Topic Extraction from News Archive Using TF*PDF Algorithm

Khoo Khyou Bun Mitsuru Ishizuka
Dept. of Information and Communication Engineering
The University of Tokyo
{kbkhoo,ishizuka}@miv.t.u-tokyo.ac.jp

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.

1. Introduction

The Web keeps growing and huge amount of new information are being posted on it continuously. Weekly, tens or hundreds of Megabytes of news stories can be added easily to the news archive of any newswire sources online. At the same time containing some influencing knowledge, this news archive may also be holding many uninteresting or trivial news. The influencing knowledge is desired but reading the news archive is rather a daunting task that will take us a lot of time and effort. And yet, this doesn't promise us that all the main topics will be discovered. So, it would be helpful if there is kind of system which would respond correctly to the generic queries such as "What's new?" or "What's important?" Unfortunately, traditional goal-driven retrieval system works well only for content-based queries. It is very useful or efficient when the user knows precisely the goal or facts he/she is seeking. However, we are at a higher level of abstraction and creating the precise goals with zero knowledge of past week's news is rather unrealistic. Hence, what would be desirable is an intelligent system that automatically summarizes us a weekly report of the main topics embedded in the archive of newswire sources on the Web.

Research on TF*PDF (Term Frequency * Proportional Document Frequency) (Eq. 1) [1, 2] algorithm has been reported else where. TF*PDF algorithm is adapted in the ETTS [1] which is useful in tracking the emerging topic in a particular information area of interest on the Web, by summarizing the change posted on it. This information area consists of a number of web domains having so called "stock" type information which is rather static. Whereas, this paper elaborates the application of TF*PDF algorithm on a number of newswire sources on the Web, which is said to be having varieties of "flow" type information.

When asked "what were the hot topics in the past week?" We will answer, "The prominent topics would be the ones that were discussed frequently in many documents from many newswire sources last week." This makes sense to the questioner. Thus, in building a system that is useful in summarizing a weekly news report of main topics, we use the novel TF*PDF algorithm to recognize the terms that try to explain the main topics. These would be the terms that appear frequently in many documents from many newswire sources concurrently. TF*PDF algorithm is designed in a way that it would assign heavy term weight to these kind of terms and thus reveal the main topics.

In section 2, we discuss some related work and fundamental ideas in event detection and tracking by conventional IR system. Section 3 delivers our approach in generating weekly news report of main topics, followed by section 4 which discusses the results of the experiments run. We conclude and brief on the research direction in section 5.

2. Related Works and Fundamental Study

Among the conventional information retrieval systems [15, 11, 5, 3, 4] working on news archive, the well-known TDT [5] is picked for detail study and discussion below. We will point out some characteristics of conventional IR system that make it unsuitable for us to fulfill the objective of summarizing the weekly report of main topics from news archive.

The goal of the pilot research in Topic Detection and Tracking (TDT) [8] was new detection and tracking within streams of broadcast news stories. The involved researchers found and quoted in [8] that "the state of art is capable of providing adequate performance for detecting and tracking of new event, but there is a high enough failure rate to warrant significant research into how algorithms can be advanced". Its successor, TDT [5] added the online news events detecting and tracking functionality as one of its main objectives. Taking advantage that the on-going event would have its keywords appeared in multiple documents in a certain time frame, the document burst detection using time window and document clustering have been the central ideas. The conventional TF*IDF [13] algorithm has been adapted by these system to judge the unique terms in a document, and thus the uniqueness of the document. I would say it judges the unique terms of a document because this algorithm tends to give the most significant weight multiplication to a term when the term appears in only one document or in its first document, by the multiplication of the IDF (inverse document frequency). Because of the large retrospective [9] corpus used (six month collection of CNN news corpus in TDT), the event keyword appears in its first document or early stage of the event document burst will be weighted significantly and thus can trigger the onset of event detection well. However, after the appearance of a number of event documents or in the midst of the event, more the consequent event documents containing the event keywords appear, less likely the document will be judged as an event document because TF*IDF algorithm always try to give significant weight to the terms that appear in less documents. Thus, we have the fear of losing the detection of an event although it was hot and taking the central stage by being discussed in many documents.

In TDT, large specified retrospective [9] corpus has been used to calculate the incremental IDF. As stated in [5], we understand that IDF works effectively for document retrieval after a sufficient number of documents have already been processed. However, as the retrospective corpus plays an important role in the calculation the term weight, it may influences the results negatively if not properly carved. In short, it is desirable if the retrospective corpus can be excluded in terms weight calculation.

Also, TDT don't take advantage on the good assumption that during the onset of an event, the event would be reported heavily from the major channels concurrently. By the time, many important terms appear in multiple documents from many different channels will be discussing the hot event. By making use of this good characteristic, we can carve an algorithm to recognize these terms and thus the hot topic accurately, even without the help of retrospective corpus. However, TDT groups up the information from different channels (Reuters and CNN news articles in TDT1) into a large corpus for TF*IDF term weight calculation.

Qualitative and quantitative test has been done on TDT with the retrospective corpus TDT1, 15863 chronologically ordered and manually segmented news stories spanning from 1 July 1994 to 30 June 1995. The histograms of events in [5] showed that TDT is suitable in event detection in retrospective corpus. However, it is not necessarily that it will have the same performance or suitable for online detection identifying the onset of new events from live news feeds in real time, due to the difference in the calculation of IDF. In real time online detection, a dynamic IDF is used; while in events detection in retrospective corpus, a fixed value of IDF calculated by the whole corpus is used. Even though the difference in characteristic and orientation, events burst in the histograms may not fail to trigger the onset of a new event, depends on the wide of the time window; however, it would take the long of a time window to detect a new event. This means that it would take at least four weeks for four weeks time window to trigger the onset of a new event.

In short, the technologies adapted in conventional IR or new event detection systems are insufficient to innovate a system to summarize us weekly news report of main topics. Thus, we were motivated to come out with this paper to introduce our approach.

3. System Architecture and TF*PDF Algorithm

3.1. Approach

Although the research goal is quite similar with some related works such as TDT, we are addressing the problem in a totally different approach. TF*PDF algorithm is used to recognize the terms that explain the main topics in each week news archive. Further, the sentences with high average weight will be classified or clustered according to their topic by using their sentence vectors. Each sentence is associated with a sentence vector consisting of unit vectors which are terms with highest term weight. Using these sentence vectors, we could classify the sentences according to respective topic by examining their component unit vectors. Each sentences cluster will be representing a topic. The sentences in each topic cluster are then arranged chronologically to form a summary of the topic. The flow of information in the system is illustrated in Figure 1.

3.2. Novel Concept of TF*PDF

Our system works on the basic concept that whenever there is a hot topic in the air, the topic will be discussed frequently in many news documents from majority newswire sources. Thus, instead of collecting the information from all sources into a "large" mixed corpus and calculate its term weight with a standard TF*IDF algorithm like what TDT does, we rather give equal importance to the information from each newswire sources and channel them to our system in parallel. The terms that explain the hot topics

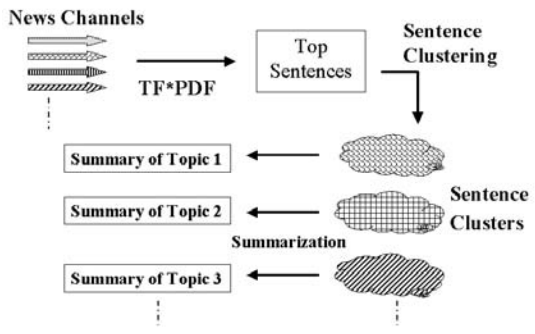


Figure 1. System Information Flow

should appear frequently in many documents in each channel will be weighted significantly. Whenever the majority channels contain a common term with heavy term weight concurrently, that should be the terms that explaining the main topics discussed broadly.

Also noted that, terms are basically content words. Stop words like prepositions (i.e. in, from, to, out) and conjunctions (i.e. and, but, or) are eliminated via a stop word list.

3.3. TF*PDF Algorithm

Thus, in order to fulfill our objective to recognize the terms that explain the hot topics, TF*PDF is innovated to count the significance (weights) of the terms. Different from the conventional term weight counting algorithm TF*IDF [13], 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 term's weight from each channel as follows.

$$W_j = \sum_{c=1}^{c=D} |F_{jc}| \exp(\frac{n_{jc}}{N_c}), \qquad (1)$$

$$|F_{jc}| = \frac{F_{jc}}{\sqrt{\sum_{k=1}^{k=K} {F_{kc}}^2}}, \qquad (2)$$

where, W_j=Weight of term j; F_{jc}=Frequency of term j in channel c; n_{jc}=Number of document in channel c where term j occurs; N_c=Total number of document in channel c; k=Total number of terms in a channel; D=number of channels

There are three major compositions in TF*PDF algorithm. The first composition that contributes to the total weight of a term significantly is the "summation" of the term weight gained from each channel, provided that the term deems to explain the hot topic discussed generally in majority of the channels. In other words, the terms that deem to explain the main topic will be heavily weighted. Also, larger the number of channels, more accurate will be this algorithm in recognizing the terms that explain the emerging topic.

The second and third compositions are combined to give the weight of a term in a channel. The second composition is the normalized term frequency of a term in a channel $|F_{jc}|$ as showed in Equation 2. The term frequency needs to be normalized because when different channel has a different size of archive, the term from a channel with more documents has a proportionally higher probability that it will occur more frequently. We want, however, to give equal importance or equal weighting to the same term from each channel; thus normalization would be carried out.

The third composition is the PDF (proportional document frequency) of a term in a channel $exp(n_{jc}/N_c)$. It is the exponential of the number of documents containing the term to the total number of documents in the channel. Here, terms that occur in many documents are more valuable (or weighted) than ones that occur in a few. Hence, the term that occurs more frequent in many documents in a channel would be the term that deems to explain the main topic in the channel. In real world, PDF has been experimentally proved to work well in such a way that it should grow exponentially in respect to the number of documents containing a term, instead of linearly, so that we can give a more significant (different) weight to the term that occurs in many documents compare to the one occurs in just a few. Mathematically, larger the number of documents containing a term in a channel, higher will be the grow rate of the PDF of the term in the channel. Numerically, this PDF has a value ranges from 1 (e^0) to 2.718 (e^1) exponentially (base e).

The total weight of a term (W_j) is equal to the summation of the weight of the term in each channel respectively. Readers may ask why W_j is calculated in this way instead of treating all the documents from all the channels the same and counts the W_j by multiplying the overall term frequency and the overall proportional document frequency. The paragraph above (describing the first composition in TF*PDF algorithm which contributes significantly to the weight of a term) has been explaining part of the answer to this question but we would like to emphasize again that, the terms that discuss the main topic in a weekly news archive would be the terms that explain the hot topic being discussed in majority of the channels. In other words, they are the terms that occur frequently in majority of the channels. Thus, W_j needs to be counted as described above. Otherwise, if there is any channel with a large number of documents containing certain terms of significant weight, the results would be deviated from having terms that explain the hot topics in majority channels.

In short, TF*PDF algorithm give significant weights to the terms that explain the common hot topic in majority channels.

3.4. Sentence Vector

As depicted in Figure 1., after TF*PDF term weight counting, sentence clustering will be carried out as the next step towards the summarization of the main topics. Each resulted cluster of sentences will be discussing a specific topic. Each sentence vector in the clusters might consist of a different combinatin and number of unit vectors. The top 30 TF*PDF terms with highest weight would be the unit vectors.

As the basic rule, when a sentence vector of a sentence has an acute angle not larger than 35.26 degree with any combination of unit vectors of another sentence vector (we refer this as reference vector), the sentence will be classified in the same cluster with the reference sentence. For example, the sentence vector No. 5 (having unit vectors “Palestinian”, “official” and “Arafat”) in Table 2 has an acute angle of 35.26 degree with the reference vector of terms “Palestinian” and “official” from sentence No. 1 (See Figure 2 below). So, sentence No. 5 would beclustered together with sentence No. 1. In the same way, the sentence No. 9 can be clustered with either sentence No. 1 or sentence No. 11. And instead, all the sentences No. 1, 5 and 9 can be clustered with sentence No. 11. In this way, sentence clustering will be carried out.

The acute angle of 35.26 degree is the optimal derived after emprical testing. This is the angle to possible the clustering of a sentence with 3 unit vectors into another sentence containing at least 2 same unit vectors. There have been noises to cluster a sentence to the sentences containing only half the number (50%) of its unit vectors. Consequently, sentences with 2 unit vectors can only be clustered with sentences containing both its unit vectors, while sentence with 4 unit vectors can only be clustered with sentences having at least 3 same unit vectors and so on.

There are four categories of vector sentences after or 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 but wrongly. 3. **FS** (fail sentence): sentence belong to an existing topic cluster but fail to be clustered in. 4. **NC** (not clustered sentence): sentence not belonging to any existing topic cluster.

We might be able to generate many clusters of sentences with each concerning a different topic. However, there are possilities that the number of sentences in different cluster may vary largely. This is rather normal than a odd especially we are here to start the clustering process from the top sentence with highest average weight, to a certain extent in descending order.

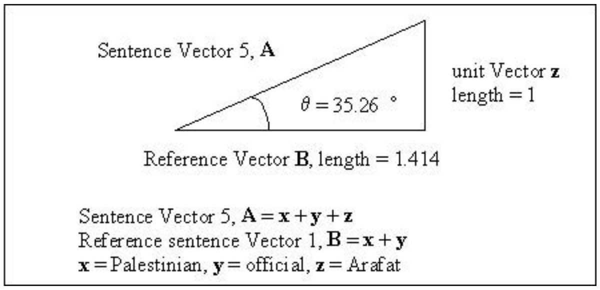

Sentence Vector 5, **A**
$\theta = 35.26$ °
unit Vector **z**
length = 1
Reference Vector **B**, length = 1.414
Sentence Vector 5, $\mathbf{A} = \mathbf{x} + \mathbf{y} + \mathbf{z}$
Reference sentence Vector 1, $\mathbf{B} = \mathbf{x} + \mathbf{y}$
$\mathbf{x}$ = Palestinian, $\mathbf{y}$ = official, $\mathbf{z}$ = Arafat

Figure 2. Maximum Acute Angle

After the process of sentences clustering, all the sentences in each cluster will be arranged chronologically for a topic summary. Lastly, due to the possible large different size of cluster, we may need to reduce the length of summary text by using some dedicated summarization techniques, should it be compression, compaction or condensation [6]. However, in this paper, the last step is not elaborated. Thus, it is not strange if the output summarization text presented in Section 4 contains some dangling anaphors.

4. Samples Run

4.1. Corpus

Two experiments have been run on the online news archive taken in from 4 newswire sources concurrently: Associated Press (AP), The New York Times (NYT), Reuters and USATODAY. Section 4.2 describes the first experiment done on the news archive dated from May 13, 2002 to May 19, 2002. Then, Section 4.3 explains on the experiment done on the news archive dated from May 6 2002 to May 12, 2002. We have been trying to summarize weekly report in two consequence weeks. The collected weekly news archives consist about 400-600 important news documents.

4.2. Experiment on archive dated from May 13 to May 19

Table 1 shows the top 30 most heavily weighted TF*PDF terms. Table 2 shows the top 25 sentences with highest average weight. Only these 25 sentences are used in the sentences clustering for elaboration. The top terms (unit vectors) contained are highlighted in bold. For each respective sentence, it’s sentence unit vectors, status, date and source are stated in Table 3.

After the sentences clustering process, 20 sentences were clustered successfully into two clusters. The first cluster consists of the sentences 2, 3, 7, 12, 15, 18, 21, 23, 24

Table 1. Top TF*PDF Term (May 13 to May 19)

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

and 25 in the original ranking. These sentences were then arranged chronologically to form a summary as shown in Section 4.2.1. This cluster and thus the summary 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. Another cluster consists of sentences 1, 4, 5, 9, 10, 11, 13, 17, 19 and 22 is related to the continuing suicide bombings in Israel and the calling of reforms in Arafat's administration. The summary of this cluster is displayed in Section 4.2.2.

4.2.1 Summary 1

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. 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. 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. 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. 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. The White House also acknowledged on Friday that security officials had prepared a presidential order for a campaign to dismantle al Qaeda. 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. 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. 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.

4.2.2 Summary 2

In Washington, a senior Bush administration official declined comment on the Likud resolution, but said President Bush remains committed to the establishment of a Palestinian state. Arafat also condemned Israel's six-week incursion into West Bank cities to root out Palestinian militants that also destroyed much of the infrastructure of the Palestinian security forces. Wolfowitz repeated the Bush administration's view that an end to Israeli military occupation of Palestinian territory and a Palestinian state was key to solving the Arab-Israeli problem. When Israeli troops moved into Palestinian towns last month, confining Arafat to his Ramallah office, many Palestinians gave him their backing, viewing the Israeli action as part of a larger attack on Palestinian aspirations for statehood. The New York Times on Thursday quoted a senior Israeli official telling reporters in Washington on condition of anonymity that reforming the Palestinian security forces could not be accomplished while Arafat was in charge. "It is the time for change and reform," Arafat told the Palestinian Legislative Council in a speech on the day Palestinians mark as the Nakba of the founding of Israel in 1948, which displaced hundreds of thousands of Palestinians. Arafat responded to widespread pressure from ordinary Palestinians, Israel and foreign leaders by calling a speech to the Palestinian parliament on Wednesday for elections and reforms. Israeli forces have encircled Palestinian cities in the West Bank and set up checkpoints across Palestinian territories which they say are meant to prevent attacks on Israelis. Meanwhile, 15 Palestinian Cabinet ministers offered to resign Saturday, officials said, a gesture to spur reforms in the Palestinian Authority , headed by Arafat. Palestinian officials have been pondering elections and reform in the face of internal, international and Israeli demands for a restructuring of the Palestinian Authority and its security forces.

Table 2. 25 Highest Weighted Sentences (May 13 to May 19)

No.	Top Sentences	Weight
1	In **Washington**, a senior **Bush** administration **official** declined comment on the Likud resolution, but said President **Bush** remains committed to the establishment of a **Palestinian** state.	263.02
2	**White House official**s confirmed that **Bush** was told in a briefing a month before the **attack**s that Bin **Laden**'s al-**Qaeda** network had discussed hijacking **American** planes.	248.95
3	A **White House official** said on Saturday U.S. **intelligence official**s have detected "enhanced activity" that points to a potential new **attack** against the United States or **American** interests abroad.	233.07
4	**Palestinian official**s have been pondering elections and reform in the face of internal, international and **Israeli** demands for a restructuring of the **Palestinian** Authority and its **security** forces.	216.72
5	Meanwhile, 15 **Palestinian** Cabinet ministers offered to resign Saturday, **official**s said, a gesture to spur reforms in the **Palestinian** Authority , headed by **Arafat**.	206.36
6	Days after the summit concluded, **Israel** began **attack**ing **Palestinian** territory in revenge for **Palestinian** suicide **bomb**ings on **Israeli** targets.	202.87
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 **official**s in several departments and then approved by the top cabinet **official**s on Sept. 4.	202.78
8	CHICAGO - President **Bush** will keep pushing for the creation of a **Palestinian** state despite a vote by **Israel**'s ruling right-wing Likud party never to accept one, the **White House** said on Monday.	198.67
9	Wolfowitz repeated the **Bush** administration's view that an end to **Israeli** military occupation of **Palestinian** territory and a **Palestinian** state was key to solving the Arab-Israeli problem.	198.06
10	When **Israeli** troops moved into **Palestinian** towns last month, confining **Arafat** to his Ramallah office, many **Palestinian**s gave him their backing, viewing the **Israeli** action as part of a larger **attack** on **Palestinian** aspirations for statehood.	194.94
11	The New York Times on Thursday quoted a senior **Israeli official** telling reporters in **Washington** on condition of anonymity that reforming the **Palestinian security** forces could not be accomplished while **Arafat** was in charge.	193.83
12	That plan, which was drawn up by high-ranking **official**s among several Cabinet departments, was awaiting President **Bush**'s review when the World Trade Center and Pentagon were **attack**ed.	192.28
13	**Israeli** forces have encircled **Palestinian** cities in the West Bank and set up checkpoints across Palestinian territories which they say are meant to prevent **attack**s on **Israeli**s.	191.40
14	**Bush**'s brother Jeb **Bush**, the Florida governor, faces re-election this year and also is depending on **Cuban American**s, who vote heavily **Republican**.	188.67
15	On Wednesday night, the **White House** said President **Bush** was **warn**ed by **American intelligence** agencies in early August of Mr. bin **Laden**'s desires to hijack airplanes.	187.19
16	**Carter** told Castro and leading **Cuban** scientists that he had asked **White House**, State Department and **intelligence official**s specifically if **Cuba** was transferring technology or other information that could be used in **terrorist** activities.	184.71
17	”It is the time for change and reform,” **Arafat** told the **Palestinian** Legislative Council in a speech on the day **Palestinian**s mark as the Nakba of the founding of **Israel** in 1948, which displaced hundreds of thousands of **Palestinian**s.	184.10
18	United States **intelligence official**s 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.	183.17
19	**Arafat** responded to widespread pressure from ordinary **Palestinian**s, **Israel** and foreign leaders by calling a speech to the **Palestinian** parliament on Wednesday for elections and reforms.	182.48
20	Indian **intelligence official**s say Dawood Ibrahim, a Bombay crime boss wanted in India, is back in **Pakistan** and plotting retaliatory **attack**s with Pakistani **intelligence official**s.	182.43
21	The **White House** also acknowledged on Friday that **security official**s had prepared a presidential order for a campaign to dismantle al **Qaeda**.	178.97
22	**Arafat** also condemned **Israel**'s six-week incursion into West Bank cities to root out **Palestinian** militants that also destroyed much of the infrastructure of the **Palestinian security** forces.	174.92
23	WASHINGTON, May 17 The **White House** began an aggressive **attack** on **Democrat**s in Congress today as President **Bush** tried to contain the political fury over a **warn**ing he received last August that Osama bin **Laden** might be planning a hijacking.	174.19
24	Key members of Congress have asked whether the government had information pointing to the **attack**s 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.	172.13
25	In response to the uproar after the disclosure of the August **warn**ing to Mr. **Bush**, **White House official**s insisted that they had no serious evidence last summer that Al **Qaeda** was considering a suicide hijacking.	171.90

Table 3. Sentence's Unit Vector, Status, Date and Source

No.	Unit Vectors	Status	Date, Time (May)	Source
1	Washington Bush official Palestinian	CS	13, 5:45 PM	AP
2	White House official Bush attack Laden Qaeda American	CS	17, 5:56 AM	USAToday
3	White House official intelligence attack American	CS	19, 3:29 PM	Reuters
4	Palestinian official Israeli security	CS	19, 3:36 PM	Reuters
5	Palestinian official Arafat	CS	18, 5:11 PM	AP
6	Israel attack Palestinian bomb Israeli	FS	18,10:54 AM	Reuters
7	official Laden	CS	17,8:56 AM	NYT
8	Bush Palestinian Israel White House	FS	13,10:46 AM	Reuter
9	Bush Israeli Palestinian	CS	15, 2:11 PM	AP
10	Israeli Palestinian Arafat attack	CS	15, 6:50 PM	AP
11	Israeli official Washington Palestinian security Arafat	CS	16, 3:30 PM	Reuter
12	official Bush attack	CS	17, 2:55 PM	NYT
13	Israeli Palestinian attack	CS	17, 6:36 PM	Reuters
14	Bush Cuban American Republican	NC	19, 7:36 PM	AP
15	White House Bush warn American intelligence Laden	CS	16, 2:55 PM	NYT
16	Carter Cuban White House intelligence official Cuba terrorist	NC	14, 1:36 PM	AP
17	Arafat Palestinian Israel	CS	16,6:08 AM	USA Today
18	Intelligence official Qaeda attack	CS	18, 2:52 PM	NYT
19	Arafat Palestinian Israel	CS	16, 6:49 PM	Reuters
20	Intelligence official Pakistan attack	MS	14, 2:55 PM	NYT
21	White House security official Qaeda	CS	17, 5:26 PM	Reuters
22	Arafat Israel Palestinian security	CS	15, 1:49 PM	AP
23	White House attack Democrat Bush warn Laden	CS	18,8:51 AM	NYT
24	attack White House Bush intelligence Laden	CS	16, 7:31 PM	Reuters
25	warn Bush White House official Qaeda	CS	18, 2:52 PM	NYT

4.2.3 Discussions

Two FS (fail sentences): 20 sentences which is 80 % of the all 25 sentences were clustered successfully. However, there are 2 FS: sentences 6 and 8. The content of these two sentences are related to the topic in second summary but they fail to be clustered because their unit vectors couldn't make up with the minimum acute rules to be classified into the cluster. But coincidentally, the sentence 8 is repeating the content of sentence 1 in the cluster.

One MS (miss sentence): The sentence 20 is a miss sentence with the combination of unit vectors that caused it to be classified wrongly into the first summary cluster (not included in the summary for easy understanding). This sentence talks about the dispute between India and Pakistan. FS may make no harms to our output summary but not the MS. Miss sentence will be included in the weekly report automatically by the system and it may mislead the reader regarding the main topic.

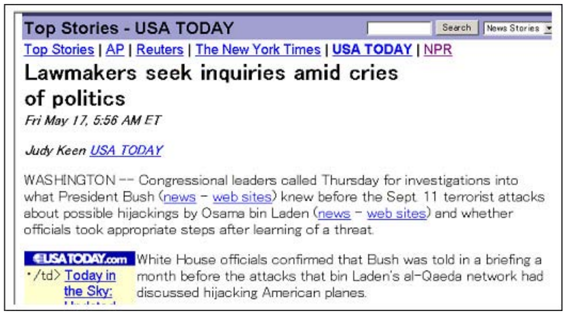
Two NC (not clustered sentences): sentence 14 and 16. These sentences are related to the raising issue of Cuban being accused as a terrorist activities propagator state and U.S. is going to tighten the economic pressure on it. A NC doesn't mean that it will be un-clustered forever. These two sentences are un-clustered most probably because there is not enough sentences have been used in sentences clustering elaboration. So, if we evaluate more sentences, the topics regarding the Cuba, India-Pakistan dispute and even the remarkable Russia-US(Nato) consensus would have been clustered out. However, when we go deeper down the list of sentences, the first two clusters may grow bigger with contents repeat. The TF*PDF terms in Table 1 gives a good hints on the discoverable topics suggested.

Note: The sentence (sentence 8 and 23) begins with "all capital letter" word is the first sentence in the document. The "all capital letter" words are excluded in weight counting.

4.2.4 Page Extraction

The top three pages with highest average weight are as Figures 3, 4 and 5 below. These three pages are ranked at the top (same sequence) when both counted by using the top 30 or all the TD*PDF terms, but with different page weight. Using only the top 30 terms weighted the top three pages at 102.18, 97.46 and 86.58; while using all the TF*PDF terms weighted the three pages at 132.34, 130.66 and 122.70.

The Figure 3 tells that the U.S. congressional leaders called for investigation into what President Bush knew before the Sept. 11 terrorist attacks about possible hijackings by Osama bin Laden. The second sentence in this document is the second sentence in Table 2 with 248.95 average sentence weights. Figure 4 tells that Arafat asked for the Israeli troops withdrawal before holding a new Palestinian election. Sentence 13 is one of the sentences in this page. The sentence No. 8 in Table 2 matches the first sentence in



Lawmakers seek inquiries amid cries of politics

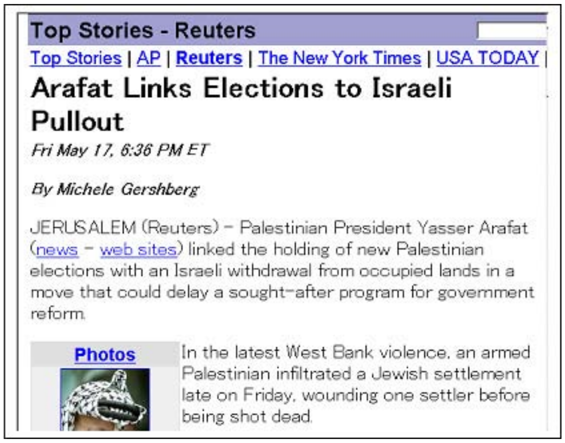
Fri May 17, 5:56 AM ET

Judy Keen USA TODAY

WASHINGTON -- Congressional leaders called Thursday for investigations into what President Bush (news - web sites) knew before the Sept. 11 terrorist attacks about possible hijackings by Osama bin Laden (news - web sites) and whether officials took appropriate steps after learning of a threat.

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.

Figure 3. USA Today, May 17


Top Stories - Reuters
Top Stories | AP | Reuters | The New York Times | USA TODAY |
Photos

Arafat Links Elections to Israeli Pullout

Fri May 17, 6:36 PM ET

By Michele Gershberg

JERUSALEM (Reuters) - Palestinian President Yasser Arafat (news - web sites) linked the holding of new Palestinian elections with an Israeli withdrawal from occupied lands in a move that could delay a sought-after program for government reform.

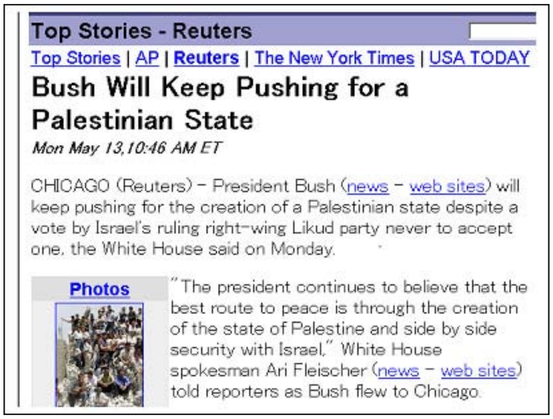
In the latest West Bank violence, an armed Palestinian infiltrated a Jewish settlement late on Friday, wounding one settler before being shot dead.

Figure 4. Reuters, May 17

the Figure 5 page.

4.3. Experiment on archive dated from May 6 to May 12

Table 4 shows the top 30 most heavily weighted TF*PDF terms. Figure 6,7 and 8 illustrate the top three pages extracted. All three pages are related to the Israel-Palestinian issues. It was a week of high tension in that region. We can see from Table 4 that the terms Palestinian and Israeli gain the highest term weight. Instead, majority of the top ten terms explain the Palestinian-Israeli issues. However, in the lower rank, we can find some important terms concerning issues such as pipe bomb student Hedler and Enron case. Overall, this is a week dominated by Palestinian-Israel issues, followed by few sub-topics like Hedler and Enron. We would be able to generate a summary for each of these topics by the sentences clustering techniques described in the previous section.



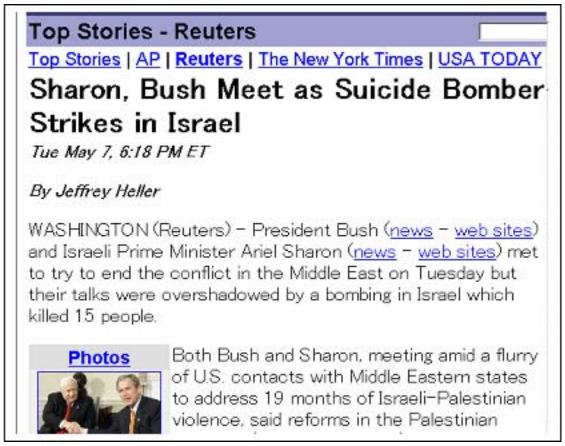
Bush Will Keep Pushing for a Palestinian State

Mon May 13,10:46 AM ET

CHICAGO (Reuters) - President Bush (news - web sites) will keep pushing for the creation of a Palestinian state despite a vote by Israel's ruling right-wing Likud party never to accept one, the White House said on Monday.

"The president continues to believe that the best route to peace is through the creation of the state of Palestine and side by side security with Israel," White House spokesman Ari Fleischer (news - web sites) told reporters as Bush flew to Chicago.

Figure 5. Reuters, May 13



Sharon, Bush Meet as Suicide Bomber Strikes in Israel

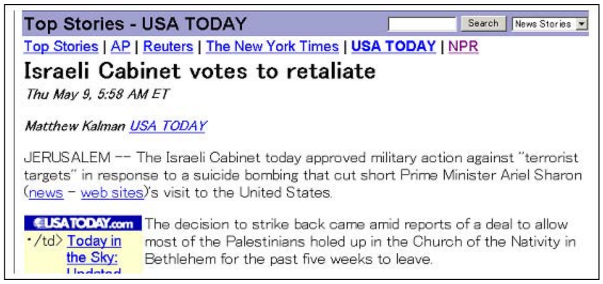
Tue May 7, 6:18 PM ET

By Jeffrey Heller

WASHINGTON (Reuters) - President Bush (news - web sites) and Israeli Prime Minister Ariel Sharon (news - web sites) met to try to end the conflict in the Middle East on Tuesday but their talks were overshadowed by a bombing in Israel which killed 15 people.

Both Bush and Sharon, meeting amid a flurry of U.S. contacts with Middle Eastern states to address 19 months of Israeli-Palestinian violence, said reforms in the Palestinian

Figure 6. Reuters, May 7



Israeli Cabinet votes to retaliate

Thu May 9, 5:58 AM ET

Matthew Kalman USA TODAY

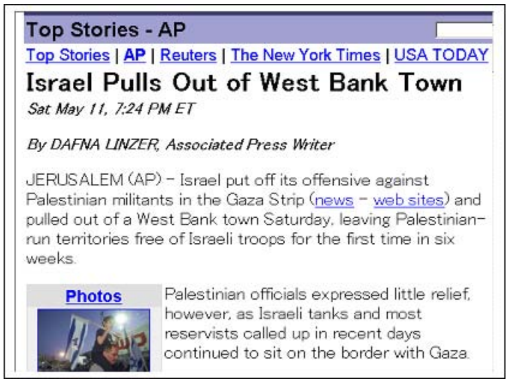
JERUSALEM -- The Israeli Cabinet today approved military action against "terrorist targets" in response to a suicide bombing that cut short Prime Minister Ariel Sharon (news - web sites)'s visit to the United States.

The decision to strike back came amid reports of a deal to allow most of the Palestinians holed up in the Church of the Nativity in Bethlehem for the past five weeks to leave.

Figure 7. USA Today, May 9

Table 4. Top TF*PDF Term (May 13 to May 19)

Term	Weight	Term	Weight	Term	Weight	Term	Weight	Term	Weight
1 Palestinian	844.29	attack	357.77	church	223.44	talk	175.59	House	163.43
2 Israeli	641.22	Sharon	334.56	Gaza	215.6	federal	173.4	Home	159.43
3 official	594.86	Arafat	281.56	Point	188.4	Washington	170.57	Student	150.5
4 Bush	469.92	American	269.9	Helder	186.6	Enron	168.83	Troop	150.11
5 bomb	431.2	kill	249.1	Peace	179.48	West	168.75	Minister	145.64
6 Israel	398.23	security	235.76	suicide	175.73	White	166.73	Letter	145.49



Top Stories - AP

Top Stories | AP | Reuters | The New York Times | USA TODAY

Israel Pulls Out of West Bank Town

Sat May 11, 7:24 PM ET

By DAFNA LINZER, Associated Press Writer

JERUSALEM (AP) - Israel put off its offensive against Palestinian militants in the Gaza Strip (news - web sites) and pulled out of a West Bank town Saturday, leaving Palestinian-run territories free of Israeli troops for the first time in six weeks.

Photos

Palestinian officials expressed little relief, however, as Israeli tanks and most reservists called up in recent days continued to sit on the border with Gaza.

Figure 8. AP, May 11

4.4. About Evaluation

Many type of measures [10, 12, 14, 16] have been proposed to evaluate Information Retrieval systems. Same to the automatic text summarization, evaluation is mainly developed by the developers to test their own systems. TIPSTER SUMMAC [7] was the initial work on developer-independent evaluation of automatic summarization system. However, our "cross-disciplines" system involves the information retrieval and summarization. The efficiency of TF*PDF algorithm has been tested. Summarization using sentence vector clustering is novel. The results from the samples run has proved to us the feasibility of this system doing more than fulfilling the minimum objective stated in this paper's title.

5. Conclusions

Our system is useful in summarizing the main topics in news archive periodically, should it be weekly, weeks-ly or monthly. We have shown that the TF*PDF algorithm performs well in extracting the terms explaining the main topics, by taking advantage on the concept that whenever there is a hot topic on air, the terms that explain the hot topics will appear frequently in many documents from multiple newswire sources. Also, we have introduced our approach in summarizing the main topics: sentence vector clustering. We have shown the effectiveness of our system by proper experiments done.

References

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

[2] K.B. Khoo and M. Ishizuka: "Information Area Tracking and Changes Summarizing in WWW" In: Proc. of WebNet 2001, International Conf. on WWW and Internet, pp. 680-685, Orlando, Florida. 2001

[3] S. Dharanipragada, M. Franz, J.S. McCarley, K. Papineni, S. Roukos, T. Ward, W.-J. Zhu: "Statistical Models for Topic Segmentation", In: Proc. of SIGIR '00

[4] J. Lafferty, D. Beeferman and A. Berger: "Statistical Models for Text Segmentation" In: Machine Learning, special issue on Natural Language Learning, C. Cardie and R. Mooney eds., 34(1-3), pp. 177-210, 1999

[5] Yiming Yang, Jaime G. Carbonell, et al.: "Learning Approaches for Detecting and Tracking News Events" In: IEEE Intelligent Systems, 1999, pp. 32-43.

[6] I. Mani and M. T. Maybury (eds): "Advances in Automatic Text Summarization" In: MIT Press, 1999.

[7] I. Mani, T. Firmin, D. House, G. Klein, B. Sundheim and L. Hirschman: "The TIPSTER SUMMAC Text Summarization Evaluation" In: Proc. of EACL-99, Bergen, Norway, June 1999

[8] J. Allan, R. Papka, and V. Lavrenko: "Online New Event Detection and Tracking" In: Proc. of SIGIR '98: 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, ACM Press, New York, 1998, pp. 37-45

[9] Yiming Yang, T. Pierce, and J. Carbonell.: "A study on retrospective and on-line event detection" In: Proc. of SIGIR '98: 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, ACM Press, pp. 28-36, 1998

[10] K.S. Jones and P. Willett, editors: "Readings in Information Retrieval" In: Morgan Kaufmann Publishing, San Francisco, 1997. Chapter 4, pp. 167-256

[11] B. Masland, G. Linoff, and D. Waltz : "Classifying news stories using memory based reasoning" In: Proc. of SIGIR '92, pp. 59-65, 1992

[12] J. Tague-Sutcliffe : "Measuring the informativeness of a retrieval process" In: Proc. of SIGIR '92, pp. 23-36, 1992.

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

[14] G. Salton: "Automatic Text Processing" In: Addison Wesley Publishing Co., Massachusetts, 1989.

[15] P. Hayes, L. Knecht, and M. Cellio: "A News Story Categorization System" In: Morgan Kaufmann Publishing, San Francisco, pp. 518-526, 1997. Originally appeared in Proceedings of the 2nd Conference on Applied Natural Language Processing 1988.

[16] http://trec.nist.gov/