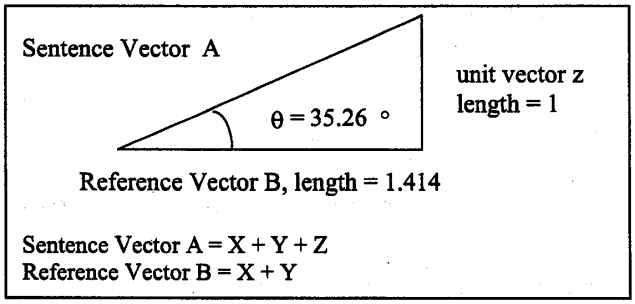
Wj = Weight of term j; Fjc = Frequency of term j in channel c ; njc = Number of document in channel c where term j occurs ; Nc = Total number of document in channel c; k=total number of terms in a channel ; D = number of channels

TF*PDF is innovated to give the significant weight to the terms that explain the main topic. Different from the conventional term weight counting algorithm TF*IDF [2], in TF*PDF algorithm, the weight of a term from a channel is linearly proportional to the term's within-channel frequency, and exponentially proportional to the ratio of document containing the term in the channel. The total weight of a term will be the summation of the term's weight from each channel. With this, the terms devoted to the main (hot) topics will gain significant weight.

4. Sentence Vector

The top 30 TF*PDF terms will be used as unit vectors. Each sentence may contain a different distribution of unit vectors. When a sentence vector has an acute angle not bigger than 35.26 degree with reference vector of another sentence, the two sentences will be classified in the same cluster. This is illustrated in Figure 2.

Figure 2: Minimum acute angle for sentence clustering



There are 4 categories of sentences during the clustering process:

1. CS (Cluster Sentence): sentence clustered to a certain topic correctly
2. MS (Miss sentence): sentence clustered to a certain topic wrongly
3. FS (Fail sentence): sentence belong to an existing topic cluster but failed to be clustered into
4. NC (Not clustered sentence): sentence not belonging to any existing topic cluster

More the sentences we evaluate in the clustering process, more clusters we will produce, with each concerning a different topic. However, the clusters may be consisting a large different number of sentences. Thus, some summarization jobs, should it be compression, compaction or condensation, are needed to be done. However, this is not elaborated in this paper.

† Dept. of Info. & Comm. Eng.,The University of Tokyo

5. Samples Run

An experiment has been done on the online news archive from 4 newswire sources: Associated Press (AP), The New York Times (NYT), Reuters and USA Today. The following subsection elaborates on the results from the experiment done on the news archive dated from May 13, 2002 to May 19, 2002.

5.1 Experiment on archive from May 13 to May 19

Table 1: Top TF*PDF Terms (May 13 to May 19)

Term	Weight	Term	Weight	Term	Weight
official	718.28	Security	212.04	Democrat	165.55
Bush	602.48	Carter	210.38	priest	164.66
Palestinian	588.78	Cuba	193.24	Qaeda	162.7
attack	452.58	intelligence	189.57	home	160.9
American	396.67	Republican	184.84	bomb	160.35
House	346.34	Terrorist	176.96	threat	154.15
Arafat	279.4	Israel	175.04	Cuban	152.1
Israeli	266.24	Washington	174.49	Pakistan	145.35
kill	260.66	Russia	169.3	warn	142.22
White	256.43	Laden	166.47	sign	138.93

Table 1 shows the top 30 most heavily weighted TF*PDF terms. These are the terms used as sentence unit vectors. 25 top sentences with highest average weight were elaborated in the sentence clustering process. As a result, 20 sentences were clustered successfully into 2 clusters. The first cluster consists of sentences 2, 3, 7, 12, 15, 18, 21, 23, 24 and 25 in the original sentences ranking. These sentences were then re-arranged chronologically to form a summary of the first topic as shown in Table 2. This cluster concerns the scrutiny of Bush administration handling of intelligence data on suspected terrorists in the United States months before the Sept. 11 hijack attacks on World Trade Center, and the possibilities of the next wave of terrorists attacks on American.

Table 2: Summary from first topic cluster

No.*	RN.**	Summary (sentences)
1	15	On Wednesday night, the **White House** said President **Bush** was **warned** by **American intelligence** agencies in early August of Mr. bin **Laden**'s desires to hijack airplanes.
2	24	Key members of Congress have asked whether the government had information pointing to the **attacks** on America after the **White House** disclosed President **Bush** had had an **intelligence** briefing in early August that included concerns bin **Laden**'s group might try to hijack a passenger plane.
3	2	**White House officials** confirmed that **Bush** was told in a briefing a month before the **attacks** that bin **Laden**'s al-**Qaeda** network had discussed hijacking **American** planes.
4	7	In the interview, the **official** described a plan to act against Mr. bin **Laden** that was developed in August, approved at the level of the "deputies" the No. 2 **officials** in several departments and then approved by the top cabinet **officials** on Sept. 4.
5	12	That plan, which was drawn up by high-ranking **officials** among several Cabinet departments, was awaiting President **Bush**'s review when the World Trade Center and Pentagon were **attacked**.
6	21	The **White House** also acknowledged on Friday that **security officials** had prepared a presidential order for a campaign to dismantle al **Qaeda**.
7	23	May 17 The **White House** began an aggressive **attack** on **Democrats** in Congress today as President **Bush** tried to contain the political fury over a **warning** he received last August that Osama bin **Laden** might be planning a hijacking.
8	25	In response to the uproar after the disclosure of the August **warning** to Mr. **Bush**, **White House officials** insisted that they had no serious evidence last summer that Al **Qaeda** was considering a suicide hijacking.
9	18	United States **intelligence officials** said that they began to intercept communications among **Qaeda** operatives discussing a second major **attack** in October, and that they have detected recurring talk among them about another **attack** ever since.
10	3	A **White House official** said on Saturday U.S. **intelligence officials** have detected "enhanced activity" that points to a potential new **attack** against the United States or **American** interests abroad.

** Original sentences ranking number
* Chronological order of sentences in cluster/summary

5.2 Discussion

In Table 2 of first topic summary, the highlighted terms are the top 30 TF*PDF terms used as unit vector. These terms explained the same topic and hence enable the sentence clustering for the first topic. Also, there is a second topic being clustered successfully but not listed here. In total, 20 sentences or 80% of the 25 evaluated sentences were clustered successfully.

6. Conclusions

In this paper, a web based information system useful in generating summary of main topics from weekly news archive has been proposed. Results from the experiments have shown to us the viability of our system. Having this system to report us the summary of main topics automatically, we can keep track of the main happening events without much effort.

Reference:

[1] K.B. Khoo and M. Ishizuka, Emerging Topic Tracking System, Proc. of Web Intelligence (WI01), LNAI 2198 (Springer), pp. 125-130, Maebashi, Japan. 2001

[2] G. Salton and C. Buckley : Term-Weighting Approached in Automatic Text Retrieval, Information Processing and Management, Vol. 4, No. 5, pp. 513-523 (1989)