

Topic Trend Detection and Mining in World Wide Web  
( WWW 上でのトピックトレンドの探知とマイニング )

by

Khoo Khyou Bun  
(48-17035)

A thesis presented to The University of Tokyo  
in fulfillment of the  
requirement for the degree of  
PhD  
in  
Information and Communication Engineering

Tokyo University, Japan

Supervisor  
Professor Mitsuru Ishizuka

©Khoo Khyou Bun, 2004

## Acknowledgements

For this dissertation, I am extremely fortunate to have worked with my research supervisor Professor Mitsuru Ishizuka. He is a great supervisor and provided me with constant encouragement and guidance. His willingness to show me the clear research direction so that I could excel in my work was always inspiring.

In the regular lab meetings, the fruitful and valuable ideas gained from the lab members on complementing this research are highly appreciated. The encouraging discussion was always the best way in reaching an optima solution for system flaw and the sources of fun as well.

I would like to take this chance to send my greatest respect to Professor Hitoshi Iba, Hiroshi Dohi Sensei, Fujita Tameko san and Dr. Helmut Prendinger for helping me up throughout the years. Also, my highest gratitude to all my seniors, Dr. Yasufumi Takama, Dr. Santi Saeyor, Dr. Itsuro Saito, Dr. Yutaka Matsuo, Dr. Matsumura Naohiro, Dr. Ma Shunli, Fukushima Shinichi, Tsutsui Takayuki, Inomata Kentaro, Yamamoto Tetsu and Yuan Zong. They have been providing me with constructive comments and advice.

Not forgetting, to my other lab members, Dr. Maxim S. Loukianov, Hironori Tomobe, Shin Ando, Nao Tokui, Tarou Yabuki, Adam Jatowt, Mori Junichiro, Makoto Iwashita, Naoaki Okazaki, Nakasone Arturo, Yang Zenglu, Ratanachai Sombatsrisomboon, Sohei Aya, Gakuto Kurata, Yohei Asada, Tomoya Taniguchi and Yustinus Juli, thank you very much from the bottom of my heart.

Last but not the least, to the friendships (name not specified) that company me and made me brave in walking through all the rain and shine in these few years of campus life are highly appreciated.

## Abstract

Ever since the Web information proliferation provides huge dynamically changing textual data online freely, the research in online topic detection and tracking is getting more important and challenging. The technology capable of capturing and analyzing the changes on the Web is no doubt vital, in providing the needed information in time for one to stay competent in this fast changing information age. This dissertation presents an approach toward the automatic journalism of new information (changes) on the Web. These information changes on the Web can be classified into two types: “flow” and “stock”. “Flow” type information (i.e. news) come to the Web constantly and regularly, at a rather fast pace. “Stock” type information, mainly the static web pages, change unpredictably doesn’t know at when and in what form. Our system aims to innovate the technology and use a new TF\*PDF (Term Frequency \* Proportional Document Frequency) algorithm to detect the prominent topics in the changes. In the framework and domain of problem addressed, this algorithm is more superior than the conventional TF\*IDF algorithm in a way that it doesn’t need retrospective corpus, besides posing minimal risk of losing the tracks of detection and tracking of popular topics. Also, our system requires less computational complexity while offering more flexibility. It crawls the Web, collects the changes and journalizes a summary of popular topics to the user. It does more than the conventional web tracking systems that just acknowledges the URLs of changed pages. It can become our personalized e-journalist on the Web and periodically provide us with the collection and e-publication of currently popular events.

## Table of Contents

Chapter 1 Introduction.....	1
1.1 Introduction .....	1
1.2 Structure of Dissertation .....	2
Chapter 2 Web Intelligence and Data Mining.....	4
2.1 Overview .....	4
2.2 Web Change Monitoring and Tracking Tools.....	6
2.2.1 ChangeDetector .....	6
2.2.2 TopBlend -- HtmlDiff .....	8
2.2.3 ChangeDetect.....	9
2.2.4 Webspector .....	10
2.2.5 WebBeholder .....	11
2.2.6 WebCQ.....	13
2.2.7 Infosphere Project .....	14
2.3 Web Crawling.....	15
2.3.1 General-purpose Web Crawler .....	15
2.3.2 Topic-Focused Web Crawling .....	16
2.3.3 Adaptive Crawling .....	18
2.4 Web Community Farming and Mining.....	18
2.4.1 Definition of Web Mining .....	18
2.4.2 Web Content Mining .....	19
2.4.3 Web Structure Mining .....	20
2.4.4 Web Usage Mining .....	21
2.5 Conclusion.....	22
Chapter 3 Topic Trends Detection and Tracking .....	24
3.1 Introduction .....	24
3.2 Information Retrieval, Extraction and Management.....	24
3.2.1 Difference of Information Retrieval and Extraction .....	24
3.2.2 Vector Space Model and Term Weighting .....	25
3.2.3 Relevance Feedback.....	25
3.2.4 Visual-Based Presentation.....	26
3.3 The Architecture of Topic Trends Detection and Tracking Systems.....	26

3.3.1 The Essential Components .....	27
3.3.2 Input Corpus and Topic Features Attributes.....	27
3.3.3 Morphological Analysis .....	28
3.3.4 Visualization.....	31
3.3.5 Evaluation.....	32
3.4 Overview of the Main Approaches.....	33
3.4.1 ThemeRiver .....	33
3.4.2 TimeMines .....	34
3.4.3 HDDI (Hierarchical Distributed Dynamic Indexing) .....	36
3.4.4 TDT – Topic Detection and Tracking.....	38
3.4.5 PatentMiner .....	39
3.5 Future Directions in TDT.....	41
Chapter 4 Automatic Online Journalism .....	43
4.1 Introduction.....	43
4.2 Document Clustering and Classification .....	43
4.2.1 Hierarchical Clustering .....	43
4.2.2 Partitioning Relocation Clustering.....	44
4.2.3 Density-Based Partitioning.....	45
4.3 Multi-document Summarization Methods .....	45
4.3.1 Statistical Techniques – Sentence Extraction and Ordering Heuristics .....	46
4.3.2 Natural Language Processing Approaches.....	47
4.3.3 Graph and Links Analysis Methodologies .....	47
4.4 Working Systems .....	48
4.4.1 Google News .....	48
4.4.2 NewsBlaster.....	49
4.4.3 NewsInEssence.....	49
4.5 Conclusion .....	50
Chapter 5 “Flow” Type Information Topic Detection and Summarization.....	51
5.1 Introduction.....	51
5.2 Fundamental Study on Topic Detection in News Archive.....	53
5.3 Approaches .....	56
5.3.1 Overview .....	57

5.3.2 Implementation of TF*PDF Algorithm.....	58
5.3.3 Sentence Vector Clustering .....	60
5.4 First Sample – Topic Selection and Topic Terms Ranking by TF*PDF Algorithm (Compare with TF) .....	62
5.4.1 Corpus (2003 July 20 ~ December 20).....	62
5.4.2 Topic Selection Algorithm .....	62
5.4.3 How about using TF (Term Frequency)? .....	68
5.4.4 Topic Terms Ranking using TF*PDF and TF .....	72
5.4.5 Evaluation for Terms Extraction by TF*PDF and TF.....	78
5.4.6 Graphical Ranking of TF*PDF and TF in Time Series.....	84
5.4.7 Conclusion.....	89
5.5 Second Sample – Weekly News Topics Summarization .....	90
5.5.1 Corpus .....	90
5.5.2 Experiment on Archive from May 13 to May 19.....	90
5.5.3 Experiment on archive dated from May 6 to May 12 .....	98
5.5.4 About Evaluation .....	100
5.6 Third Sample - Generating a Better-Coverage Summary of News Topics using Time Features and Sentence Clustering.....	101
5.6.1 A Better-Coverage Summary.....	101
5.6.2 Term Weight Acceleration .....	104
5.6.3 Topic Summary from Feb 1 to Feb 9 .....	106
5.6.4 Result Summary for the Time Frame from Feb 1 to Feb 9.....	107
5.7 The Uniqueness of Our Approach .....	108
5.8 The Merits of Our Approach.....	110
5.9 Conclusion.....	111
Chapter 6 Emerging Topic Tracking System (ETTS).....	113
6.1 Introduction .....	113
6.2 Related Works and Motivations .....	114
6.3 System Architecture.....	116
6.3.1 Area View System .....	117
6.3.2 Web Crawling.....	122
6.3.3 Change Summarizer .....	123

6.4 Samples Run .....	126
6.4.1 First Sample (Oct 1 to Dec 15, 2003 on the keywords “Asia Economy”) .....	126
6.4.2 Second Experiment Sample.....	130
6.4.3 Third Experiment Sample .....	134
6.5 Comparison to Related Works .....	140
6.6 Conclusion .....	142
Chapter 7 Conclusion.....	144
Publication Lists .....	153

## List of Figures

Figure 1 : Output Display of ChangeDetector .....	7
Figure 2 : Frame View of TopBlend .....	9
Figure 3 : Interface of ChangeDetect .....	10
Figure 4 : Interface of Webspector.....	11
Figure 5 : System Overview of WebBeholder .....	12
Figure 6 : Output Example of WebBeholder .....	13
Figure 7 : Sample Output of WebCQ.....	14
Figure 8 : Harvest Rate of Ordinary Crawler (source: UCB) .....	17
Figure 9 : Harvest Rate of Focused Crawler (source: UCB) .....	17
Figure 10 : ToughGraph .....	26
Figure 11: General architectural overview of TTD system .....	27
Figure 12 : Sample of DET Curve (source:TDT).....	32
Figure 13 : Sample Output of ThemeRiver.....	33
Figure 14 : Sample Output of TimeMines .....	35
Figure 15 : sLoc Process Discovering Concept Region in HDDI.....	37
Figure 16 : Sample Output of TDT .....	39
Figure 17 : Sample Output of PatentMiner.....	41
Figure 18 : Risk of TF*IDF – lost of tracks in the midst of hot topic .....	55
Figure 19 : System Information Flow.....	57
Figure 20 : Cosine Angle for Sentence Clustering.....	61
Figure 21 : TF*PDF Weight of Selected Terms from July 20 to Dec 20, 2003.....	63
Figure 22 : TF*PDF Weight of 5 Topic Terms from the First Topic.....	64
Figure 23 : TF*PDF Weight of 5 Topic Terms from the Second Topic.....	65
Figure 24 : TF*PDF Weight of 5 Topic Terms from the Third Topic .....	66
Figure 25 : TF Weight of Selected Terms from July 20 to Dec 20, 2003.....	68
Figure 26 : TF Weight of 5 Topic Terms from the First Topic.....	70
Figure 27 : TF Weight of 5 Topic Terms from the Second Topic.....	71
Figure 28 : TF Weight of 5 Topic Terms from the Third Topic .....	72
Figure 29: Example document with the term “amin” appearing many time.....	74
Figure 30: Example document with the term “syria” appearing 33 times .....	76
Figure 31: 100 – Ranking of “blackout” (constant top).....	85

Figure 32: 100 – Ranking of “power” (constant top).....	85
Figure 33: 100 – Ranking of “city” (constant top).....	85
Figure 34: 100 – Ranking of “amin” (drop out).....	86
Figure 35: 100 – Ranking of “schwarzenegger” (drop out).....	86
Figure 36: 100 – Ranking of “palestinian” (drop out).....	87
Figure 37: 100 – Ranking of “area” (push up).....	87
Figure 38: 100 – Ranking of “night” (push up) .....	88
Figure 39: 100 – Ranking of “failure” (push up) .....	88
Figure 40: 100 – Ranking of “line” (push up) .....	88
Figure 41 : USATODAY, May 17.....	96
Figure 42 : Reuters, May 17 .....	97
Figure 43 : Reuters, May 13 .....	97
Figure 44 : Reuters, May 7 .....	99
Figure 45 : USATODAY, May 9.....	99
Figure 46 : AP, May 11 .....	100
Figure 47 : Precision/Recall Mapping.....	101
Figure 48 : TF*PDF Weight of Selected Terms from Jan 29 to Feb 15.....	103
Figure 49 : ETTS System Architecture .....	117
Figure 50 : Salient Page Determination.....	119
Figure 51 : Domain Tree .....	121
Figure 52 : Information Cone Traversing Flow Chart.....	122
Figure 53 : CNN News May 18, 2000.....	134
Figure 54 : CNN News July 13, 2001 .....	134
Figure 55 : Percentage of Information Cones Vs Suitability .....	137
Figure 56 : CNN News April 17 2001 .....	140
Figure 57 : CNN News May 25 2001 .....	140

## List of Tables

Table 1 : Example Attributes Used for Topic Detection and Tracking .....	28
Table 2 : TF*PDF Weight of 5 Topic Terms from the First Topic .....	63
Table 3 : TF*PDF Weight of 5 Topic Terms from the Second Topic.....	65
Table 4 : TF*PDF Weight of 5 Topic Terms from the Third Topic.....	66
Table 5 : The value of W (Topic Weight) obtained using TF*PDF and TF.....	67
Table 6 : TF Weight of 5 Topic Terms from the First Topic .....	69
Table 7 : TF Weight of 5 Topic Terms from the Second Topic.....	70
Table 8 : TF Weight of 5 Topic Terms from the Third Topic .....	71
Table 9: Term Ranks at the peak time of First Topic (August 17).....	73
Table 10: Term Ranks at the peak time of Second Topic (Oct 9).....	75
Table 11: Term Ranks at the peak time of Third Topic (Dec 16).....	77
Table 12: Documents containing the term “area” (August 15~19, AP) .....	79
Table 13: Documents containing the term “area” (August 15~19, source NYT) .....	80
Table 14: Documents containing the term “area” (August 15~19, source Reuter) .....	82
Table 15: Documents containing the term “area” (August 15~19, source USATODAY) .....	82
Table 16: Ratio of topic document containing the term “area” (August 15~19) .....	83
Table 17 : Top TF*PDF Terms ( May 13 to May 19).....	91
Table 18 : 25 Highest Weighted Sentences (May 13 to May 19) .....	91
Table 19 : Sentence’s Unit Vector, Status, Date and Source .....	93
Table 20 : Top TF*PDF Terms (May 6 to May 12).....	98
Table 21: TF*PDF Weight of Selected Terms from Jan 29 to Feb 15 .....	103
Table 22 : Weight Acceleration of Selected Terms from Jan 31 to Feb 15.....	104
Table 23 : Logic for Calculating Term Weight Acceleration .....	105
Table 24 : Top 30 TF*PDF Terms from Feb 1 to Feb 9.....	106
Table 25 : Sentence’s Unit Vector, Date and Source (Feb 1 to Feb 9).....	107
Table 26 : A comparison between systems.....	110
Table 27 : A comparison of TF*IDF and TF*PDF .....	110
Table 28 : Merits of TF*PDF in Topic Detection and Tracking.....	111
Table 29 : Salient Pages in First Sample Experiment (keywords: “Asia economy”).....	127
Table 30: Top 20 TF*PDF Terms in two weeks interval change from Oct 1 to Dec 15, 2003 .....	128
Table 31: Sentences Extracted from change (Oct 1~Oct 15, 2003) .....	128

Table 32: Sentences Extracted from change (Oct 16~Oct 31, 2003) .....	129
Table 33: Sentences Extracted from change (Nov 1~Nov 15, 2003).....	129
Table 34: Sentences Extracted from change (Nov 16~Nov 30, 2003).....	130
Table 35: Sentences Extracted from change (Dec 1~Dec 15, 2003).....	130
Table 36 : Second Experiment Salient Pages (keywords: “nuclear weapon”).....	132
Table 37 : TF*PDF Term Weights (keywords: “nuclear weapon”).....	133
Table 38 : Resultant Summary for the keywords of “nuclear weapon” .....	133
Table 39 : Third Experiment Salient Page (keyword: “e-commerce”) .....	135
Table 40 : TF*PDF Term Weights (period between Oct 3 and Nov 3, 2000).....	138
Table 41 : TF*PDF Term Weights (period between Nov 3 and Dec 4, 2000) .....	138
Table 42 :Resultant Summary for the keyword of “e-commence” .....	139



# Chapter 1 Introduction

## 1.1 Introduction

In this fast changing information age, knowledge of emerging trends in the a particular area of interest is essential for ones to stay current or be competitive, should they be companies, research scholars, movie or film producers, fashion designer and follower, and etc regardless of their profession. A company market analyst may always want to mine and review the new information on some products for their trends from different points of view. For example, the mobile phone in Japan trends to be flip-able, color screen and having camera with high resolution; humanoid robot has been the popular trend and gained much attentions. Also, by knowing the preference of the consumer from different age group or different interests, the company analyst may predict the market trend of a product according to the population trend of each age group. In general, a user can know the way some particular pages have changed since it was last viewed; consumers can use the technology to compare new products, services or auctions on E-commerce websites; researchers of a particular subject area can keep updated by monitoring websites of related conferences, journals, or scholars; companies can monitor evolution of their competitors' websites to discover their new directions or offerings and etc.

Research on topic trends detection and tracking exploits many similar techniques used in data mining, for example linguistic and statistical features, learning algorithms, clustering and classification technologies, and visualization. This research field has become important after the Web, which itself the most popular communication media nowadays, has evolved into the largest information pool in the world. The information in this pool or the Web is changing dynamically and speedily. This nature of the Web provides great challenges in the research field to invent the state-of-art technology, which is able to discover the evolving features and concepts of interesting topics from

the changes by its content, structure and distribution over time. This research area is regarded as a branch of Web Mining research which taking into consideration the textual data's time features.

Basically, topic trend detection software applications input a collection of temporal textual data and recognize the topic trend in time series. It can be viewed as a branch of research in Text or Web Mining involving the time features analysis. Many researchers have proposed different measurement as the significant of a topic at a particular time point. These measurements can be information gain, topic term weight, topic cluster size, number of documents containing the topic keyword, and etc. Higher the value of this measurement at a certain time, more important the topic is likely to be. These topics significant are usually measured towards a bigger time unit, for example day, week, month or year. This is different to the time-sensitive stock prices that may move drastically in minutes or even seconds. We don't report a topic of "stock price" movement in minutes range, but we may track the topic of the stock's drastic move in minutes range if it become popular and become news for a certain period.

Recent research shows much effort and progress have been made in innovating the automatic topic trend detection and tracking. While many of the systems are semi-automatic and require user interaction in the process, some of them are fully automated. The semi-automatic system needs user input and feedback to start and produce output. On the other hand, fully automated system would generate output without much user intervention. However, regardless of full or semi-automatic, most of these systems rely on human domain expert to judge if the topic trend is emerging and interesting to the users. Seeing that the Web keeps growing rapidly, both in its size and complexity, the research in this area will remain important and pose great challenges continuously.

## **1.2 Structure of Dissertation**

This dissertation is organized into 7 chapters:

**Chapter 1:** Overview and introduction to the needs and importance of Topic Tracking and Mining in WWW.

**Chapter 2:** The state of the art of some Web change monitoring and tracking systems are studied. Further, Web crawling as well as web community forming and mining technologies are summarized.

**Chapter 3:** This part focuses on topic trends detection and tracking technologies. The nuts and bolts technologies required for topic trends detection and tracking are presented, followed by the insights of a few representative systems.

**Chapter 4:** What and why automatic online journalism is explained, together with the two major processes of document clustering and multi-document summarization. The advanced systems of Google News, NewsBlaster and NewsInEssence are also discussed with some details.

**Chapter 5:** “Flow” type information topic detection and summarization methodologies are proposed. With details and supportive experimental results, Topic selection and TF\*PDF algorithms for topic weighting and topic terms detection will be introduced, followed by illustration of sentence clustering methodology and a better-coverage concept in generating a temporal news topic summary.

**Chapter 6:** In this chapter, we propose the Emerging Topic Tracking System (ETTS), which is useful in tracking and summarizing the changes in an information area of our interest on the Web, with details of its system architecture and samples run.

**Chapter 7:** Conclusion, importance and significance of this work, and ideas are presented.

## Chapter 2 Web Intelligence and Data Mining

### 2.1 Overview

The Web (World Wide Web) is a global network of hypertext documents. These documents are also called web pages. They are basically text files of content text intervened in structural HTML (Hypertext Markup Language) tags. These HTML tags describe how the text of a web page should be formatted when it is displayed on our computer screen. Each of these web pages has a unique URL (Uniform Resource Locator) and thus can be linked between each other. These links are called hyperlinks. By keying in a web page's URL in a Web browser (i.e. Microsoft Internet Explorer), we can retrieve and view the content of the page. Thus, URL is the address of a web page on the Web. The URL of a web page can be tagged and embedded into other web pages, which instead can also be embedded in other web pages. As a result, this produce a web of huge number hypertext documents where users can surf by clicking on the hyperlinks to look for the information needed.

This Web of hypertext documents spread speedily in 1990s to all over the world. Now, it is not only the most important global information infrastructure or information channel for disseminating knowledge, it is also the biggest information storage and service platform ever. It contains billions of web pages and this number is still growing. The "simple to create" characteristic of web page is one of the main reasons responsible for the rapid wide-spreading of the Web. It allows everyone from anywhere anytime to author and pose personal web pages on the Web for sharing globally. This makes the Web highly diversified with all kind contents and complicated link structure. Thus, the Web provides the greatest challenges for information science researcher in 21<sup>st</sup> century, in order to bring more intelligence and introduce the Web, by then the Web of Intelligence, into our society seamlessly.

Towards achieving a Web of Intelligence, more information automation needed to be invented to accomplish efficient information search and knowledge queries. The Web is a now vital medium for conducting business, disseminating, sharing and publishing information through many portal-based decentralized service provider. Many ideas have been proposed by researchers to make these Web services collaborative and communicative in order to form a Web of Intelligence where its content and services understandable to machine, not just to human where the current Web do. To this end, the Web inventor, Tim Berners-Lee invented the idea of Semantic Web [TBLee01], the next generation Web where semantic markup languages defined to represent machine-understandable semantic content. This Semantic Web bases on XML (eXtensible Markup Language) together with ontology would empower the next generation Web which possible more information automation. By then, it would be possible for us to have a virtual secretary, who can, for example, know our time schedule, search and compare the airfare according to our budget, make a reservation and also book a hotel room according to our preference, without human intervention. At the same time made inroads into semantically machine-understandable and Web services automate-able, the Web will have its current pool of HTML contents/knowledge, which is designed for human remained and growing as well, since this is the simpler way for information publishing and sharing. Thus, although the Web and Web services will become more autonomous, the core technologies for mining the Web involving techniques like crawling, searching, aggregation, classification, filtering and pattern recognizing would still be important and need advancements.

As we know, one of the main growing aspects of the Web will still be the proliferation of its HTML contents/knowledge. When the Web keeps growing, with new information dynamically posted on it, the technologies for acquiring the latest meaning information will become more and more important. Therefore, Web change monitoring and tracking techniques need be enhanced to

cope up with the difficulties in providing us the valuable new information from the change, in order to keep us updated and competitive all the time.

## **2.2 Web Change Monitoring and Tracking Tools**

Web surfing, reading, blogging, trading, publishing and many others Web activities have become a part of our daily life, after a decade of information booming letting many individuals and organization have their own Web site. In order to stay ahead, some corporations even see having their own web site a must for information servicing, such as announcing company's latest movements or make a place for easy information exchanging. Hence, many new web pages will appear at the same time many others will be updated frequently. As a result, there are many information portals constantly providing dynamically changing "fresh" information on the Web. This fresh information can be very valuable for the users, so there is a need for the users to know this newest information as soon as the change happened, by checking the particular web pages constantly and frequently. This kind of activity of checking the web pages for changes is called "Web Monitoring".

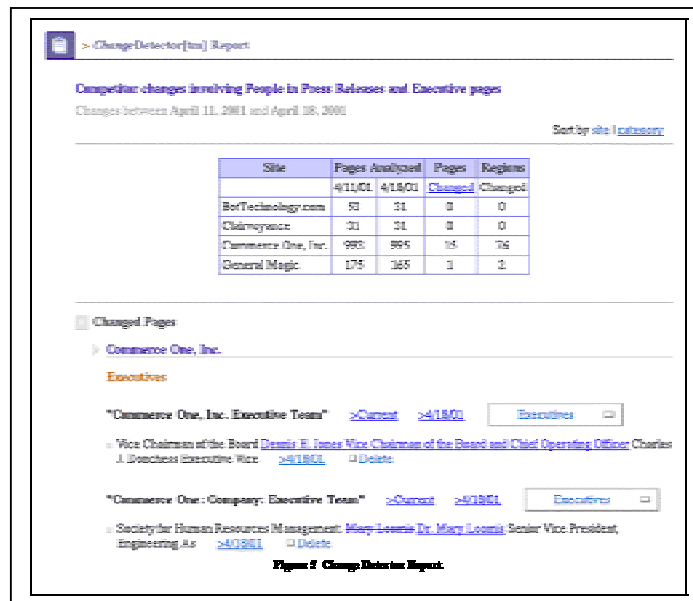
However, monitoring web pages manually is not efficient. This is because human tempts to make errors and may miss the important changes, especially when checking many pages in a frequent manual is rather impossible. Thus, we should rely this important job on some intelligent Web monitoring agents or software robots.

### **2.2.1 ChangeDetector**

ChangeDetector [boyapati02] is a system not only can track the changes on a page basis, it is designed to monitor a whole website. For any changes happen within the web site, it would be detected and reported to the user. In this way, it can inform some changes which is difficult for user to find manually, for example company group organization restructuring or the end of some production lines. The ChangeDetector technology is based on machine learning. By adapting intelligent crawling

techniques, ChangeDetector could gather all the related information of a web site efficiently even though the web site is huge. Then from the information gathered, ChangeDetector will retrieve the new information (or changes) according to its topic by using entity extraction, semantic filtering and categorization techniques. The prototype system of ChangeDetector could monitor more than 2000 web sites in one week.

The collected web pages are classified and indexed into 12 predefined categories, such as press release, contact information, recruiting information, business and etc. Based on the labeling from page index and website's directory structure, ChangeDetector would automatically generate a sitemap, which is useful for intelligent crawling, so that it would give priorities to some important pages and crawl the website more efficient next time.



The screenshot shows a web browser window displaying a report titled "> ChangeDetector[Inc] Report". The main heading is "Competitor changes involving People in Press Releases and Executive pages" with a subtitle "Changes between April 11, 2001 and April 28, 2001". A "Sort by site | category" link is visible. Below this is a table with columns: Site, Pages Analyzed, Pages Changed, and Pages Deleted. The table lists four sites: BoTechnology.com, Claimsync.com, Commerce One, Inc., and General Magic. Below the table, there is a section for "Changed Pages" with a sub-section for "Commerce One, Inc." and a sub-sub-section for "Executives". This section lists specific changes, such as "Commerce One, Inc. Executive Team" and "Vice Chairman of the Board Dennis H. Jones", each with a "Current" link, a date, and a "Delete" button. The report is captioned "Figure 1: ChangeDetector Report" at the bottom.

Site	Pages Analyzed	Pages Changed	Pages Deleted
BoTechnology.com	93	31	0
Claimsync.com	31	31	0
Commerce One, Inc.	993	995	76
General Magic	175	165	1

**Figure 1 : Output Display of ChangeDetector**

Therefore, ChangeDetector could monitor the whole website for changes and report it to the users via email or web browser. Figure 1 illustrates a sample report by ChangeDetector. From this report

display, user can know how many pages have been analyzed and how many have incurred changes in a website. At the bottom of the output display, user can find a detailed list of changes happened in the website.

### **2.2.2 TopBlend -- HtmlDiff**

TopBlend [chen00] is a HTML differencing tool, which can detect the changes happened on a page by doing page comparison by HCS (Heaviest Common Subsequence). TopBlend was implemented in Java and used the fast Jacobson-Vo [jacobson92] algorithm to solve the HCS problem for page changes comparison. One of the merits of TopBlend is it allows comparisons to be performed either on the server or client side, which could offload busy servers duty in performing heavy computations. TopBlend proposes two methodologies in displaying the output, either a merged HTML view or a more convenient side-by-side view for web pages. Side-by-side comparison view is suitable for easy changes detection on the web pages with complex graphics designs. TopBlend has evolved from and been integrated with the AT&T Internet Difference Engine (AIDE) [dougkis98] from AT&T research lab. AIDE adapted the LCS (Logest Common Subsequence) [hirschberg77], which is used in the well-known Unix diff program. TopBlend can be considered a re-implementation of AIDE after introducing Java applet, Jacobson-Vo algorithm and side-by-side comparison display methodology.

Figure 2 shows an output frame view of TopBlend generated by HCS algorithm. The upper frame shows the list of difference. The lower frame shows a comparison of the new and old pages, by highlighting the changes with a difference colors (pink and gray in this example, given a reason that these colors are not popular and rarely used by most of the web pages).



**Figure 2 : Frame View of TopBlend**

### **2.2.3 ChangeDetect**

ChangeDetect [changedetect] is a free but efficient web changes monitoring service. It is a replacement for a few former commercial tracking tools, which have been discontinued, such as Netmind, Mindit and Syponit. It saves user's favorite web pages, monitors content for changes and sends an automatic email notification to the user whenever the web pages are updated. What has actually changed on web page text will be marked with color-coded highlights for easy understanding. web page change notifications can be delivered to the user via email, ICQ, text message or even pager.

ChangeDetect provides a registration "shortcut", a small utility script in the form of a browser's toolbar icon or a bookmark in Favorites Menu, on which when the user click on will register the page being surfed automatically. This is convenient because the user has no longer needed to visit the server's homepage and register the page to be monitored by keying in its URL. This is one of its good features that so far no other monitoring tools serving.

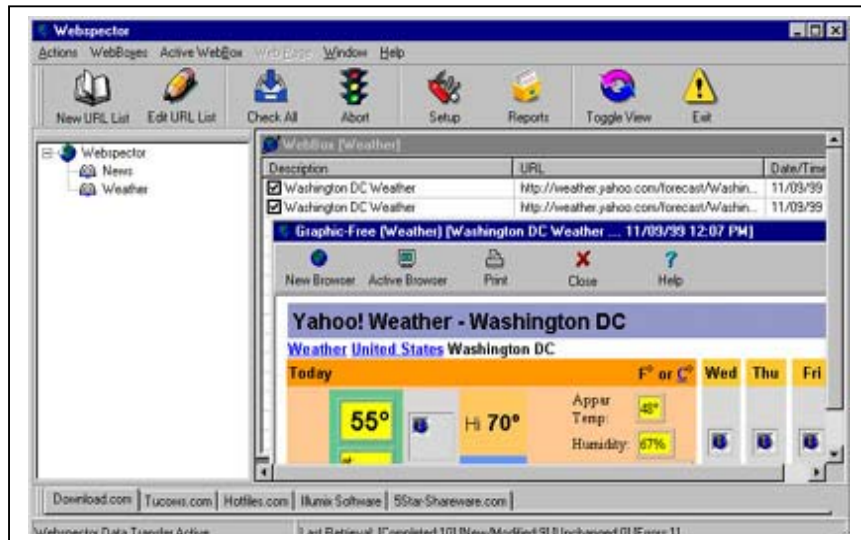
Besides, ChangeDetect allows user to set the notification trigger, by specifying some keywords combination with AND/OR rules. This allows user to decide whether to be notified when the changes on the monitored pages comprise or without a certain phrase. Figure 3 shows the interface of ChangeDetect for user to input the URL of the page to be monitored.



**Figure 3 : Interface of ChangeDetect**

#### **2.2.4 Webspector**

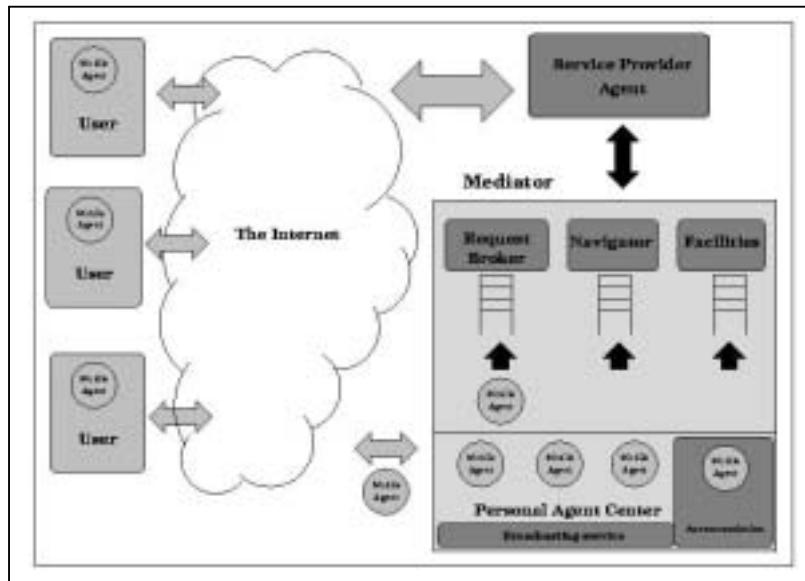
Webspector [webspector] is another tool useful for tracking web pages for content changes. Different from other server-based monitoring tools, Webspector is a client application that can be installed in user's personal computer and thus allows unlimited number of web pages monitoring depends on the computational power of the PC. Besides being able to send notification emails, Webspector also provides an interface to registering web pages URL and displaying the changes results. In addition, with Webspector user can set the time schedule for the checking task to be executed, and also have the changes or special keywords highlighted for intuitive visualization. Figure 4 is an example output display of Webspector.



**Figure 4 : Interface of Webspector**

### 2.2.5 WebBeholder

WebBeholder [santi98] is a multi agent web monitoring system developed in Ishizuka-lab from Tokyo University. System structure of WebBeholder is as shown in Figure 5. It consists of Service Provider Agent, a number of mobile agents and Mediator. The Service Provider Agent is responsible for monitoring the changes on the Web, while the mobile personal agent will be useful for customizing user requests. Next, Mediator will be responsible for dealing and negotiating the services between mobile agents and Service Provider.



**Figure 5 : System Overview of WebBeholder**

Mediator contains three service modules. They are Request Broker, Navigator and Facilitator. The Request Broker is needed to convey the service requests from mobile agents to the Service Provider, at the same time coordinating the common requests among the mobile agents. Then, the task of the Navigator is to deliver the information about other WebBeholder communities to the mobile agents. Lastly the Facilitator is purposed to provide various capabilities to the newly added mobile agents.

WebBeholder bases on LOCTAGS (Longest Common Tag Sequence) algorithm to retrieve the changes on HTML pages. Different tags in the changes will have different weight depends on their importance, for example a change in the URL will gain more points than a change on some text. When total points of changes on a page exceed the trigger threshold, the user will be notified. Figure 6 is an output example from WebBeholder, the deleted information is cancelled with cross lines while the new information will be placed at the near side.



Figure 6 : Output Example of WebBeholder

### 2.2.6 WebCQ

WebCQ [webcq, ling00] is Web information changes detecting and delivering system developed in Georgia Tech. It is capable of monitoring and tracking the changes happening on both the static and dynamic web pages. Types of changes on web pages are classified in sentinels, and this possible the more granular detecting and result presentation. Therefore, this system can deliver user with personalized changes results, coz the users are allowed to decide the kind of sentinels for monitoring.

WebCQ explores some object extraction algorithms to locate and identify the object of interests on web pages. Algorithms used can detect changes to arbitrary objects and compute how a page has changed. In addition, proxy cache service is provided to minimize access latency, reduce the workload of remote information servers, and achieve higher robustness of its change monitoring service. Lastly, Notification service is offered with a set of enhanced capabilities to provide both

server-initiated push delivery and client-initiated pull delivery of changes, at the same time presenting rich and pleasant display to the user. Figure 7 is an sample output from WebCQ. It tells the changes on weather report by displaying old and new page side-by-side with the changes highlighted.



**Figure 7 : Sample Output of WebCQ**

### 2.2.7 Infosphere Project

Infosphere project [infosphere] is part of DARPA Information Technology Expedition program. This project has been developed at Georgia Tech led by Calton Pu. Infosphere project focuses on bringing fresh information from a variety of sensors to the user as personalized fresh information delivery. They develop concepts, techniques and tools to support end-to-end quality of service (QoS), in terms of freshness, performance, availability and maintainability. Part of the concepts and techniques were implemented in the WebCQ introduced the previous session. In contrast to the well known of remote procedure call (RPC), the concept of Infopipe was proposed. In addition to their basic function of transporting information, Infopipes can control the delivery properties of

information such as freshness and performance mentioned above. They call the software architecture implementing this information flows in Infopipes as *producer/consumer* architecture, in contrast to the traditional client/server architecture. Information generated by a producer is carried to the consumers by Infopipes.

## **2.3 Web Crawling**

WWW provides us with great amounts of useful information electronically available as hypertext. This large pool of hypertext is changing dynamically and semantically unstructured, making us finding the related and valuable information difficult. Therefore, a web crawler for automatic discovering of valuable information from the Web, or Web Mining is important for us nowadays. In reality, this web crawler is a program, which automatically traverses the web by downloading documents and following links from page to page. They are mainly used by search engine to gather data for indexing. Other possible applications include page validation, structural analysis and visualization, update notification, mirroring and personal web assistants/agents etc. Web crawlers are also known as spiders, robots, worms etc.

### **2.3.1 General-purpose Web Crawler**

General-purpose web crawlers collect and process the entire contents of the Web in a centralized location, so that it can be indexed in advance to be able to respond to many user queries. In the early stage when the Web is still not very large, simple or random crawling method was enough to index the whole web. However, after the Web has grown very large, a crawler can have large coverage but rarely refresh its crawls, or a crawler can have good coverage and fast refresh rates but not have good ranking functions or support advanced query capabilities that need more processing power. Therefore, more advance crawling methodologies are needed due to the limited resources like time and network

bandwidth. Topic-Focused Crawling and Adaptive Crawling will be described in the following two sub-session.

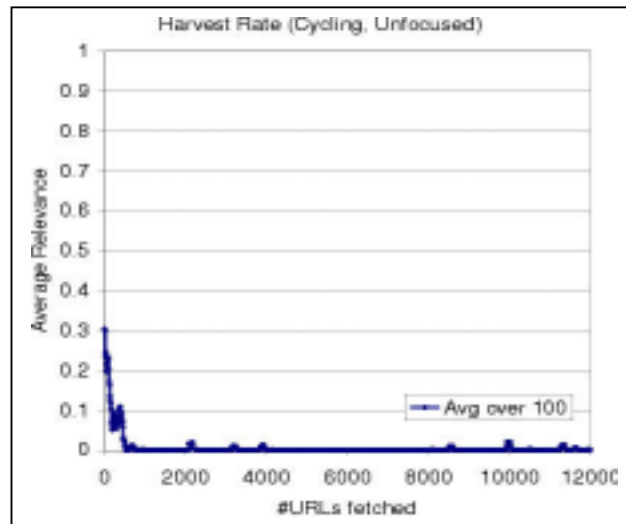
### **2.3.2 Topic-Focused Web Crawling**

Topic-Focused Web Crawling [chakra99, michelan00, chakra02] initiation was motivated by the fact the Web is huge with an unprecedented scaling problem, but most people are only interested in a small fraction of the Web. The main objective is to only crawl on a small fraction of the Web to discover the set of pages covering a certain topic. This is essential because of the finite crawling resources such as time, network bandwidth and storage as mentioned above. The major web crawlers harness dozens of powerful processors and hundreds of gigabytes of storage using superbly crafted software, and yet cover 30-40% of the web. Scaling up the operation may be feasible, but rather not appropriate because the sheer diversity of content at a generic search site snares all but the most crafty queries in irrelevant results.

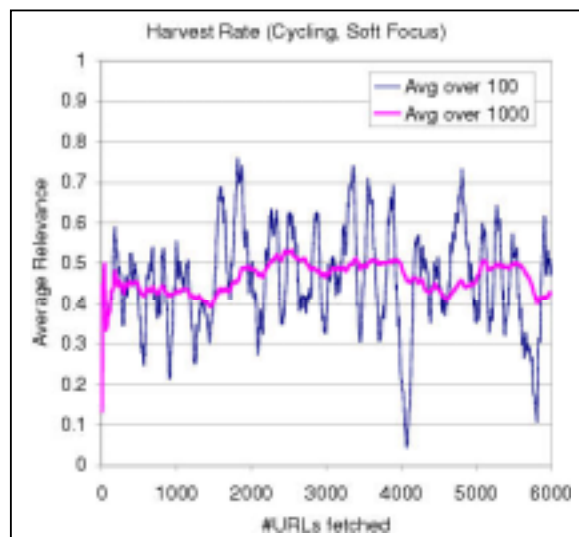
Therefore, automatic focused crawler can, starting from a handful of examples on a specific topic, while minimizing crawling time and space in irrelevant and/or low-quality regions of the web, perform a lightweight crawling activity at the level of individuals and interest groups. There can be hundreds of thousands of focused crawlers covering diverse areas of web information, tailored to the needs of specific communities.

One of the basic components of focused crawler would be hypertext topic classifier that calculates the score or relevance of a page. This relevance score reflects how interesting the page is, given the topic of the crawl. Different focused crawler uses different approach or hypothesis to calculate the importance of a page. Having continually calculates this rating can help the crawler to estimate the benefit of crawling out from the pages' out-links. The main challenge is to ensure a high *harvest rate*: the fraction of page fetches which are relevant to the user's interest. Without the hypertext classifier,

an ordinary crawler would be very difficult to keep on track. Figure 8 below shows the decaying harvest rate of an ordinary crawler, while Figure 9 illustrates that almost half of the pages fetched by focused crawler are relevant.



**Figure 8 : Harvest Rate of Ordinary Crawler (source: UCB)**



**Figure 9 : Harvest Rate of Focused Crawler (source: UCB)**

### **2.3.3 Adaptive Crawling**

Adaptive crawler [edward00] is classified as an incremental type of crawler which will continually crawl the entire web, based on some set of crawling cycles. The adaptive model used would use data from previous cycles to decide which pages should be checked for updates. Adaptive Crawling can also be viewed as an extension of focused crawling technology. It has the basic concept of doing focus crawling with additional adaptive crawling ability. Since the web is changing dynamically, adaptive crawler is designed to crawl the web more dynamically, by additionally taking into consideration more important parameters such as freshness or up to date-ness, whether pages are obsolete, the way pages change, when pages will change, how often pages change and etc. These parameters will be added into the optimization model for controlling the crawling strategy, and contribute to defining the discrete time period and crawling cycle. Therefore, it is expected that more cycles the adaptive crawler goes in operation, more reliable and refined will the output results. This is different from the result shown in the previous sub-section, where the focused crawler will gain an average harvest rate of 0.5.

## **2.4 Web Community Farming and Mining**

A web community is a collection of web pages sharing the same interest. Members of a web community may be unaware of the existence of each other (and may be even unaware of the existence of the community). Identifying web communities and understanding emergence and evolution of web communities are very important. The tasks of identifying and understand the communities on the web can be regarded as web farming and mining.

### **2.4.1 Definition of Web Mining**

The conventional word Mining means extracting something useful or valuable from a baser substance, such as mining gold from the earth. Analogously, Web Mining means extracting valuable

information from the data gathered by traditional data mining methodologies and techniques over the World Wide Web. Basically, Web mining technologies look for patterns in data through three main aspects, which are content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines, crawlers or spiders. Structure mining is used to examine useful linkage information related to the structure of a particular Web site. Lastly, web usage mining is used to examine data related to a particular user's browser, its browsing history as well as data gathered by forms the user may have submitted during Web transactions.

#### **2.4.2 Web Content Mining**

Web content mining focuses on automatic discovery of information resource available online, and involves mining web data content. Unlike Web Usage Mining or Web Structure Mining, Web Content Mining emphasis on the content of the web page which is not necessary only the text, but also the multimedia content such as image, metadata and hyperlinks.

In the Web mining domain, web content mining is regarded as an analog of data mining techniques for relational databases, because it is possible to find similar types of knowledge from the unstructured data residing in Web documents. Some of the Web documents are semi-structured such as HTML documents, or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of Web data makes the Web content mining a complicated and challenging task.

Web content mining can be approached from two different ways: Information Retrieval and Database. Kosala [kosala00] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words to represent unstructured text, and take single word found in the training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some

utilized the hyperlink structure between the documents for document representation. As for the database view, in order to have the better information management and querying on the Web, the mining always tries to infer the structure of the Web site and transform a Web site to become a database.

Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from the large online multimedia sources. Multimedia data mining on the Web has gained many researchers' attention recently. Working towards a unifying framework for representation, problem solving, and learning from multimedia is really a challenge, this research area is still in its infancy indeed, many works are waiting to be done.

### **2.4.3 Web Structure Mining**

The goal of Web structure mining is to discover the structural information about the Web. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

In general, if a Web page is linked to another Web page directly, or the Web pages are neighbors, we would like to discover the relationships among those Web pages. The relations maybe fall in one of the types, such as they related by synonyms or ontology, they may have similar contents, both of them may sit in the same Web server therefore created by the same person. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlink in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain, therefore the query processing will be easier and more efficient.

## PageRank Algorithm - Google

Google [google], based on PageRank algorithm [brin98], is today's successful and pioneer search engine. It has taken over the market by using links as the primary method of determining the value and thereby the deserved visibility of a web site. Google interprets links to a web page as objective, peer-endorsed and machine-readable signs of value.

Google indexes links between web sites and interprets a link from A to B as an endorsed of B by A. Links may have different values. If A has a lot of links to it, and C has few links to it, then a link from A to B is worth more than a link from C to B. The value determined in this way is called a page's PageRank and determines its placement in search results. (The PageRank is used in addition to conventional text indexing to generate highly accurate search results.) Links can be analyzed more accurately and usefully than traffic or page views, and have become both measure of success and dispensers of rank.

PageRank is not simply based upon the total number of inbound links. The basic approach of PageRank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally. A document ranks high in terms of PageRank, if other high ranking documents link to it. So, within the PageRank concept, the rank of a document is given by the rank of those documents which link to it. Their rank again is given by the rank of documents which link to them.

### **2.4.4 Web Usage Mining**

Web usage mining technology is used to discover user navigation patterns from web data, tries to discover the useful information from the secondary data derived from the interactions of the users

while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with Web.

#### 2.4.4.1 User Profiling

Users leave footprint in the web server when they access the web site. These footprints record the important information about the users browse record such as date, client IP, time taken, request, protocol and so on. All these information is helpful in analyzing user interests and profile. Many online stores make use of user profiling to promote and target their products to the specified user, for an example, Amazon tends to recommend a user with the books or music, which are sought by other users who have similar user profile.

#### 2.4.4.2 Web Log Mining

Many information left behind when user visits a web site. Server log files, which are simple text files will be generated automatically every time someone accesses a Website. Every "hit" to the Web site, including each view of a HTML document, image or other object, is logged. This web log file contains information about who was visiting the site, where they came from, and exactly what they were doing on the web site. Together with the Web content and Web site topology, Web log can be used for doing prediction of the user's behavior within a site, and comparison between expected and actual Web site usage, and hence possible to adjust a Web site to best suit the interests of its users.

## 2.5 Conclusion

A number of Web changes monitoring tools and systems offer solutions to detect and report the dynamic information changes on the Web. Using the state-of-art technology of Web mining: Web content mining, Web structure mining and Web usage mining, these systems are useful and productive for us in revealing the various meaningful changes on the Web. However, the capability of

23

most of these systems is rather limited to reporting the well-defined type of changes, such as acknowledging the URL of the page containing a certain keyword in its changes. Seeing that the Web will continue booming and producing much new information on it, active research should continue on better ways to create intelligence to understand the changes well and report more useful knowledge embedded in the changes.

## **Chapter 3 Topic Trends Detection and Tracking**

### **3.1 Introduction**

The automatic detection of emerging trends and new events online has become an important research topic as the huge amount of electronically available information on the Web keeps on changing dynamically. At anytime, there would be new information coming to the Web and worth to be tracked and reported to the user who will be interested in it. However, achieving this goal needs broad-based technology covering from Information Retrieval, Information Extraction and Management.

### **3.2 Information Retrieval, Extraction and Management**

The huge amount of information on the Web keeps on proliferating at a fast speed. This information appears in many forms (images, text, video, and speech), and its increase leads to information overload because there are no means for separating relevant from irrelevant information. In order to make use of this information, whether for business or leisure purpose, Information Retrieval (IR), Information Extraction (IE) and management techniques (tools) are required to allow for fast, effective and efficient access to large amounts of stored information.

#### **3.2.1 Difference of Information Retrieval and Extraction**

As an instance of Web mining, IR deals with resource or document discovery on the Web. It is a task of automatic retrieval of all relevant documents while at the same time retrieving as few of the non-relevant as possible. On the other hand, IE has the goal of transforming a collection of documents, usually with the help of an IR system, into information that is more readily digested and analyzed [kosala00, cowie96]. Thus, while IE's goal is to extract relevant facts at the same time taking care of

the structure and presentation of a document, IR tries to select relevant documents by mainly considering the document keywords.

### **3.2.2 Vector Space Model and Term Weighting**

Vector Space Model is a representation of documents and queries where they are converted into vectors. The features of these vectors are usually words in the document or query, after stemming and removing stop words. Stemming is the process of removing prefixes and suffixes from words; stop words are words such as a preposition or article that has little semantic content. The vectors of document and query are weighted to give emphasis to terms that exemplify meaning. Then, after the query vector is compared to each document vector, those that are the closest to the query are considered to be similar, and are returned as a result. SMART [salton71] is the most famous example of a system that uses a vector space model.

Term weighting is the process of giving emphasis to the parameters for more important terms. In a vector space model, this is applied to the features of each vector. A popular weighting scheme is TF\*IDF [salton89]. Other possible schemes are Boolean (1 if the term appears, 0 if not), or by term frequency alone. In a vector model, the weights are sometimes normalized to sum to 1, or by dividing by the square root of the sum of their squares.

### **3.2.3 Relevance Feedback**

Relevance Feedback is a process of refining the results of a retrieval using a given query. The user indicates which documents from those returned are most relevant to his query. The system typically tries to find terms common to that subset, and adds them to the old query. It then returns more documents using the revised query. This can be repeated as often as desired iteratively. It can also be recognized as "find similar documents" or "query by example".

### 3.2.4 Visual-Based Presentation

Visual-based presentation is one of the most effective and intuitive ways to deliver retrieval result to the user. Graphical user interface (GUI) of a few changes tracking tools has been shown in section 2.2. These GUIs are important for the user to look for the wanted information in a short time. Figure 10 below is an output example of TouchGraph [touchgraph], which is useful for user to visualize a network of inter-related information, such as a web community. In TouchGraph, the networks of information are rendered in interactive graphs, where user can interact with mouse and explore different ways of displaying the network's components on screen for easy navigation. By engaging in this visual image, user can navigate through large networks and exploit the mutual-relationship among the members in a community.

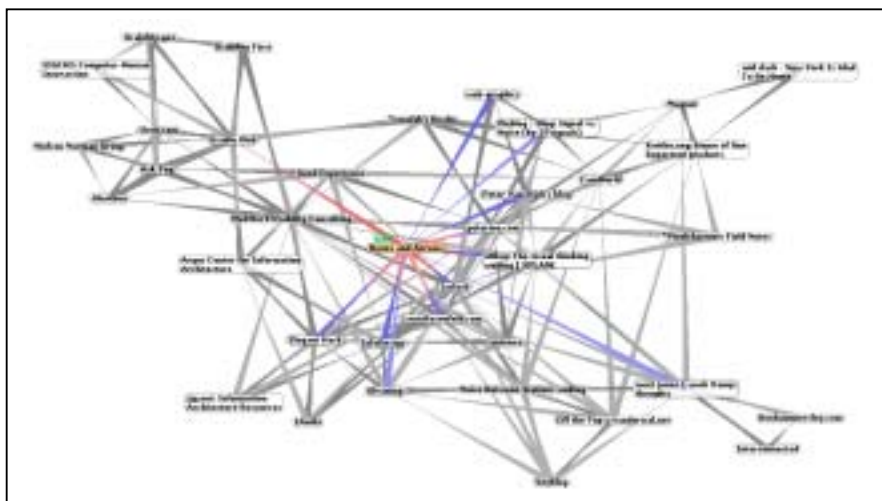


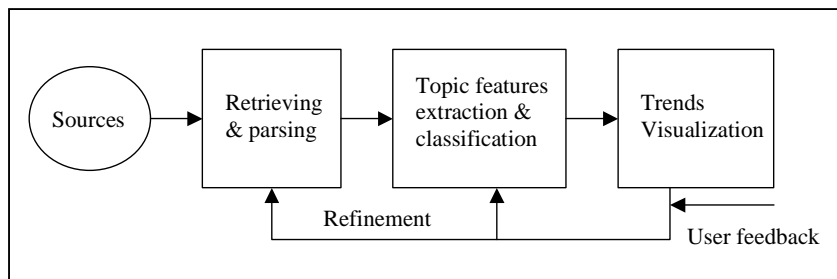
Figure 10 : ToughGraph

### 3.3 The Architecture of Topic Trends Detection and Tracking Systems

In order to cope with the diversity of digital information available nowadays, for example web logs, news, hypertext, database and etc, many approach and architecture have been invented by researchers to examine and track the topic trends in their domain accordingly. However, most of them need to

deal with the same set of problem and go through some similar processes. They first need to carry out an IR task to retrieve the needed information from their source to build the input corpus. Then IE techniques are used to analyze the corpus, should it be statistical or linguistic methodologies. After this, mostly of the system may also apply some learning algorithms or AI techniques to deduce the important topics and trends. Lastly, a visualized presentation for the result will be a plus.

### 3.3.1 The Essential Components



**Figure 11: General architectural overview of TTD system**

Figure 11 above summarizes the fundamental building of a general Topic Trend Detection and Tracking system. Basically, there are 4 necessary components: sources or input corpus, retrieving & parsing, topic features extraction & classification and trends visualization. Some advanced systems supports the user relevant feedback facility, and able to refine the output result according to user's interest.

### 3.3.2 Input Corpus and Topic Features Attributes

Many test data corpus sets are available for research use in addressing problem of discovering patterns and trend analysis. These test corpuses consist of data in different format, vary from patent and text databases, tagged news and speech documents, to hypertext. Evolving from IBM Patent Server, Delphion [delphion] is an independent company formed jointly by IBM and Internet Capital Group (ICG) to service information on US patents database (and others), which has been widely used

to exploit technology trends discovering and analyzing. Another well-known technical literature database is INSPEC [inspec], which provides accesses to the bibliography of scientific information. Also, Topic Detection and Tracking (TDT) corpus is a popular and commonly used test data. This TDT corpus has been developed since 1997 as a project under DARPA. It is built from data streams such as newswire and broadcast news from Reuters and CNN. Several test data sets (TDT-pilot, TDT2 and TDT3) have been created. Each test sets are sets of news stories and event descriptors, which is assigned a relevance judgment by human to indicate the relevance of the given news stories to the event. Later, algorithms and researches are developed to detect the event by threading the related documents.

The input data set together with the selected topic features attributes is an important part of TTD system. Recognizing and processing the topic features attributes over time is the ultimate step in doing the detection and tracking. Table below shows a few possible attributes.

**Table 1 : Example Attributes Used for Topic Detection and Tracking**

<b>Attributes</b>	<b>Description</b>
Named Entity	Name Person, Location, Organization and etc
n-Grams	Phrase/sequence of n words, can have different definition such as noun phrases, noun and adjectives combinations, nouns with gaps and etc
Unigram	Single word, significance usually measured in weight basis
Time	Granularity depends on system, can be hour, day, month, year and etc
Concept	Special knowledge calculated and defined by certain IE techniques or learning algorithms depends on the system

### 3.3.3 Morphological Analysis

Morphology is word formation and morphological analysis is the study of syntax or sentence structure, which is an important process in features extraction, especially from raw text, using NLP

techniques. There are three major types of morphological process: inflection, derivation and compounding. Inflections are the systematic modifications of a root form by means of prefixes and suffixes to indicate grammatical difference like singular and plural. Inflection process will not change word class or meaning significantly, but varies features such as tense, number, and plurality. Differently, derivation is less systematic and may results more radical change of syntactic category, and possibly cause a change in meaning. Some examples of derivations: the suffix *-en* transforms adjectives in verbs (sharp-en, dark-en), the suffix *-able* transforms verbs into adjectives (access-able, read-able), and the suffix *-er* transforms verbs into nouns (heat-er, sing-er). Lastly, compounding is about the merging of two or more words into a new word. English has many noun-noun combinations, nouns phrases that consist of two other nouns or more. Examples are cable car, computer desk, or glass bottle. Although written as separate words, they are pronounced as a single word, and denote a single semantic concept, which we may wish to list in the lexicon.

### 3.3.3.1 Statistical Name Tagger

Tagging is a task of labeling the words in text with appropriate predefined tags. Although the ultimate goal of research NLP is to parse and understand language, Statistical Name Tagging is just an intermediate task merely makes sense of the structure inherent in language without complete understanding. Therefore, tagging is a common first step needed to prepare a precisely tagged corpus for use in features extraction in the next step of topic features (terms) detection and tracking. The information that can be revealed by the tags is like article, preposition, adjective and adverb just to name some of them. Different tagging scheme will have a different number and definition of tag set. For the purpose of topic tracking, it is also needed to have tags that reveal time information of the data, for example the time when the new information appeared, or the time some information changed and evolved. One of the most influential tag sets is the one used for tagging the American Brown corpus (the Brown tag set). More recently are the tag sets used for tagging British National Corpus

(c5) and Penn Treebank, which is a simplified version of the Brown tag sets. These tag sets are all for English and in general, tag sets incorporate morphological distinctions of a particular language, and so are not directly applicable to other languages.

### 3.3.3.2 Name Entity Recognizer

One important capability for more precise text understanding is that name recognizer must know what words (names) to extract and what grammars to be deployed for this purpose. So, a way to address name entity recognition in the topic detection and tracking task needs us to produce a grammar that includes all the names that we want to trace. However, comprehensive name recognizer is possible to be deployed with the grammar generated from database or other information source. [fien03] introduces a memory-based approach to learning names in un-annotated newspaper. Some grammatical rules would be deployed to make a list of name, which is then used to construct an additional feature for training the machine learning algorithms. Borthwick [borthwick98] has been contributing by introducing MENE (Maximum Entropy Named-Entity), which uses statistical methodologies and able to achieve high scores comparatively. This work was presented in the 7<sup>th</sup> Messages Understanding Conference (MUC-7), in which the task was to identify all the “names” falling into one of the seven categories: person, organization, location, data, time, percentage, and momentary amount. Although been the simplest messages understanding task, Name-Entity Recognition is very important for Information Extraction System and the same too to topic detection and tracking system.

### 3.3.3.3 Hidden Markov Model

Markov Model is a probabilistic function of a Markov process, which is useful in modern speech recognition systems, and thus equally important in help to recognize the topic for detection and tracking in text processing. More particular, English is a language that can be modeled with n-gram

models or Markov chains. These models are ones assumed to be having a limited memory, and the probability of the next word depends only on the previous word or a limited number of words. For an example, when given the preceding  $k$  words, we can try to guess what is the next word in a sentence, and this has been proved as fairly easy. Also, on the basis of having looked at a lot of text, we can know which words tend to follow other words. What we need is a method of grouping histories that are similar in some way so as to give reasonable predictions as to which words we can expect to come next. Markov assumption that the last few words would affect the next word is one possible way to do the grouping. Therefore, Markov model (also called  $n$ -gram model) can be implemented to extract the phrases ( $n$ -grams) with the features that we want, for detecting and tracking the appearing of the features over time. HMM (Hidden Markov Model) operates at a higher level of abstraction by postulating additional “hidden” structure, and allows us to look at the order of categories of words. In an HMM, we don’t know the state sequence that the model passes through, but only some probabilistic function of it. Therefore, HMM is useful when we can think of underlying events probabilistically generating surface events and one widespread use of this is tagging.

### **3.3.4 Visualization**

Great effort and progress in the research of visualization-based topic detection and tracking have produced many precise and efficient techniques. When a user is trying to analyze and summarize a large amount of data, a system that can present an overview, at multiple levels of detail will be very helpful. Histogram can be one of the simplest approaches, with its bars indicate discrete values of the significant topic attributes over time for intuitive understanding. Information visualization is meant to help the user to understand the trend of a topic better, by plotting the patterns along a timeline allows one to see the rate of change of a pattern over time. In section 3.4, visualization approaches of some popular system will be introduced.

### 3.3.5 Evaluation

Precision and recall has been the standard and formal evaluation methodology in IR problem. Since the nature of the problem is similar, evaluation of many topic detection and tracking systems have also be measured with precision and recall. Precision: Percentage/ratio of the selected items that the system got right; recall: the proportion of the target items that the system discovered. Higher the value of precision and recall, better is the performance of the system. Opposite to precision and recall, some system use miss (false negative) and false alarm (false positive) for evaluation, by using Detection Error Tradeoff (DET) curves [martin97]. DET curves highlight how miss and false alarm rates vary with respectively when plotted on a plane. A perfect system with zero miss and false alarm would be positioned at the origin, thus, closer a result curve to the origin better is the result. Figure 12 below is a sample output of DET curve.

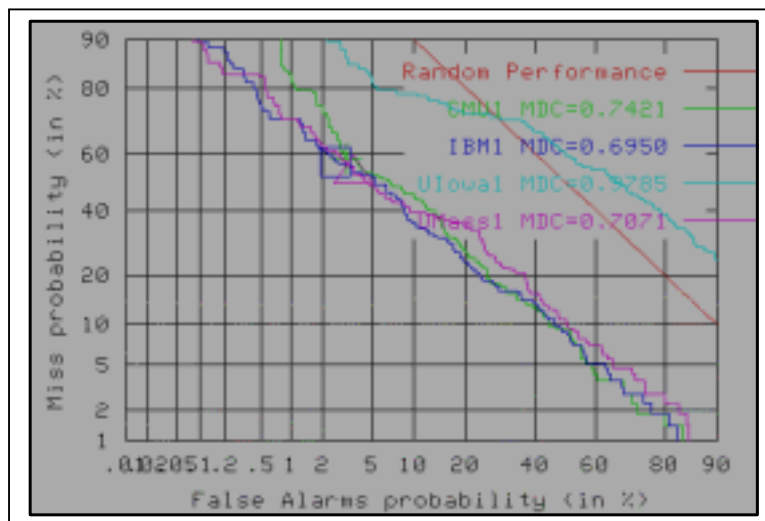
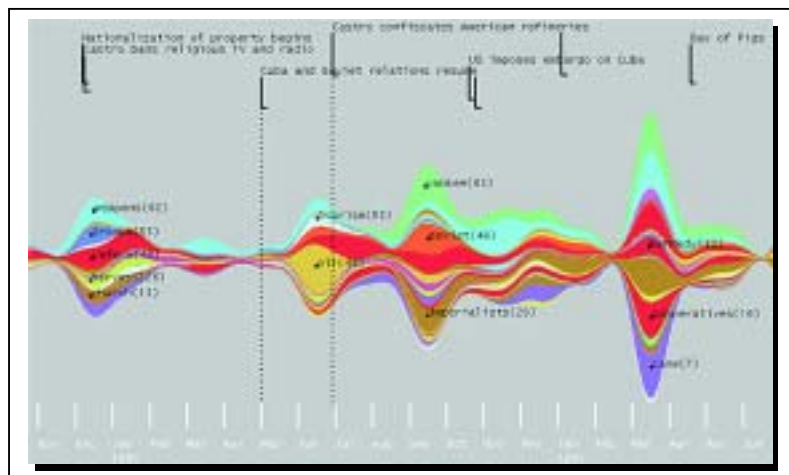


Figure 12 : Sample of DET Curve (source:TDT)

### 3.4 Overview of the Main Approaches

#### 3.4.1 ThemeRiver

ThemeRiver [havre02] is a system that search and visualizes the trends, patterns and relationships of thematic variations over time across a collection of documents. The “river” flows through time, changing width to depict changes in the thematic strength of documents temporally collected. Themes or topics are represented as colored “currents” flowing within the river that narrow or widen to indicate decreases or increases in the strength of a topic in associated documents at a specific point in time. The river is shown within the context of a timeline and a corresponding textual presentation of external events. Output from ThemeRiver is shown in Figure 13.



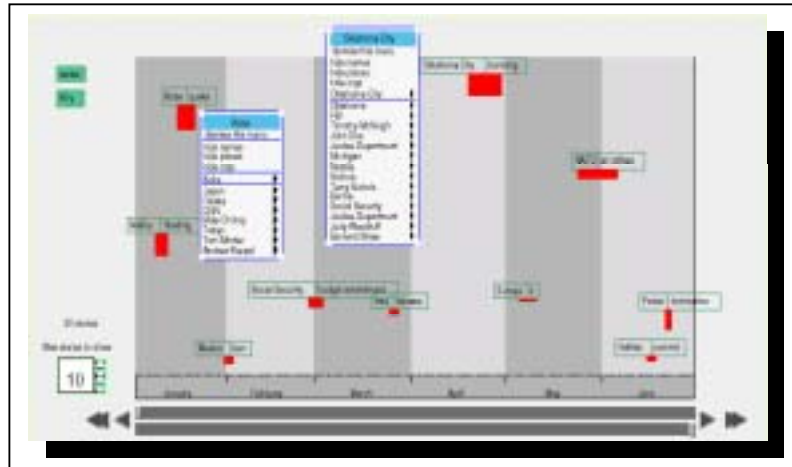
**Figure 13 : Sample Output of ThemeRiver**

ThemeRiver allows the user to track the importance of a topic over time. Figure 13 shows the output of ThemeRiver for the corpus of speeches by Cuba's Fidel Castro in years 1960 and 1961. The output shows that Castro frequently talked about oil just before American oil refineries were confiscated in 1960, indicated by the second line in the figure. The topic of oil is represented by the largest “circle”, which is preceding the second vertical line in the figure.

The central idea of ThemeRiver was to use graphical representation for topics detection and tracking. The topics (called theme words) were generated automatically from the corpus (but no detail of the methodology was specified). Of each topic, a subset of attributes was chosen manually. The counts of number of documents containing a particular theme word for each time interval become the input for the graphical presentation. ThemeRiver provides a view of the data that an experienced domain expert can use to confirm or refute a hypothesis about the topic data.

### **3.4.2 TimeMines**

The main idea of TimeMines is to present a graphical overview timeline of statistically significant topics from a corpus, which is free text with explicit date tags. Figure 14 below illustrates a sample graphical output from TimeMines [swan00]. This timeline is generated automatically for visualization of temporal locality of topics and the identification of new information within each topic. The x-axis represents time while the y-axis represents the importance of a topic. At the top of the figure are the most significant topics and the less significant topics will appear in lower part. There will be many “topic blocks” in the visualization interface that contain the terms explaining their topic respectively. These terms are usually named entity or n-gram. Users are allowed to interface by clicking on the terms, for a pop up menu of associated features of the topics. Any of these features can be selected as a label for each topic. Furthermore, users can view more information of a feature in its sub-menu.



**Figure 14 : Sample Output of TimeMines**

As for the input corpus for the experiment, TDT and TDT-2 were used. Badger IE system [fisher95] was applied to generate an initial list of all “name entities” and noun phrases. Named entity is defined as a specified person, location, or organization, while noun phrases are n-grams match the regular expression (N\$|J)\*N for up to 5 words, where N is noun, J is an adjective, \$|J is associate, and \* indicates zero or more occurrences.

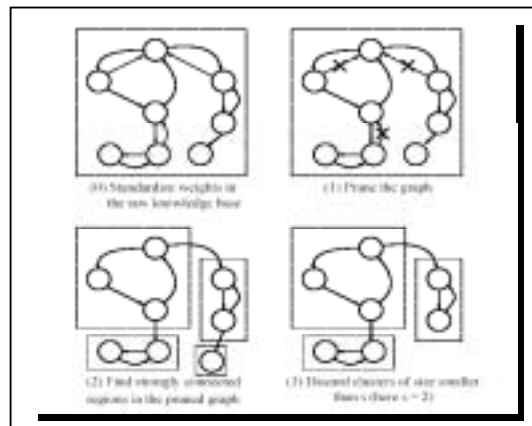
The main techniques in TimeMines are Information Extraction and Natural Language Processing (NLP) for processing the input data, and for grouping the relevant topics into a given time frame, hypothesis-testing techniques was used. This system employed a default model of features distribution on a base rate of occurrence, which does not vary with time. For each document, its feature will be compared to the default model and a statistical test will be carried out to decide if the document feature is “unexpectedly” different from the default model. Then, these features will be tested again to find out if they appear in the same time frames to form a single topic. Lastly, only the topics with importance higher the threshold will be displayed in the timeline user interface (Figure 14). This system is fully automated and able to generate a graphical representation of topics for a given time-tagged corpus.

### 3.4.3 HDDI (Hierarchical Distributed Dynamic Indexing)

HDDI [pottenger01] is an approach to organizing large quantities of unstructured data in a loosely coupled distributed environment under development at Lehigh University and at the National Center for Supercomputing Applications. The approach is based on the algorithmic creation of subtopic regions of semantic locality in sets of distributed documents; this allows automatic discovery of similarities at a fine level of granularity amongst concepts within documents. The following steps are performed: concept identification/extraction, concept co-occurrence matrix formation, hierarchy construction, knowledge base creation, identification of regions of semantic locality and hierarchy mapping. First, part of speech tagging approach is done or identifying various parts of the speech. Next, a finite-state machine extract complex noun phrases (concepts) according to the regular expression  $C?(G\$|P\$|J\$)*N+(I*D?C?(G\$|P\$|J\$)*N+)^*$ , where C is a cardinal number, G a verb (gerund or present participle), P a verb (past participle), J an adjective, N a noun, I a preposition and D a determiner, ? indicates zero or one occurrence,  $\$|$  indicates union, \* indicates zero or more, and + indicates one or more occurrence. In the concept co-occurrence matrix formation, “Co-occurring” defines concepts that occur within the same item, which is for example abstracts, titles, web pages, patents and etc.

Given the extracted concepts, they compute concept frequency and co-occurrence matrices. Then, systematic filtering, pruning, and meshing were carried out to produce a higher level of combined matrices, in a recursively manual. Then, for each concept in each matrix they compute a similarity with other concepts for a mapping that quantitatively determines how similar they are semantically. The resultant mapping is a graph in which nodes are concepts and arc weights are similarity measures. This concepts graph consists of regions of high-density clusters of concepts - subtopics regions of semantic locality. These regions consist of clusters of concepts that commonly appear together and collectively create a knowledge neighborhood. In the next step, for grouping similar concepts together,

the sLoc algorithm was used. They called sLoc the contextual transitivity in the similarity relation. It also means that a threshold is chosen based on the structure and distribution of the similarities and transitivity is constrained accordingly.



**Figure 15 : sLoc Process Discovering Concept Region in HDDI**

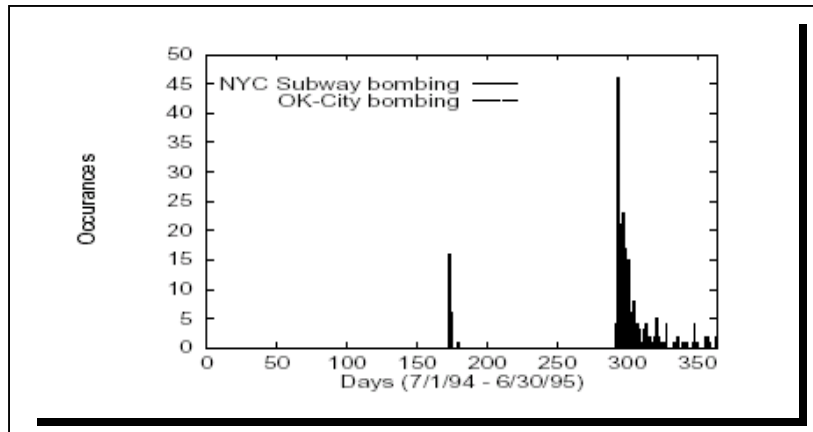
Figure 15 shows the steps in the sLoc process. Prior to the first step, the weights in graph (also named knowledge base) are normalized (step 0 in the figure). The first step in sLoc is to statistically prune the input graph. Arcs of weight smaller than a certain threshold  $T$  are virtually pruned. The second step involves the identification of the clusters within the graph. Tarjan's algorithm [tarjan72] is applied to find strongly connected regions, by using a variant of depth-first search. At this stage each strongly connected region is a cluster. The size of a given cluster is the number of nodes (concepts) it contains. During the third and final step, clusters of size smaller than parameters  $s$  are discarded. They interpret the remaining clusters as regions of semantic locality in the knowledge base. The greater  $T$ , the more arcs are cut off, and therefore the smaller in size the strongly connected regions. Thus the greater  $T$  the smaller in size and the more focused will be the regions of semantic locality. A good value of  $T$  could yield an optimum clustering. They used statistical heuristic to

determine the optimum  $T$  as a function ( $T: T(x) = \text{Max}(w) - x * SD$ ) of mean  $M$  and standard deviation  $SD$  derived from the arc weight distribution.

This system automatically detect emerging topic by tracing changes over time in concepts frequency and association, by taking a snapshot of the statistical state of a collection at multiple points in time. Two features of an emerging topic are: first it should be semantically richer at a later time, and second it should occur more frequently as an increasing number of items (documents). The semantic richness is judged by the number of other concepts that appear in the same region of semantic locality. Machine learning techniques were adapted to evaluate the concept for an emerging topic. In their experiment, the input features (i.e. number of occurrences of a concept in the trial year, the year before, two years before and etc) were inputted to a  $7*10*2$  neural network. This was a learning model for achieving better recall (just as a radar system depends on good recall) because domain experts will do the final filtering for an emerging topic.

#### **3.4.4 TDT – Topic Detection and Tracking**

TDT project sponsored by DARPA began in 1997. TDT research develops algorithm for detecting new event in data stream from online and broadcast news. The TDT data repository has been the most commonly used corpus, which was initially in English, and then developed to include Chinese and Arabic in TDT2 and TDT Phase 3 (TDT3). The pilot TDT corpus comprises of 15,863 news stories spanning from July 1 1994 to June 30 1995, from Reuters news and CNN broadcast.



**Figure 16 : Sample Output of TDT**

TDT detects the news stories that discuss an event that has not been reported in earlier stories. They used statistical algorithm to extract the keywords in a news story and compare the story with earlier stories. The decision of a new event detection on a news document based on a single pass clustering algorithm [allan98]. The content of each story is represented as a query and if the story query does not trigger any previous query by exceeding its threshold, the story will be defined as a new story. Figure 16 illustrates an event burst of a news story. The x-axis represents time in terms of days and the y-axis the story count per day. News stories discussing the same event tend to be in temporal proximity. Therefore, lexical similarity and temporal similarity are utilized to do document clustering according to topic, and stories that appear in the same timeframe are likely to be matched.

### 3.4.5 PatentMiner

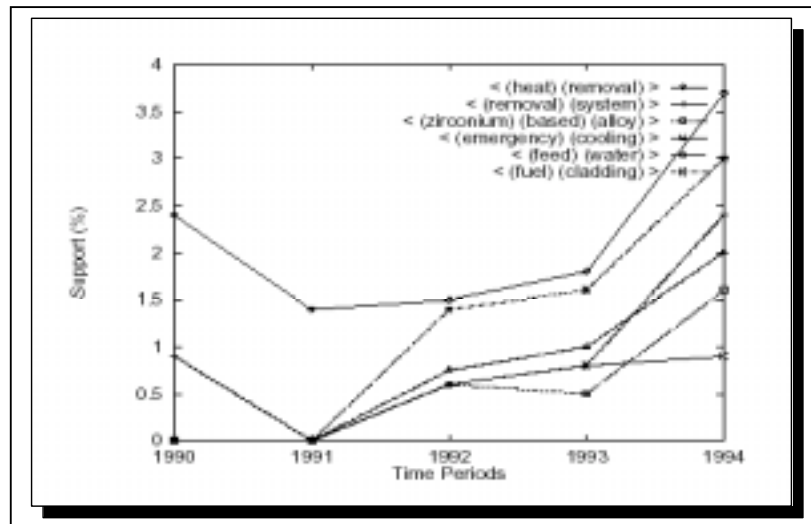
The PatentMiner [brian97] system was developed to discover trends in patent data. The system uses IBM DB2 database (Delphion) containing all granted United States patents. The two main components of the system are: phrases identification using sequential pattern mining, and trend detection using shape queries. Several procedures were done to prepare data processing like stop words removing, identification of positions and occurrences of sentences, paragraphs and section boundaries in documents. Next, a subset of patents is categorized by category and date range, and

GSP algorithm selects user-defined attributes, called “phrases”. GSP stands for Generalized Sequential Patterns [srikant96]. By phrase they understand any sequential of words with a minimum or maximum gap between any of the words, where gap can be described in terms of words, sentences, paragraphs or sections. For example, if the minimum sentence gap is one for the phrase “data mining”, then words “data” and “mining” must occur in separate sentences. Finally, the Shape Definition Languages (SDL) [agrawal95] specifies what kind of trends will be displayed. The possible type of shapes can be, for example upwards or spikes.

For a given database D of documents, each document consists of one or more text fields and a timestamp. The unit of text is a word and a phrase is a list of words. Associated with each phrase is a history of the frequency of occurrence of the phrase, obtained by partitioning the documents based upon their timestamps. The frequency of occurrence in a particular time period is the number of documents that contain the phrase. A trend is a specific subsequence of the history of a phrase that satisfies the users’ query over the histories. For example, the user may specify a “spike” to find those phrases having frequency of occurrence increased and then decreased.

PatentMiner adapts a sequential pattern matching techniques that is frequently used in data mining systems. The pattern matching system searches for frequently occurring word patterns. The word can be adjacent or separated by a variable number of other words predefined by user. The used techniques identify the frequently co-occurring terms and treat them as a single topic. The resulting topic set of words is regarded as a phrase. As with TimeMines, documents in the input data set are categorized into various groups based on their date information. Next, they extract phrases from each group and the frequency of occurrence of each phrase in all groups is calculated. Shape query was used to determine which phrases to extract, based on the user’s inquiry. This shape query allows user to graphically define various shapes (such as “recent upwards trend”, “recent spike in usage”, “downwards trend”, and “resurgence of usage”) for trend detection and retrieves the phrases with

frequency distributions that match the query. Alternatively, users can define their own shape by using a visual shape editor.



**Figure 17 : Sample Output of PatentMiner**

Figure 17 is an example output from PatentMiner, showing some trends found from U.S. Patents classified in the category “Induced Nuclear Reactions: Processes, Systems, and Elements”. These example phrases matched a shape query that represented an increasing trend of their usage in recent years. Without knowing a priori the kind of patents filed in this category, they were able to look at the trends and determine some of the popular topics of recently granted patents.

### 3.5 Future Directions in TDT

Although the future of any research area is difficult to be predicted, we are here to outline a few future prospects of Topic Detection and Tracking that need promotions and are likely to take place. Although there are much interesting works of visualization, it can be anticipated that more state-of-the-art intuitive GUI will be innovated. We expect a sophisticated interface that would allow users to access the relevant data sources, manipulate the underlying text, and get a display of the results in a

meaningful way. The visualization scheme should be able to draw the user's attention and attract them to drill into the data to find out what have led the topic trends development. Other than visualization supports, efforts are seen being put in developing topic detection and tracking algorithms with high precision and recall achievement. Up to date, statistical methodologies have been widely applied because of its simplicity, but with the advancements of natural language processing techniques nowadays, linguistic approaches should be taken into consideration for implementation for better results. In addition, most of the systems that have been developed are designed for processing the manually-build corpus with much metadata information, if not only the precise time tags. Therefore, it is no doubt that the robust techniques for topic trends detection and tracking in raw text such as hypertext are getting more importance. Lastly, at the time shifting into the ubiquitous computing environment, a good reporting scheme is also needed to automatically alert the user anytime anywhere, when new developments are happening in a specific area of interests, so that the user can be all-time aware of the new development which is critical for decision making.

## Chapter 4 Automatic Online Journalism

### 4.1 Introduction

Since the advent of the Web, access to unlimited news sites/feeds is just a click away from us. However, at the same time facing the problem of wasting time re-visiting the same piece of news information, we may be drowned and miss the interesting pieces in the mountains of information. Realizing this, works were started for technology innovation to overcome this information barrier, by creating some AI tools to automatically find, gather, sort, refine and present the intelligences behind the online news to the user efficiently – Automatic Online Journalism. In order to achieve this objective, there are two major tasks involved: first is IR, which is basically a document clustering or classification task, to group the relevant news articles from thousands of sources online, and second is to summarize the event in each topic group (multi-document summarization).

### 4.2 Document Clustering and Classification

Document clustering techniques are used to classify a collection of text into groups or clusters of documents having similar contents. The similarity between documents is usually measured with the associative coefficients from the vector space model (Section 3.2), e.g., the cosine coefficient. The widely used document clustering method is hierarchical clustering algorithm described in the next sub-section. Besides, partitioning relocation clustering and density-based clustering method will also be briefed at the following.

#### 4.2.1 Hierarchical Clustering

Hierarchical clustering builds a cluster hierarchy. This cluster hierarchy is in fact a tree of clusters, which is also known as a dendrogram. This tree of clusters can be built either bottom-up, by starting

with one-point (singleton) clusters and grouping the most similar ones recursively, or top-down, whereby one starts with all the objects and divides them into groups so as to maximize within-group similarity. These bottom-up and top-down methodologies are also called agglomerative and divisive way of clustering [jain88, kaufman90]. These agglomerative and divisive processes will continue until a stage where, usually a pre-defined number  $k$  of clusters formed. Hierarchical clustering is good for its embedded flexibility regarding the level of granularity and ease of handling of any forms of similarity or distance. However, the greatest drawback of hierarchical clustering algorithm is the vagueness of termination criteria and the clusters will not be re-visited once constructed.

#### **4.2.2 Partitioning Relocation Clustering**

Unlike hierarchical method, partitioning clustering can gradually improve clusters by applying greedy heuristics and optimize the clusters iteratively. More specifically, partitioning relocating clustering methods divide data into several subsets, and then apply relocation schemes to iteratively re-allocate them in between. The division of these subsets starts out with partitions based on randomly selected seeds (one seed per subset or cluster). Most algorithms will employ several passes to refine the clusters whereas hierarchical algorithms do only once. Then the question is how many passes needed or when the iterative refinery process should stop. This can be determined based on a measure of goodness or the cluster's quality, such as group-average similarity or mutual information between adjacent clusters. Nevertheless, the most important stopping criteria can be the measure of goodness improves after each iteration, and stop when the curve of improvement flattens or when goodness starts decreasing. Among many of the partitioning relocation clustering methodologies and algorithms,  $k$ -means [hartigan75] and  $k$ -medoids [kaufman90] have been popular and widely implemented.

### **4.2.3 Density-Based Partitioning**

In density-based partitioning, a cluster is defined as a connected dense component, which grows in any direction that density leads. Therefore, density-based algorithms are capable of and good in discovering clusters or arbitrary shapes. It can be an implementation of the idea of Euclidean space been divided into a set of its connected components, which requires concepts of density, connectivity and boundary.

## **4.3 Multi-document Summarization Methods**

Recently with the growth of available online information new summarization problems have appeared. As a further step of document summarization researchers have begun to search for methods suited for summarizing several documents simultaneously. Multi-document summarization problem requires extraction of similarities and differences between multiple documents in order to generate single output. The difficulty of multi-document summarization increases together with increased content diversity of documents in question. Intuitively, input documents should be topically close for obtaining the meaningful output. The number of documents to be summarized can range from couples of documents to the large collections. Multi-document summarization puts more stress on issues of speed, size, text unit selection and user goals than single-document summarization. Sentence ordering arouses to be challenging problem mainly in case of documents discussing events in different time orders. Additionally, documents constituting the collection may have different coverage, style and quality hence farther complicating the task.

There are different types of relations between documents discussing common topic. Common ways to characterize inter-document relationships are diversities of perspectives, details or temporal features of documents discussing the same event. Events can be described in different perspectives, levels of detail or in different points in time. There are important issues that need to be tackled in order to generate satisfying results. Due to large quantities of available text it is significant to

eliminate redundancy between texts. Another question is an ordering method for summarization, which would be based on temporal or semantic differences among documents. In case of www space there is already a powerful relationship clue to be exploited in form of the linking information between entities belonging to the collection. However text documents are treated separately and their relationships can be usually discovered by statistical or nature language processing methods.

#### **4.3.1 Statistical Techniques – Sentence Extraction and Ordering Heuristics**

For domain dependent summarization purposes techniques derived from information extraction research field can be employed. Mckeown and Radev [mckeown95] advocate a template-based technique for multi-document summaries. In their example documents about terrorist attacks are parsed so that pre-defined information slots are filled. Then relationships between different events are established through the comparison of templates. In last step, authors aggregate templates to construct higher-level merged templates. In spite of the high effectiveness of the approach, the main method's drawback is its domain dependent. Moreover, it is important to note that only specific types of documents can be applied to template creation.

Maximal marginal relevance (MMR) algorithm was proposed for domain independent summarization tasks [goldstein98]. The key points of this method are redundancy limiting and diversity maximizing proposals. For topically related documents the degree of redundancy is quite high in the comparison to a single document so appropriate techniques should be employed to increase diversity of partial text units, which will finally constitute the summarization output. Such redundancy often appears in the form of repetition of common content especially in the case of articles discussing one event from different points of time.

Another similar in character to MMR is the clustered-based summarization. MEAD system [radev00\_2] generates a centroid cluster for document collection composed of central terms for all articles. In next step system identifies summary sentences by comparing them with a centroid cluster.

#### **4.3.2 Natural Language Processing Approaches**

In opposition to above mentioned methods, which follow the text extraction path, there was also a natural language generation system proposed by Barzilay et. Al [barzilay99]. Their technique is called “Theme Intersection” and relies on semantic network usage for paragraph alignment for different documents devoted to common event. The semantic network is employed to identify phrases, which have the same semantic meaning. From phrase clusters new sentences are generated by natural language generation module and ordered in a chronological order.

Mani and Bloerdon [mani99\_3] focus on providing user-defined summaries by extracting common inter-document cohesion relationships between terms. Terms are understood here as words, phrases and proper nouns. The authors use spreading activation algorithm in order to detect which information from each document is relevant to user’s query. Query-related terms in documents can be common for many documents but can be also unique for some of them. These terms are compared across all documents and common or unique sentences are extracted for summary building.

#### **4.3.3 Graph and Links Analysis Methodologies**

Several algorithms for detecting documents’ relationships were recently proposed. Usually those approaches were based on text extraction and fusion techniques and only few proposals used natural language generation methods. Common approach to multi-document summarization exploits connectivity model of documents. Basically it means that the more strongly connected paragraph or any text unit to other paragraphs (units) the more important it is from the point of view of summarization problem. To measure the similarity between parts of documents several features are

usually selected like terms, sentences, n-grams and etc. Salton et al. [salton97] compares document text units basing on the term similarity measure. The paragraphs, which are similar to many other paragraphs from different documents and which also create densely connected points in the connectivity graph (text relationship map) are regarded as salient ones. Consequently they form resulting summary.

## **4.4 Working Systems**

In the following subsections, we introduce the three most popular working systems of news extracting/summarizing : Google News, Newsblaster and NewsInEssence.

### **4.4.1 Google News**

One of the famous automatic news extraction systems is a service provided by Google News. Google News retrieves and groups the news articles from approximately 4,500 news sources worldwide and presents the latest important news on its main page. The source and compiled time of each news (usually the most recent articles appeared few minutes or hours ago) are displayed, followed by its description, which is usually the first paragraph of the news article. For each news topic, Google News may propose a thread or cluster of tens or hundreds of related articles, using their undisclosed grouping methodologies. Users are allowed to do “sort by date” on the articles in a group of a given topic, in order to trace the history of the developing issue. This will arrange the stories in chronological order, with the most recent report placed first. Therefore, users can get to the most recent news and the related articles easily. However, for a user who doesn't follow the Google News daily, he may not know what were the main topics in the past week, because the topic may not appear as the "cover news" again.

#### **4.4.2 NewsBlaster**

Another similar application is Columbia NewsBlaster [newsblaster], an automatic system for event tracking and summarization developed by the Columbia NLP (natural language processing) Group. In addition to clustering the news articles, NewsBlaster implements multi-document summarization methodology [barzilay02] to generate a topic summary for each cluster. They propose that both the chosen clustering algorithm and linguistic text features for better news documents clustering [vasileios00], and evaluated the performance with the DARPA's Topic Detection and Tracking training corpus (TDT2) of 22410 articles. Basically, both Google News and NewsBlaster group daily news articles into existing threads and present the “hot” topics in their main page. It is observed that summary of “hot” topic generated by NewsBlaster is a combination of sentences from multi documents [barzilay01], and has a link to the original document at the end of each sentence. However, NewsBlaster as well as Google do not identify the location of an event and not make use of time information, like what [yang98, allan01] do. [allan01] is a sentence-based system. This system extracts and ranks a single sentence from each event as the temporal summaries of news.

#### **4.4.3 NewsInEssence**

Another similar system for finding and summarizing clusters of related news articles from multiple sources on the Web is NewsInEssence [essence, radev98, radev00], developed by the Computational Linguistics And Information Retrieval (CLAIR) group at the University of Michigan. NewsInEssence's search agent, called NewsTroll, searches for stories related to the same event. The agent then enters keywords into search engines of news sites and produces summaries of a subset of stories that it finds. One of the strong points of this system compared to the two above is its ability of producing multi-length summaries.

## **4.5 Conclusion**

We have seen that significant initiatives have been taken to realize the automation of online journalism using techniques like AI and clustering. One of the main challenges these system is to scale up with hundreds of news sources and thousands of users. It was claimed that users seem satisfied with output accuracy and the convenience for news searching, but these artificial intelligence summarization systems NewsInEssence and Newsblaster are far from perfect. Summaries aren't always as coherent as those written by human editors. Thus, these systems are not to replace human editors. Rather, they provide a complementary tool to help humans cope with the exploding quantity of information on the Web in a timely fashion. Even with errors, they are useful and helpful in this way.

## Chapter 5 “Flow” Type Information Topic Detection and Summarization

### 5.1 Introduction

Topic detection and tracking has become increasingly important after the web information proliferation provides huge dynamically changing textual data online. This task is responsible to discover the evolving features and concepts of interesting topic from the mass of textual data by investigating its content, structure and distribution over time. Basically it takes as input a collection of temporal textual data and recognizes the topic trend in time series. It is regarded by many researchers as a branch research of Web Mining that involves the time features analysis. Recent research has showed much advancement in studying the information that contains time components, and innovating the state-of-art automatic topic trend detection and tracking task. Examples of them who use graphical interpretation to render topic trend detection and tracking are TimeMines, ThemeRiver and PatentMiner presented in Chapter 3. Other researches [pottenger01, allan98, yang98, yang99]] on this topic use a combination of the techniques like linguistic and statistical features, learning, clustering and etc. as their main approaches. However, most of these systems were developed to work on structural data in tagged corpus or databases. These structural data tend to reside in fields with pre-defined semantics and hence easier to handle in many ways, compared to the free text like raw html pages and real-time news articles. More importantly for us, they merely display some features (words, n-grams, document frequency) associated with each detected topic, which show little details about the story lines of each topic. Therefore, a summary of each topic will be more useful for the users to understand the flow of each topic. Lastly, most of them apply clustering techniques to aggregate topic documents and use the cluster size to measure the significance of a topic, resulted in need of some

computational complexities. Therefore, some effort should be made to innovate a more efficient algorithm to do the topic detection and features extraction.

Therefore, our goal is to address the measures for detecting and summarizing the important topics in news archive, given a range of news channels. Our topic detection algorithm TF\*PDF [kbkhoo02] is innovated to take on this objective. It works on the idea that the topics being discussed in several channels concurrently are likely to be “hot” and important. Thus, this algorithm is designed in a way to give significant weight to the terms that explain the important topics in many documents in many news channels concurrently. Later, by exploiting the weight variances of these topic terms, we are getting to know the topic time frame by measuring the information “surprise” in the term weight. The terms of a hit topic should present an acceleration value during the rise of the topic and a deceleration value before the topic fades. Later, we will do sentences clustering on the important sentences appearing in the topic time frame, based on their sentence vector by using the extracted topic term as unit vector. The topic cluster will be the prime cluster generated. In this way, we can generate a better-coverage summary of the topic emerged in a certain time frame. Another way in reverse where we want to summarize all the main topics existing in a certain period (for example the pass week or month), similarly we will use our TF\*PDF algorithm to extract the topic terms for the week and then do clustering on the important sentences in the week. In this scenario, we will produce a range of sentences clusters each devoted to a topic. Sentences in each cluster will be arranged chronologically to form a topic summary respectively. As a result, we could present automatic weekly news topic journalism to the users.

Works on this topic detection and summarization or automatic journalism is of important seeing that news archives of various newswire sources are overwhelming. While providing some important knowledge, these news archives may additionally contain many uninteresting or trivial news. The influencing information is naturally desirable but digesting all the news archive can be a time and

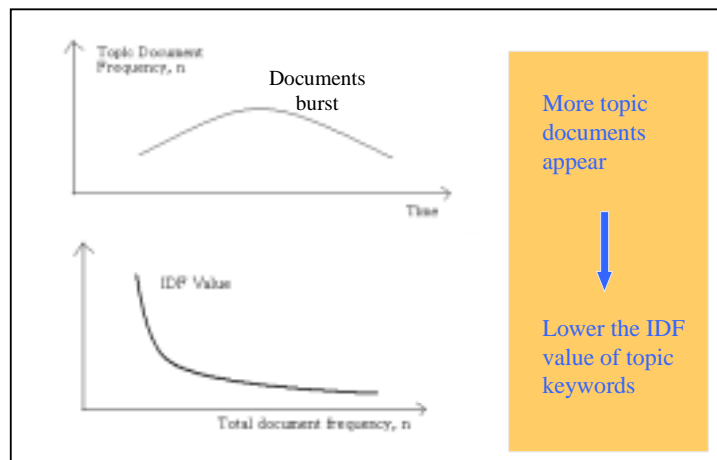
efforts consuming task. Therefore, discovering all the main topics appearing in the archive manually sounds impossible. As a result, it would be helpful if there is a system, which is able to respond correctly to the generic queries such as “What is new?” or “What is important?” Unfortunately, traditional keyword-based retrieval systems work well only for searches using queries with precisely stated goal. Therefore, it is necessary for a user to know beforehand the kind of information or facts he is searching for. However, we are at a higher level of abstraction and creating the precise goals without any knowledge of past weeks' events is rather unrealistic. Hence, what would be desirable is an intelligent system that automatically summarizes a report of the main topics embedded in the archive of newswire sources on the Web.

Research on TF\*PDF (Term Frequency \* Proportional Document Frequency) (Equation 1) [kbkhoo01] algorithm has been reported in some academic publications. TF\*PDF algorithm was implemented in the ETTS system [kbkhoo01, kbkhoo01\_2, kbkhoo01\_3] which is an application designed for tracking the emerging topic in a particular information area of interest on the Web. The information area is represented by a set of web domains that contain the so-called “stock” type information, which is rather static. Whereas, this chapter presents the implementation of TF\*PDF algorithm on a number of news sources on the Web, which are regarded as “flow” (regularly changing) type of information.

## **5.2 Fundamental Study on Topic Detection in News Archive**

Among the conventional information retrieval systems [hayes97, masland92, yang99, dhara00, lafferty99] based on news archive, the famous TDT [yang99] has been selected for detailed study and discussion. We would like to focus on some significant characteristics of conventional IR system that often contribute to insufficiency in fulfilling our objective of summarizing the weekly report of main topics from news archive.

The goal of the pilot research in Topic Detection and Tracking (TDT) [allan98] was a detection of new events and tracking within streams of broadcasted news stories. The involved researchers found 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” [allan98]. Its successor, TDT [yang99] added the online news events detecting and tracking functionality as one of its main objectives. Taking advantage of the fact that the on-going event would have its keywords appearing 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 [salton89] algorithm has been adapted by these systems to find the unique terms in a document, and thus the uniqueness of the document. This algorithm tends to assign the most significant weight to a term when it appears in only one document or in the first document of the event. Because of the large retrospective [yang98] corpus used (six month collection of CNN news corpus), the event keyword appearing in its first document or an early stage of the event document burst, will be weighted significantly and thus can trigger the onset of event detection. However the more consequent event documents containing the event keywords appear, the less likely the document will be judged as an event document. This is because TF\*IDF algorithm always try to assign significant weight to the terms that appear in few documents. Therefore, more event documents appear (document burst), lower will be the value of IDF (see Figure 18 below for reference), and thus lower the weight of the event terms. Thus, we have the risk of losing the detection of the event although it is important and widely discussed in many documents in various newswire.



**Figure 18 : Risk of TF\*IDF – lost of tracks in the midst of hot topic**

In TDT [yang98], large retrospective corpus has been used to calculate the incremental IDF values. As stated in [yang99], IDF algorithm works effectively for document retrieval after a sufficient number of documents have been processed. However, as the retrospective corpus plays an important role in the calculation of term weight, it may influence the results negatively if it is not properly shaped to suit its objectives.

TDT doesn't take advantage of the assumption that during the onset of an event, the event would be reported heavily in majority of the news channels concurrently. Many important terms appear in multiple documents from many different channels when discussing the hot event. By making use of this characteristic, we can create an algorithm to recognize these terms and thus the hot topics accurately, even without the help of retrospective corpus. However, TDT groups the documents 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 the retrospective corpus TDT1, which contains 15863 chronologically ordered and manually segmented news stories from 1 July 1994 to 30 June

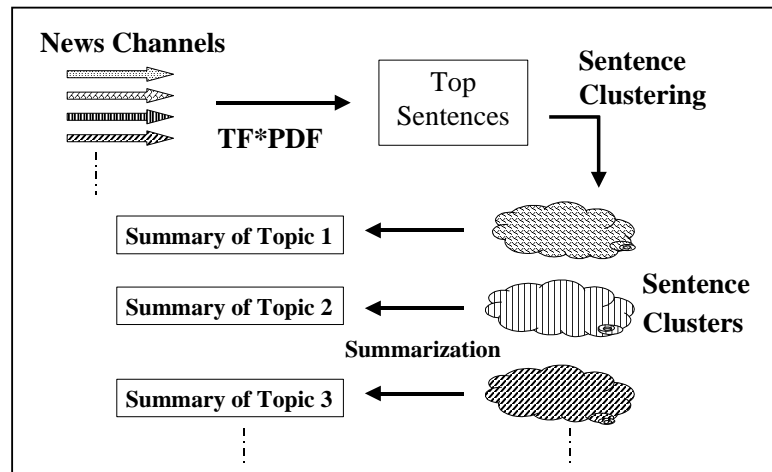
1995. The event histograms [yang99] show that TDT is efficient in event detection in retrospective corpus. However, it is not necessarily that it will have the same performance or will be suitable for online detection of the onset of new events from separate sources in real time, due to the difference in the calculation of IDF. In 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 applied. In general, the event histogram has its event burst in decremental exponential, with significant spikes occurring at the beginning edge. This fact supports the arguments that TDT is efficient in recognizing the start of a new event, but possesses the risk of losing the detection after many documents regarding the event have emerged. However, the burst of event may not fail to trigger the onset of a new event, depending on the width of the time window. This means that it would take at least four weeks for a four weeks time window to trigger the onset of a new event.

In short, the technology adapted in conventional IR and event detection systems is insufficient to construct a system summarizing weekly news report of main topics. Therefore, we were motivated to introduce our novel approach.

### **5.3 Approaches**

Although our research goal may be quite similar to some related works such as TDT, we are addressing the problem in a totally different approach. TF\*PDF algorithm is used to detect the terms that explain the main topics in the news archive and assign them proper weights. Then, the sentences with high weight score will be clustered to their topic groups respectively using their sentence vectors. Each sentence has an associated sentence vector consisting of unit vectors which are the top TF\*PDF terms. Using these sentence vectors, we could classify the sentences to the respective topic by examining their component unit vectors. Each sentences' cluster will be representing a different

topic. The sentences in each topic cluster are then arranged chronologically in order to form a summary of given topic. The flow of information in the system is illustrated in Figure 19.



**Figure 19 : System Information Flow**

### 5.3.1 Overview

Our system works on the basic concept that whenever there is a hot topic appearing, it will be discussed frequently in many news documents from majority of newswire sources. Thus, instead of grouping the information from all sources into a “large” mixed corpus and calculating its term weight with a standard TF\*IDF algorithm like in TDT, we rather give equal importance to news coming from each newswire source and channel it to our system in a parallel way. The terms that explain the hot topics appear frequently in many documents in each channel and should be weighted significantly. Whenever majority of the channels contain common terms with high term weights concurrently, these terms may explain the main topics discussed broadly.

Besides, it should be noted that, terms are supposed to be content words. Therefore, stop words like prepositions (i.e. in, from, to, out) or conjunctions (i.e. and, but, or) are eliminated via a stop word list.

### 5.3.2 Implementation of TF\*PDF Algorithm

In order to fulfill our objective of recognizing the terms that explain the recent hot topics, TF\*PDF is applied to count the significance (weights) of these terms. In contrast to the conventional term weight counting algorithm like TF\*IDF, the weight of a term in a channel is linearly proportional to the term's within-channel frequency, and exponentially proportional to the ratio of documents containing this term in the channel. Then, the total weight of the term will be the summation of the term's weight in each channel as follows:

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

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

$W_j$  = 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 ;  $D$  = total number of channel

There are three major components in TF\*PDF algorithm. The first component that contributes significantly to the total weight of a term is the “summation” of the term weight gained from each

channel, provided that the term seems to explain the hot topic discussed generally in majority of the channels. In other words, the terms that would explain the main topic will be highly weighted. Also, the larger is the number of channels, the more accurate will be our algorithm in recognizing the terms that explain the emerging important topic.

The second and third components are combined to give the weight of a term in a channel. The second part is the normalized frequency of a term in a channel  $|F_{jc}|$  as showed in Equation 2. The term frequency needs to be normalized because a given channel can have a different size of archive than other channels. In such a case the term taken from a channel with more documents would have 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 should be performed.

The third component 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 that contain 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 only a few. Hence, the term that occurs more frequently in many documents in a channel would be considered as the term that seems to explain the main topic in a given channel. The PDF value of a term in a channel, has been experimentally proved to work well in such a way that it should grow exponentially, instead of linearly, with respect to the number of documents containing it. In this way we can give a more significant weight to the term that occurs in many documents compared to the one occurring in just a few. Mathematically speaking, the larger the number of documents containing a term in a channel, the higher will be the grow rate of the PDF of the term in the channel. PDF has a value ranges from 1 ( $\exp 0$ ) to 2.718 ( $\exp 1$ ) exponentially (base e).

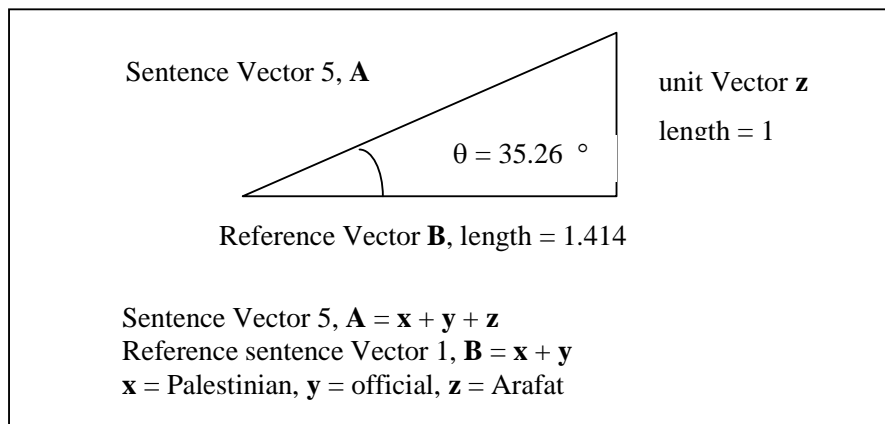
The total weight of a term ( $W_j$ ) is equal to the sum of the weight of the term in each channel respectively. Readers may ask why  $W_j$  is calculated in this way instead of treating all the documents from all the channels as one group and obtaining the value of the  $W_j$  by multiplying the overall term frequency and PDF. The reason is, if there is any channel with a large number of documents containing certain terms of high frequency, then the results would be deviated from detecting terms that explain the broadly recognized hot topics in majority of the news channels.

### **5.3.3 Sentence Vector Clustering**

As depicted in Figure 19, after TF\*PDF term weight calculating, sentence clustering will be carried out as the next step towards summarization of the main topics. Each resulted cluster of sentences will be describing a specific topic. Every sentence vector associated with a sentence might consist of a different number and combination of unit vectors. The top 30 TF\*PDF terms of highest weight in the sentence would be selected as the unit vectors.

#### **5.3.3.1 Minimum Cosine Angle**

As a basic rule, when a given sentence vector has an acute angle not larger than 35.26 degree of unit vectors from another sentence vector (reference vector), the two sentences will be classified into the same cluster. For example, the sentence vector No. 5 (with unit vectors “Palestinian”, “official” and “Arafat”) in Table 10 has an acute angle of 35.26 degree (Figure 2) with the reference vector containing the terms “Palestinian” and “official” from sentence No. 1. As a result, sentence No. 5 would be clustered 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.



**Figure 20 : Cosine Angle for Sentence Clustering**

#### 5.3.3.2 Precision and Recall

The cosine angle of 35.26 degree is the optimal derived after empirical testing. This is the angle to make possible the clustering of a sentence composed of 3 unit vectors together with another sentence containing at least 2 same unit vectors. 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 wrongly to a certain topic
3. **FS** (fail sentence): sentence belongs to an existing topic cluster but failed to be clustered in
4. **NC** (not clustered sentence): sentence not belonging to any existing topic cluster

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

After the process of sentence clustering, all the sentences in each cluster will be arranged chronologically to form a topic summary. Lastly, since the resulting clusters may have different number of sentences or repeating in content, we may need to reduce the length of the summary text by using some dedicated summarization techniques, should it be compression, compaction or condensation [mani99].

## 5.4 First Sample – Topic Selection and Topic Terms Ranking by TF\*PDF Algorithm (Compare with TF)

In this experiment sample, we introduce a topic selection algorithm and show the effectiveness of using TF\*PDF for topic detection. By surveying the experimental results, we could conclude a few advantages of using TF\*PDF towards TF.

### 5.4.1 Corpus (2003 July 20 ~ December 20)

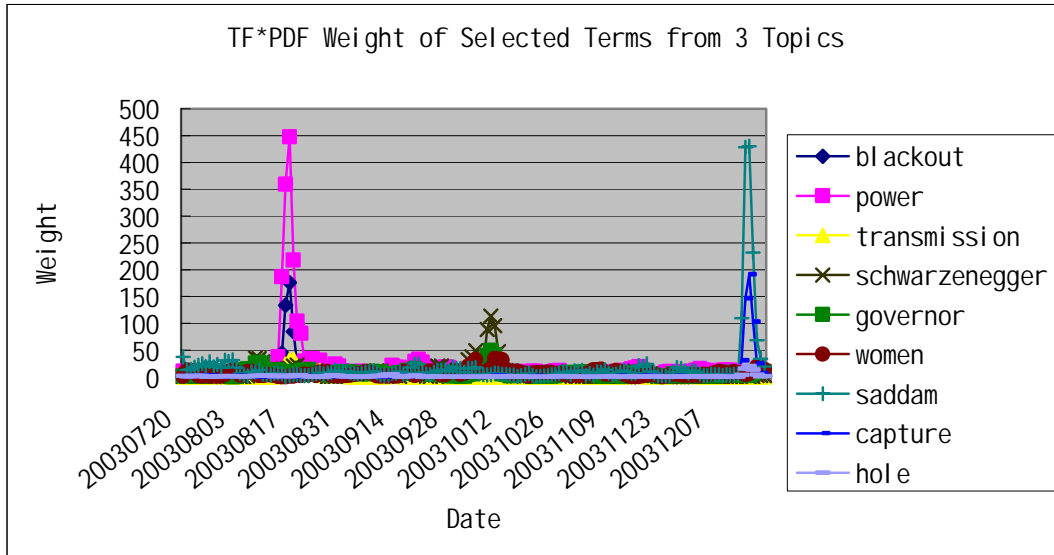
This experiment uses the four news channels: Associated Press (AP), The New York Times (NYT), Reuters and USATODAY as information sources. The news articles dated from July 20, 2003 to December 20, 2003 were collected and used as experiment corpus. A topic selection algorithm involving term's co-occurrence and weight distribution is introduced.

### 5.4.2 Topic Selection Algorithm

Figure 21 shows the TF\*PDF weight distribution of some topic terms from July 20 to Dec 20. This figure illustrates three prominent topics and the related terms happened in the experiment period. These three topics have a high TS (topic selection) value, which is calculated using the Topic Selection Algorithm shown below:

$$\text{Topic Selection, TS} = \frac{C}{C + (1 - C)W}$$

where,  $C$  = co-efficient ;  $C$  = Co-occurrence of topic terms ; Topic Weight:  $W$  = The ratio of the TF\*PDF weight of the co-occurrences terms inside a time window to the outside of the time window



**Figure 21 : TF\*PDF Weight of Selected Terms from July 20 to Dec 20, 2003**

In this experiment, we would like to show the contribution of TF\*PDF algorithm to the topic weight,  $W$ . Thus,  $\alpha$  is set to zero. A 5-days time window is used for the purpose of calculating  $W$ .  $W$  is the weight ratio of in-window weight to the out-window weight of the co-occurrence terms. As a result, we will find many groups of topic terms having different TS (Topic Selection) values at a different time. The first topic terms group with high TS value happens in the time window from August 15 to August 19. Table 2 below shows the TF\*PDF weight of 5 topic terms (“blackout”, “power”, “transmission”, “north” and “grid”) from the first topic. Figure 22 shows the weight distribution of these 5 terms, which is very similar. This first topic is concerning the massive electricity blackout in the northern part of America, where later analysts found out that it was caused by the aged electric grid and transmission problem.

**Table 2 : TF\*PDF Weight of 5 Topic Terms from the First Topic**

	blackout	Power	transmission	North	Grid

20030812	0	13.13	0	5.56	0.26
<i>20030813</i>	<i>0</i>	<i>14.43</i>	<i>0</i>	<i>6.46</i>	<i>0.26</i>
<i>20030814</i>	<i>3.84</i>	<i>38.06</i>	<i>1.69</i>	<i>7.14</i>	<i>1.76</i>
<b>20030815</b>	<b>44.05</b>	<b>186.48</b>	<b>10.51</b>	<b>22.52</b>	<b>19.25</b>
<b>20030816</b>	<b>133.61</b>	<b>359.32</b>	<b>23.01</b>	<b>30.65</b>	<b>32.5</b>
<b>20030817</b>	<b>175.78</b>	<b>447.6</b>	<b>31.58</b>	<b>35.39</b>	<b>43.28</b>
<b>20030818</b>	<b>84.13</b>	<b>217.86</b>	<b>32.78</b>	<b>50.02</b>	<b>29.59</b>
<b>20030819</b>	<b>27.1</b>	<b>103.84</b>	<b>18.15</b>	<b>46.58</b>	<b>14.12</b>
<i>20030820</i>	<i>19.4</i>	<i>81.34</i>	<i>13.11</i>	<i>43.55</i>	<i>11.45</i>
<i>20030821</i>	<i>6.02</i>	<i>28.88</i>	<i>6.17</i>	<i>15.75</i>	<i>7.74</i>
<i>20030822</i>	<i>3.16</i>	<i>21.83</i>	<i>4.41</i>	<i>9.43</i>	<i>8.02</i>
20030823	7.74	34.33	14.24	16.96	13.95

Notes: Bold in dark background = in-window weight; italic = out-window weight. Result:  $W = 33.03$

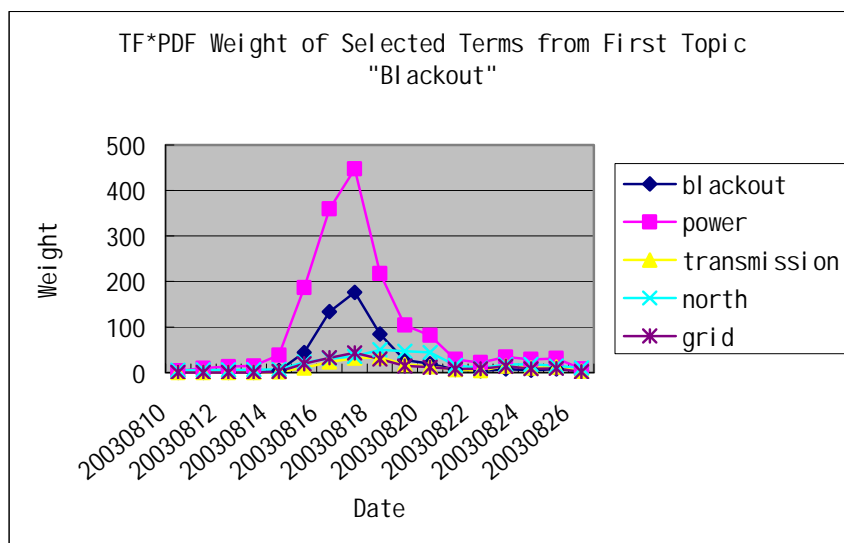


Figure 22 : TF\*PDF Weight of 5 Topic Terms from the First Topic

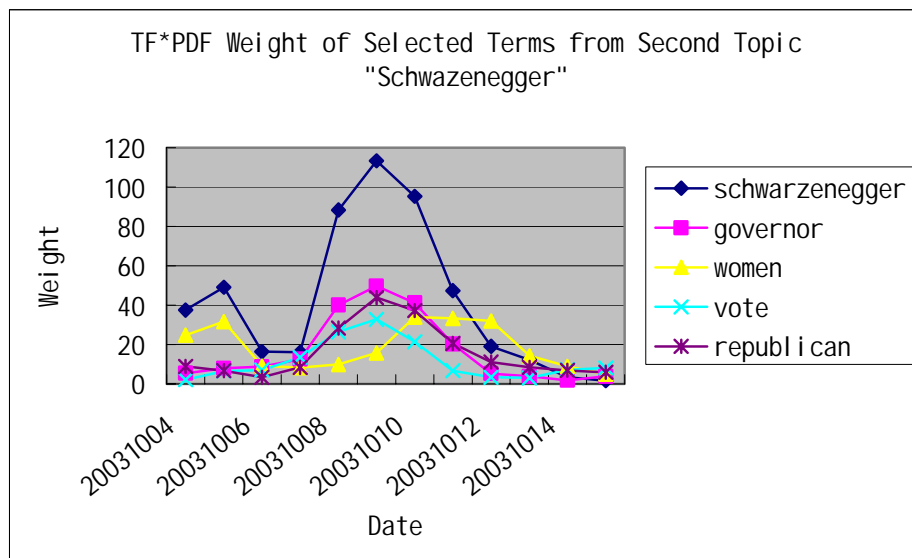
The second topic terms group with high TS value happens in the time window from October 8 to October 12. Table 3 below shows the TF\*PDF weight of 5 topic terms (“schwarzenegger”, “governor”, “women”, “vote”, “republican”) from the second topic. Figure 23 shows the weight distribution of these 5 terms, which is very similar. This second topic is related to the voting recall in

California and the republican representative Schwarzenegger, who was accused of having many women scandals, won the election and became the governor of California state.

**Table 3 : TF\*PDF Weight of 5 Topic Terms from the Second Topic**

	<b>schwarzenegger</b>	<b>Governor</b>	<b>women</b>	<b>vote</b>	<b>republican</b>
20031005	49.09	7.84	31.73	6.52	6.72
20031006	16.29	8.64	9.09	6.96	3.3
20031007	15.92	12.53	8.21	13.47	8.32
<b>20031008</b>	<b>88.28</b>	<b>40.12</b>	<b>9.85</b>	<b>26.45</b>	<b>28.37</b>
<b>20031009</b>	<b>113.37</b>	<b>49.69</b>	<b>15.72</b>	<b>32.83</b>	<b>43.79</b>
<b>20031010</b>	<b>95.31</b>	<b>41.16</b>	<b>33.89</b>	<b>21.35</b>	<b>37.17</b>
<b>20031011</b>	<b>47.38</b>	<b>20.32</b>	<b>33.32</b>	<b>6.6</b>	<b>20.55</b>
<b>20031012</b>	<b>19.02</b>	<b>5.3</b>	<b>31.91</b>	<b>3.27</b>	<b>11.05</b>
20031013	12.14	3.83	14.3	3.08	8.4
20031014	3.25	1.74	8.88	7.16	6.77
20031015	1.55	3.84	4.96	8.01	5.78
20031016	1.52	4.86	10.72	13.86	7.95

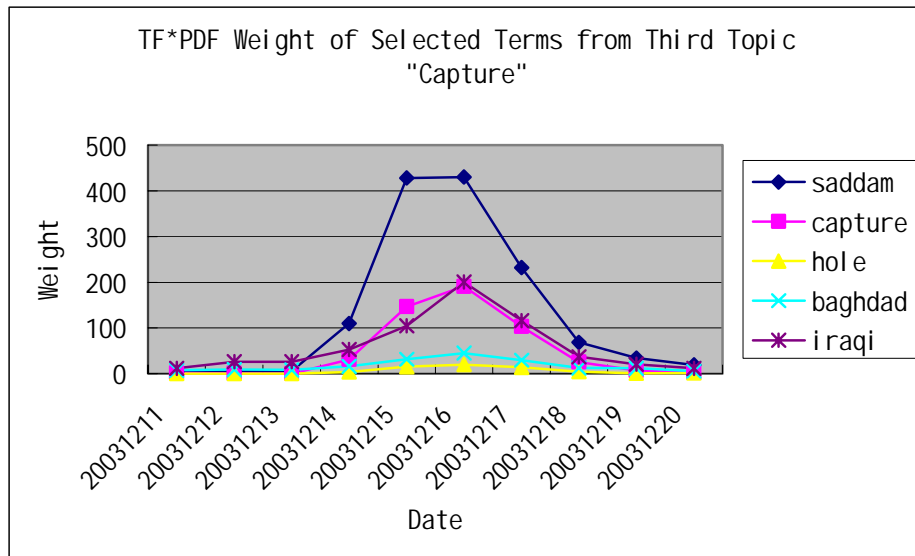
Result: **W** = 21.91



**Figure 23 : TF\*PDF Weight of 5 Topic Terms from the Second Topic**

**Table 4 : TF\*PDF Weight of 5 Topic Terms from the Third Topic**

	saddam	Capture	Hole	baghdad	Iraqi
20031210	4.25	1.44	0	3.65	4.4
20031211	3.54	1.15	0.28	7.64	11.9
20031212	4.59	0.53	0.28	8.76	26.1
20031213	5.13	0.27	0.28	8.18	26.01
<b>20031214</b>	<b>109.54</b>	<b>30.79</b>	<b>3.5</b>	<b>15.83</b>	<b>52.58</b>
<b>20031215</b>	<b>428.34</b>	<b>146.34</b>	<b>14.98</b>	<b>31.55</b>	<b>104.47</b>
<b>20031216</b>	<b>430.48</b>	<b>191.28</b>	<b>19.82</b>	<b>44.85</b>	<b>200.47</b>
<b>20031217</b>	<b>232.16</b>	<b>102.92</b>	<b>13.46</b>	<b>29.12</b>	<b>115.85</b>
<b>20031218</b>	<b>68.52</b>	<b>24.97</b>	<b>3.96</b>	<b>12.47</b>	<b>36.64</b>
20031219	33.77	7.06	0.81	11.8	20.17
20031220	19.03	2.32	1.04	5.67	11.33

Result:  $W = 92.25$ **Figure 24 : TF\*PDF Weight of 5 Topic Terms from the Third Topic**

The third topic terms group with high TS value happens in the time window from Dec 14 to Dec 18. Table 4 shows the TF\*PDF weight of 5 topic terms (“saddam”, “capture”, “hole”, “baghdad”, “Iraqi”) from the third topic. Figure 24 shows the weight distribution of these 5 terms, which is very similar. This third topic is about the capture of the Saddam in a fox hole in the north of Baghdad. A lot of Iraqi was happy with the capture and celebrated on the street.

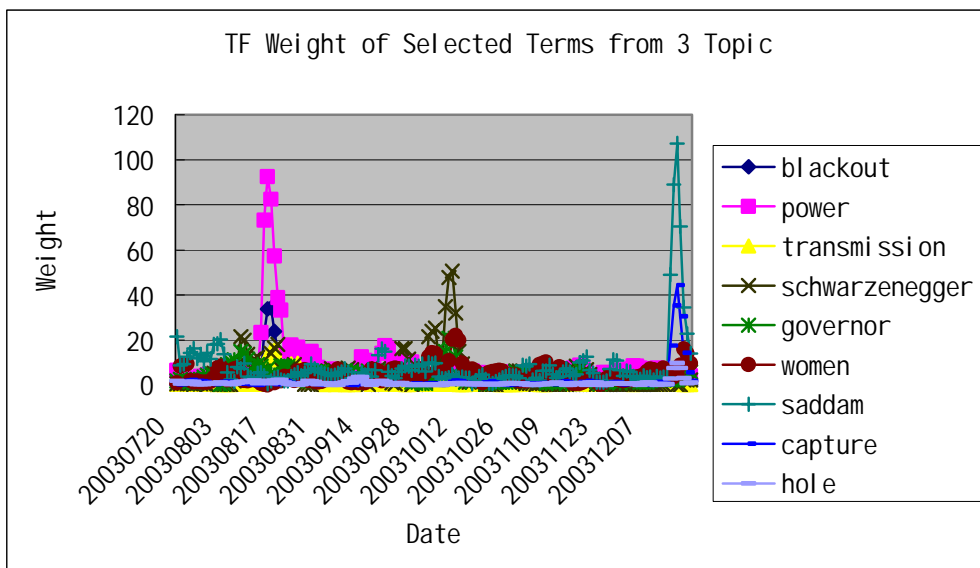
**Table 5 : The value of W (Topic Weight) obtained using TF\*PDF and TF**

	<b>Terms</b>	<b>W (Using TF*PDF)</b>	<b>W (Using TF)</b>
First Topic	Blackout	14.33	5.82
	Power	7.13	3.93
	Transmission	4.57	3.54
	North	2.25	1.57
	Grid	4.75	3.31
	<b>Total</b>	<b>33.03</b>	<b>18.17</b>
Second Topic	Schwarzenegger	7.39	5.29
	Governor	5.12	3.79
	Women	2.74	2.56
	Vote	2.34	1.51
	Republican	4.32	3.15
	<b>Total</b>	<b>21.91</b>	<b>16.3</b>
Third Topic	Saddam	19.21	7.49
	Capture	43.8	15.75
	Hole	20.71	12.2
	Baghdad	3.18	1.79
	Iraqi	5.34	2.31
	<b>Total</b>	<b>92.25</b>	<b>39.54</b>

### 5.4.2.1 Conclusion

In this subsection we have proposed a Topic Selection Algorithm,  $TS = C + (1 - \alpha)W$ . Table 5 above presents the  $W$  (Topic Weight) values of three topic terms groups with high  $TS$  (Topic Selection) values when using  $TF*PDF$  algorithm.  $W$  is the weight ratio of in-window (5 days window wide) weight to the out-window (5 days before and after the window) weight of all co-occurrence terms. The same value sets obtained when using  $TF$  (Term Frequency) is displayed at the side for comparing (detail number is tabled in next subsection). As a result, we found out that the topic weight ratio  $W$  calculated using  $TF*PDF$  will have a higher value and thus more likely to trigger the topic detection. A multiplication of co-efficient would not help to improve this topic weight ratio of  $TF$  because its off-peak values (or noise) would also multiply.

### 5.4.3 How about using $TF$ (Term Frequency)?



**Figure 25 : TF Weight of Selected Terms from July 20 to Dec 20, 2003**

$TF$  weights of selected topic terms from July 20 to Dec 20 are displayed in Figure 25. From this figure, we can see that the  $TF$  value of topic terms are relatively lower than the  $TF*PDF$  values.

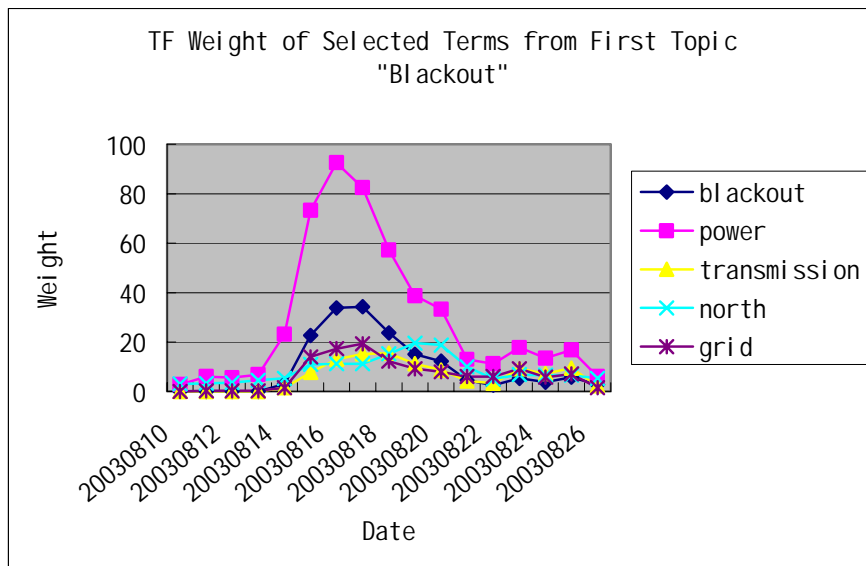
Without the effect of PDF, we found that the topic terms' weight within the topic-occurring time frame is not "outstanding". On the other hand, some low value topic terms have their values close to the noise level. The smaller values of W (Topic Weight Ratio) presented in Table 5 explains that the difference of peak and off-peak values of TF weight is smaller. The multiplication of a co-efficient can make the peak value of TF equal to peak value of TF\*PDF, but at the same time the off-peak (and noise) value will grow at the same multiple as well. Thus, the multiplication of a co-efficient doesn't help to improve the Topic Weight Ratio, W.

Table 6, 7, 8 and Figure 26, 27 28 below table and figure the TF weights of 5 topic terms from the first, second and third topics respectively. The TF weights and Topic Weight Ratio, W are relatively lower than those of TF\*PDF. A low topic weight value may fail to trigger the topic selection algorithm for alerting a topic detection.

**Table 6 : TF Weight of 5 Topic Terms from the First Topic**

	<b>blackout</b>	<b>power</b>	<b>transmission</b>	<b>North</b>	<b>Grid</b>
20030812	0	5.5	0	3.5	0.25
20030813	0	6.75	0	4.5	0.25
20030814	3	23.25	1.5	5.25	1.5
<b>20030815</b>	<b>22.75</b>	<b>73.25</b>	<b>7.75</b>	<b>11</b>	<b>14</b>
<b>20030816</b>	<b>33.75</b>	<b>92.5</b>	<b>13</b>	<b>11.25</b>	<b>17.25</b>
<b>20030817</b>	<b>34.25</b>	<b>82.5</b>	<b>15.25</b>	<b>11.25</b>	<b>19.25</b>
<b>20030818</b>	<b>23.75</b>	<b>57.25</b>	<b>15.5</b>	<b>15.25</b>	<b>12.25</b>
<b>20030819</b>	<b>15</b>	<b>38.75</b>	<b>11.25</b>	<b>19.5</b>	<b>9.25</b>
20030820	12.25	33.25	8.75	18.75	8
20030821	4.5	13	4.25	10	6
20030822	2.5	11.25	3.25	5	6
20030823	5.25	17.75	9.25	6.75	9.25

Result: W = 18.17

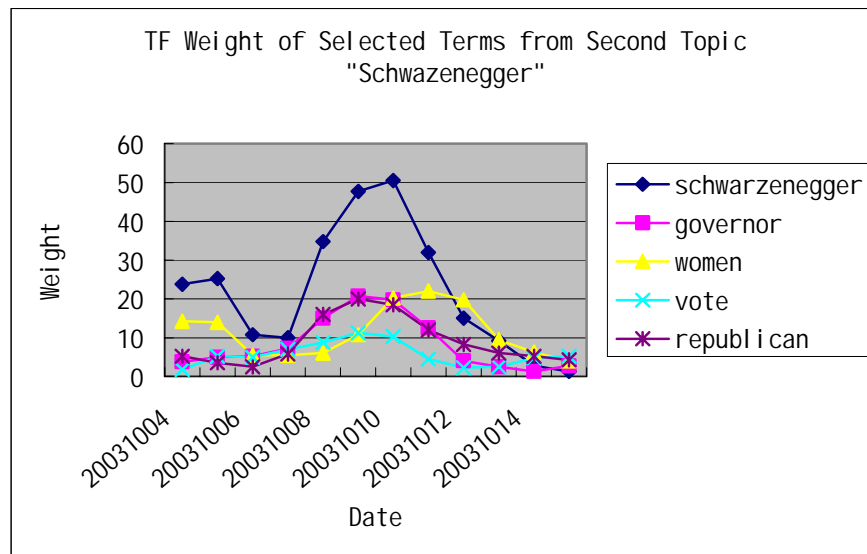


**Figure 26 : TF Weight of 5 Topic Terms from the First Topic**

**Table 7 : TF Weight of 5 Topic Terms from the Second Topic**

	<b>schwarzenegger</b>	<b>governor</b>	<b>women</b>	<b>Vote</b>	<b>republican</b>
20031005	25.25	5	14	5.25	3.5
20031006	10.75	5.25	5.5	5	2.5
20031007	10	7.25	5.5	7	5.75
<b>20031008</b>	<b>34.75</b>	<b>15</b>	<b>6</b>	<b>8.75</b>	<b>16</b>
<b>20031009</b>	<b>47.75</b>	<b>20.75</b>	<b>10.75</b>	<b>11.25</b>	<b>20</b>
<b>20031010</b>	<b>50.5</b>	<b>19.75</b>	<b>20.25</b>	<b>10.25</b>	<b>18.5</b>
<b>20031011</b>	<b>32</b>	<b>12.5</b>	<b>22</b>	<b>4.5</b>	<b>12</b>
<b>20031012</b>	<b>15</b>	<b>4</b>	<b>19.75</b>	<b>2.25</b>	<b>8.25</b>
20031013	9.25	2.5	9.5	2.5	6
20031014	2.75	1.25	6.25	4.75	5.25
20031015	1.25	2.75	4	5.25	4.25
20031016	1.25	3.5	7	8.75	5.5

Result: W = 16.3



**Figure 27 : TF Weight of 5 Topic Terms from the Second Topic**

**Table 8 : TF Weight of 5 Topic Terms from the Third Topic**

	saddam	capture	hole	baghdad	iraqi
20031210	3.75	1.25	0	3	3.5
20031211	3	1	0.25	6.25	9.75
20031212	3.5	0.5	0.25	7	16
20031213	3.5	0.25	0.25	6.5	15
<b>20031214</b>	<b>49</b>	<b>17.5</b>	<b>2.75</b>	<b>9.25</b>	<b>23</b>
<b>20031215</b>	<b>89</b>	<b>35.25</b>	<b>7.5</b>	<b>13.25</b>	<b>30.75</b>
<b>20031216</b>	<b>107.3</b>	<b>44.25</b>	<b>9.75</b>	<b>16.5</b>	<b>40.75</b>
<b>20031217</b>	<b>70.5</b>	<b>30.5</b>	<b>7.5</b>	<b>13</b>	<b>33.5</b>
<b>20031218</b>	<b>34.5</b>	<b>14.25</b>	<b>3</b>	<b>9.25</b>	<b>21.5</b>
20031219	22.75	5.5	0.75	9.5	15.25
20031220	14	1.75	1	5	8.75

Result: W = 79.08

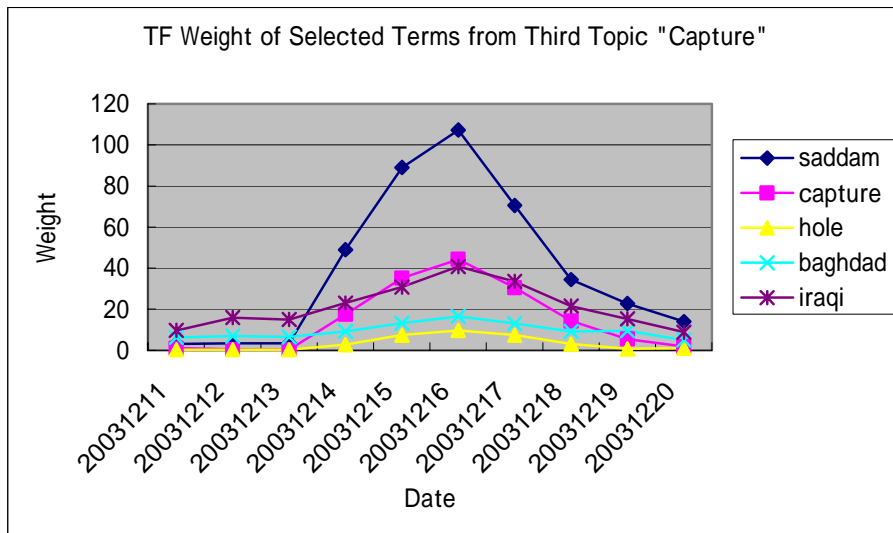


Figure 28 : TF Weight of 5 Topic Terms from the Third Topic

#### 5.4.4 Topic Terms Ranking using TF\*PDF and TF

The TF\*PDF algorithm can push the topic terms up in the list. At the same time strengthening the weight, TF\*PDF can improve topic terms' ranking. On the other hand, topic terms can drop out from the TF list while some non-topic terms would appear high in the list. A multiplication by an equation or square may enhance the terms' weight in the TF list to the same amplitude with TF\*PDF list but it wouldn't change the ranking.

##### 5.4.4.1 Terms Ranking in the First Topic

Table 9 below shows the TF\*PDF and TF term weight at the peak time of the first topic on August 17. The terms with dark-gray background are the topic terms, which were used in the example for calculating W (Topic Weight). The term "north" drops out from the TF list of top 15 terms, where it is ranked 19. At the same time, some non-topic terms which are not related to the first topic, such as "amin", "water", "bush", "schwarzenegger" and "palestinian" have a high ranking in the TF list. Although not a topic term, "amin" gains high TF weight and gone up in the TF list (but not the

TF\*PDF list) because it appears many time in one or few document in the news channel of Reuters. Figure 29 is the example document where the term “amin” appears 18 times. We should not argue that these non-topic terms are not important and should be blocked with a finer stop word list, because the non-first-topic term “schwarzenegger” appears to be the main topic term in the second topic. In fact, stop word list itself is a research work that still needs more efforts to be put into.

**Table 9: Term Ranks at the peak time of First Topic (August 17)**

No	TF*PDF		TF	
	Term	Weight	Term	Weight
1	power	447.6	power	82.5
2	city	314.42	city	44.75
3	blackout	175.78	blackout	34.25
4	electricity	64.61	grid	19.25
5	energy	46.22	electricity	18
6	night	43.61	energy	17.5
7	grid	43.28	amin	17
8	area	35.94	transmission	15.25
9	north	35.39	water	14.25
10	emergency	35.09	bush	14.25
11	detroit	34.71	schwarzenegger	14
12	failure	32.68	detroit	13.8
13	transmission	31.58	cleveland	13.37
14	line	30.87	emergency	13.14
15	electric	28.53	palestinian	13.07
	bush (rank 19, weight 27.19) water (rank 20, weight 26.53) amin (rank 37, weight 19.4) schwarzenegger (rank 40, weight 16.47) palestinian (rank 54, weight 14.57)		north (rank 19, weight 12.25)	

The term such as “transmission” and “grid” are ranked lower in TF\*PDF list compared to the TF list. However, it doesn’t mean that these terms are not significant in TF\*PDF list because there are some other topic terms such as “failure”, “energy” and “electricity” are also important. The term “night” and “city” may tells that people in New York city (or Detroit city) needed to gone through

their night without electricity. Instead of weak point, this is actually the strong point of TF\*PDF ability to push the topic terms high in the list, which is not the case for TF.

**Top Stories - Reuters**

## Former Ugandan Dictator Amin Buried in Saudi Arabia

Sat Aug 16, 4:06 PM ET

*By Paul Busharizi*

KAMPALA (Reuters) - Former Ugandan President Idi Amin, one of Africa's bloodiest despots who was blamed for killing tens of thousands of his people, was buried at a small funeral in Saudi Arabia hours after his death on Saturday.

x

Photo

Amin, dubbed "the butcher" by many Ugandans, was buried in the Red Sea coastal city of Jeddah where he had lived in a villa for much of the time since being ousted from power in 1979. He was in his late 70s.

x

Reuters Photo

The quick funeral was in keeping with Amin's Muslim faith, but the mostly family affair was a far cry from the pomp he demanded during in the 1970s when he ruled Uganda with a whimsical savagery that shocked and revolted the world.

x

Reuters Photo

Uganda remains full of stories of how Amin, who seized power in a military coup in 1971 and ruled for eight years, kept severed heads in a fridge, fed corpses to crocodiles and had one of his wives dismembered. Some say he practiced cannibalism.

**Slideshow: Ex-Ugandan Dictator Idi Amin Dies**

**Figure 29: Example document with the term “amin” appearing many time**

#### 5.4.4.2 Terms Ranking in the Second Topic

In the Table 10 below, the terms with dark-gray background are again the topic terms, which were picked up randomly for showing example in calculating W (Topic Weight) for the second topic. However, one of the terms which is “women” drops out from both the TF\*PDF and TF list, because the weight distribution of this term is a little bit lagging the peak time of the second topic. At the

same time, we have the term “vote” drop out from the TF list. So, TF\*PDF is doing better with only one drop out instead of two for the TF. Also, since the topic was about the vote recall in California, the term “recall” and “california” have been pushed a little higher in the TF\*PDF list comparatively. Nevertheless, for some of the topic terms, it is difficult to argue that which is more important and should be ranked a bit higher and so on. This is rather a subjective answer depends on the feeling of the judge. However, it is undoubted that the term “syria”, which is a non-topic term has gone up into the TF list. Figure 30 is the document (from AP source) in which the term “syria” appear 33 times. This term of “syria” doesn’t gone high in TF\*PDF list because the TF\*PDF algorithm has the ability to filter out this kind of non-topic term appearing very frequently in one or just few document.

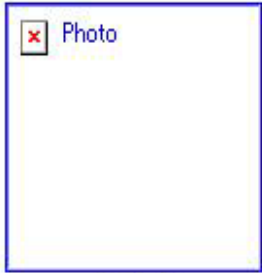
**Table 10: Term Ranks at the peak time of Second Topic (Oct 9)**

No.	TF*PDF		TF	
	Term	Weight	Term	Weight
1	schwarzenegger	113.37	schwarzenegger	47.75
2	california	71.95	bush	26.25
3	bush	68.72	california	26
4	american	60.77	american	21.25
5	recall	50.79	governor	20.75
6	governor	49.69	republican	20
7	republican	43.79	recall	19.5
8	campaign	43.24	campaign	19.25
9	iraq	36	iraq	19.25
10	washington	35.01	security	14.75
11	security	34.66	washington	14.5
12	vote	32.83	syria	12.5
13	national	28.39	office	12.25
14	office	27.09	democratic	11.75
15	house	24.57	national	11.5
	syria (ranked 39, weight 11.11)			


**Syria Criticizes U.S. Vote on Sanctions**  
 Thu Oct 9, 5:59 PM ET

*By DONNA ABU-NASR, Associated Press Writer*

DAMASCUS, Syria - The preliminary U.S. congressional approval of sanctions against Syria sparked fierce condemnation from Damascus Thursday, with one Syrian official calling it a "biased and illogical act" that would damage U.S.-Syria relations and dim chances for peace in the Middle East.

 Photo  
 AP Photo

The vote by the House International Relations Committee on Wednesday was a largely symbolic gesture — but one that could lead to more damaging U.S. measures, Western diplomats said.

 AP Video  
 U.S. Envoy Condemns Syrian President  
 (AP Video)

The bill, which accuses Syria of sponsoring terrorists, seeking weapons of mass destruction and occupying Lebanon with more than 20,000 troops, passed three days after Israeli warplanes struck an alleged Palestinian militant training camp outside Damascus. The attack came a day after an Islamic Jihad bomber killed 19 people in an Israeli restaurant.

**Figure 30: Example document with the term “syria” appearing 33 times**

#### 5.4.4.3 Terms Ranking in the Third Topic

Similarly for the third topic, in the Table 11 below, we found that one of the topic term “hole” drops out from the TF list. This term is ranked at 22 with a weight of 9.75 in the TF list. On the other hand, the non-topic term “dean” came to the TF list. Again, this is because TF doesn’t have the effect of TF\*PDF to push the main topic terms up in the rank. In other words, TF would pull up the non-topic terms that appear very frequently in one or few document.

**Table 11: Term Ranks at the peak time of Third Topic (Dec 16)**

No.	TF*PDF		TF	
	Term	Weight	Term	Weight
1	saddam	430.48	saddam	107.25
2	hussein	332.46	iraq	60.75
3	iraq	228.34	hussein	49
4	iraqi	200.47	capture	44.25
5	capture	191.28	iraqi	40.75
6	bush	115.43	bush	39.75
7	baghdad	44.85	american	33.75
8	world	39.92	world	17
9	washington	34.85	baghdad	16.5
10	general	25.41	washington	13.75
11	court	20.67	court	13
12	official	20.44	dean	13
13	international	19.97	general	12.25
14	hole	19.82	international	11.5
15	dictator	19.66	official	11
			hole (ranked 22, weight 9.75)	

#### 5.4.4.4 Conclusion

It is proved with the facts that TF\*PDF algorithm is working better than TF in ranking the topic terms. We have shown that it can practically push the topic terms up in the rank, at the same time filtering out the non-topic terms that have high frequency in just one or few documents. If we multiply the TF weight with some equation (such as square or exponential), we may make the amplitude of TF weight similar with TF\*PDF weight, but this doesn't change the rank the topic terms. Therefore, some topic terms may drop out from the TF list while the non-topic terms will climb up the TF list to the top. Besides, there should not be an argument that the non-topic terms come to the TF list should be blocked by using a proper stop word list, because we found that the term "schwarzenegger" (the main term of the second topic) was pulled up to the top in the first topic TF list but not the TF\*PDF list. To conclude, the facts being shown is the solid prove telling that

TF\*PDF algorithm is working better than TF to do topic word ranking, by its natural ability to push the topic words up in list while filtering out the non-topic terms.

#### **5.4.5 Evaluation for Terms Extraction by TF\*PDF and TF**

The standard way of doing evaluation is Precision and Recall. For the first topic, there are five non-topics terms in the TF list, which are “amin”, “schwarzenegger”, “palestinian”, “water” and “bush”. Thus, the precision for TF is 10/15 while the recall is 10/REL. It would not be difficult for us to judge if a term is non-topic term. However, it can be difficult for us to tell if the push up terms (“night”, “area”, “failure” and “line”) in the TF\*PDF list are topic terms. It is easier to accept that “night”, “failure” and “line” are topic terms (because of example such as the transmission line problem and power failure, people need to go thorough the night without electricity and officials need to work through the night to restore power). The most controversial term can be the “area”. This is because this term can be explaining other events in other area in the world, for example what is happening in Iraq northern area and so on. In this case, the most appropriate way for us to evaluate whether it is a topic term is to look into the corpus, and check how many percent of the documents containing the term “area” is are topic documents. After doing a checking on all the documents from the first topic time frame, which are containing the term “area”, we actually found out that 15 out of 24 documents from AP are topic documents (Table 12). Similarly, we found that 19 out of 25 documents from NYT containing “area” are topic documents (Table 13); 7 out of 11 from Reuter (Table 14) and 9 out of 14 from USATODAY (Table 15). Finally, Table 16 summarizes the percentage of topic documents in each channel. Overall, 67.7% or more than 2/3 of the documents containing the term “area” are related to the topic “blackout”. It is concluded that high term weight and thus ranking of the term “area” is mainly contributed by the topic means. Therefore, we can calculate in this example which is a big topic that the Precision from TF\*PDF is 15/15 and Recall 15/REL.

**Table 12: Documents containing the term “area” (August 15~19, AP)**

Topic Documents		Non-topic Documents	
File Name	Sentence	File Name	Sentence
Blackout.html	Outages ranged over an area with roughly 50 million people.	iraq_raid.html	Soldiers are still at the scene searching the area
Blackout_autos.html	Cooperation among the Big Three was a key factor to restoring power to the area sooner than expected.	car_nascar_spencer_appeal.html	After the event, Busch said his car ran out of gas near Spencer's hauler in the garage area.
Blackout_north.html	much of the metropolitan area to have power soon.	iraq_fighting.html	U.S. Army patrol passed the area about 30 minutes later
Blackout_costs.html	New York State, excluding the city, lost another \$1 billion and the other affected areas outside New York another \$3 billion.	golf_pga_championship.html	I could see things turning around and I'm starting to feel confident in certain areas of my game
Blackout_delicate_g_rids.html	utilities were loath to build new capacity in Connecticut's Fairfield County, because of high property costs even though the area holds a quarter of the state's residents and accounts for half the state's demand	canada_sars.html	The latest deaths come nearly six weeks after the WHO removed Toronto from its list of SARS-infected areas, saying the city had contained the outbreak.
Blackout_fema_chief.html	The nation's worst blackout occurred two weeks ago after cascading power failures darkened many areas of eight states	israel_military.html	Israel reoccupied most West Bank towns, and has been moving in and out of these areas repeatedly
Blackout_investigation.html	council officials said they were among five reported transmission failures in the area during a period of just over leading up to the blackout peak	afghan_independence_security.html	Omar, has improved coordination among his commanders, dividing Afghanistan into military areas of control
Blackout_midwest.html	a state of emergency declared for five southeast Michigan counties and signed an executive order to expedite nearly a million gallons of gasoline from West Michigan to the Detroit area.	afghan_fighting.html	tribesmen in the area openly say they would protect Taliban
abraham_power.html	"The (FERC) measure ... goes to the question of whether or not we would mandate and force down the throats of regional areas of the country a federal approach to deregulation of the marketplace," Abraham	tropical_storm_erika.html	Brownsville is just north of the Mexican border, farther south than areas hardest-hit by Claudette.
blackout_phones.html	Phone service in areas of the Eastern U.S. affected by the blackout		

ml	was disrupted		
blackout_sports.html	Electric service in the area was recovering more quickly.		
blackout_vignettes.html	a woman from Texas who was visiting businesses in the area and trying to get to New York.		
blackout_what_happened.html	Reports of lightning hitting a facility in the Niagara Falls area have been ruled out, as have reports that a fire at a New York City electric facility may have triggered the power disaster.		
finance_blackout.html	Automatic teller machines in the blackout areas were shut down, however.		
Blackout_new_york.html	Consolidated Edison, which provides power to the area, was not sure what caused the blackout.		

**Table 13: Documents containing the term “area” (August 15~19, source NYT)**

Topic Documents	Sentence
cheersinnewyorkcitybutpartsofmidwestarelagging.html	Service had also resumed for most people in northern Ohio and southern Michigan, two other hard-hit areas
detroitssweatswhileitwaitsforelectricity.html	power had been restored by midafternoon to nearly all the 1.5 million area residents who had lost it
expertsaskingwhyproblemsspreadsofar.html	After the 1965 blackout, the transmission system that carries power from one area to another was modified specifically
powerwasrestoredtonewyorkcity.html	Officials who oversee the power-sharing system ordered rolling blackouts cutting power for several hours in several areas
reversalofpowerflowshugeamount.html	why systems designed to confine problems to relatively small areas suddenly failed
cautioningathoroughinvestigation.html	high demand due to heat, lightning strikes in the Niagara Falls area, or a fire at a power plant
surplusenergycasedtopowerfail.html	power plants in an area are producing less electricity than consumers are demanding, the system falls below 60 cycles per second
expertsretraceastringofmishapsbeforetheblackout.html	Michigan utility began supplying power to the very same area in an effort to meet the demand
forasuddenpowerlessmanhattanasnowdayin august.html	who had been on a train passing through the New York area from Virginia to her home in Newton
fortravelersthewordiscallahead.html	Airports in more than a dozen cities in the blackout area, from Boston to Detroit, were operating

	to Detroit, were operating
ohiolinesfailedbeforeblackout.html	a second 345-kilovolt line in the same area, probably one helping to carry the load from the first failed line
oversightgroupwarnedutilitiesonpowerflows.html	The Midwest Independent System Operator is charged with helping oversee the safe generation and transmission of power in the Ohio area
partofgridsusceptibletoanticipatedflows.html	There are areas in all parts of the country that are constrained
reliabilityofthebulkelectricitysupplyinnorthamerica.html	New York City, southwestern Connecticut and Long Island were areas of concern because of a lack of power plants and choke points in transmission lines.
partsofcountrymayenterweekendwithoutpower.html	continued repairs on the grid that distributes electricity to New York and the eastern United States and tried to restore electric service to affected areas, which included parts of New Jersey, Connecticut
surgeofpowertriggeredcascadeshutdowns.html	blackout may have been triggered somewhere along Lake Erie in Ohio and spread throughout the affected area
partsofcountrymayremainwithoutpowerthroughweekend.html	electricity to the eastern United States back in operation and restore electric service to affected areas, which stretched from the East Coast to Detroit
powerrestoredtoportionsofnewyork.html	By midmorning, several blocks in the Midtown area of Manhattan, around the heavily touristed Times Square district, had power again.
powerfailurerevealsacreakysystemenergyexpertsbelieve.html	physically impossible to transmit that much power into the area along the existing lines
<b>Non-topic Documents</b>	<b>Sentences</b>
bushadministrationplansdefenseofterrorlaw.html	He works well in that area
attacksiniraqmaybesignalsofnewtactics.html	Most of the area will be without water, and now people will start saying the Americans did this
formeriraqiofficialknownaschemicalaliiscaptured.html	Mr. Hussein was still hiding in the "Sunni triangle" area to the north and west of Baghdad
frenchhealthofficialquitsoverheatwavedeaths.html	Raffarin came under fire in the spring because of his efforts to overhaul the laws governing several key areas of French public service
genetherapyusedtotreatpatientswithparkinsons.html	Another potential danger is that the virus could spread to other areas of the brain
israelsoftensstanceonwantedpalestinians.html	The agreement appeared to apply only to areas under Palestinian security control

**Table 14: Documents containing the term “area” (August 15~19, source Reuter)**

Topic Document		Non-topic document	
File Name	Sentence	File Name	Sentence
power_airlines_dc.htm 1	The FAA said traffic headed into and out of the three major New York area airports was halted,	colombia_usa_rumsfeld_dc.html	A U.S.-backed spraying program reduced the area of coca crops
power_dc.html	In New York, Citibank and J.P Morgan Chase and Co said their automatic teller machine networks had been shut down in areas affected by the blackout.	crime_greenriver_dc.html	A human skeleton unearthed in a rural area near Seattle last week was that of a teenage
power_outages_airtraffic_dc.html	Air traffic headed into the three major New York-area airports was halted on Thursday because of a massive power outage in the eastern United States and Canada	crime_shooting_ohio_dc.html	The incident took place in an office area of the Andover Industries plant about 50 miles east of Cleveland
power_regulation_dc.html	blackout in 1977 was limited to a much smaller area, primarily New York City	financial_citibank_scam_dc.html	Citibank is the No. 3 U.S. commercial bank by assets and the No. 2 retail bank in the New York City area
power_gas_dc.html	dozens of cars and motor homes blocking offramps, as people fled the darkened areas		
power_midwest_dc.html	In Akron, power was restored in all but a four-block area by 11 a.m.		
power_grid_dc.html	the system is designed to "shed load," or turn off power supply to some areas to balance the amount of power		

**Table 15: Documents containing the term “area” (August 15~19, source USATODAY)**

Topic Document		Non-topic document	
File Name	Sentence	File Name	Sentence
11598744.html	the waiting area suddenly went black	11597871.html	Fine said the FBI needs to move more aggressively in

			several areas
11598751.html	president of PJM Interconnection, a power authority for seven states, including blackout areas in New Jersey and Pennsylvania	11621381.html	similarities to the sniper attacks that killed 10 people in the Washington, D.C., area last fall
11598856.html	The New York blackout of 1977 was more terrifying, and it lasted up to 25 hours in some areas.	11635714.html	the tests will also be watching for North Korean and Chinese subs because they frequent the areas where the tests will take place
11623033.html	At least some efforts to improve the nation's electrical grid were underway before Thursday's power system crash, though not in the stricken areas	11657096.html	The string of killings bears a resemblance to the sniper attacks that terrorized the Washington, D.C., area last fall
11621335.html	Some power companies, seeking top dollar, send electricity to areas already saturated with power, and that overwhelms the system	11658837.html	it would be moved to a non-public area in the building or taken out
11623851.html	Detroit-area residents began to get power back		
11624694.html	the bigger question is why the system failed to contain the blackout to a limited area, as it is designed to do		
11623634.html	Just over an hour after the first line failed, an area covering 9,600 square miles of the Midwest, Northeast and Canada was without electricity.		
11598748.html	NERC has long worried about the "Erie Loop," the area where the blackout may have begun		

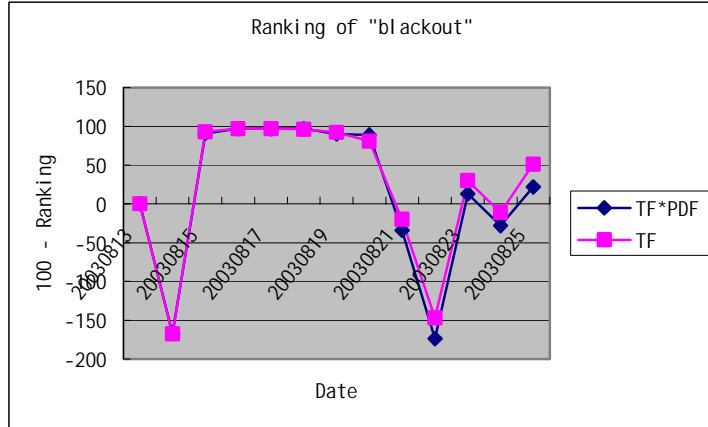
**Table 16: Ratio of topic document containing the term “area” (August 15~19)**

	Number of Topic Document	Number of non-topic document	Total	Topic Document/Total
AP	15	9	24	15/24
NYT	19	6	25	19/25
Reuter	7	4	11	7/11
USATODAY	9	5	14	9/14
Total	50	24	74	50/74 (67.7%)

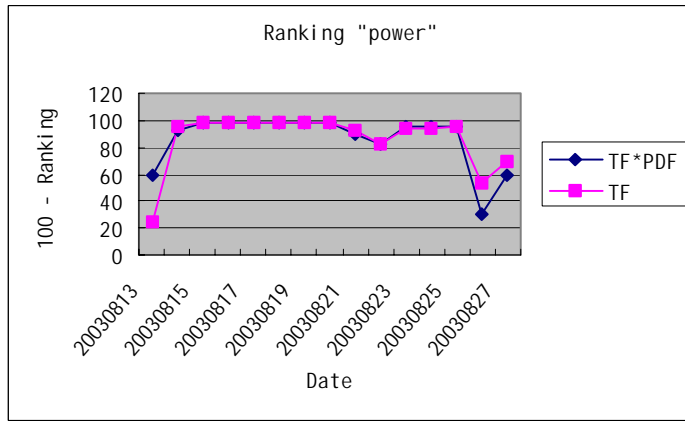
#### 5.4.6 Graphical Ranking of TF\*PDF and TF in Time Series

In order to look at the characteristic of TF\*PDF algorithm from a different view of point, graphical plots of the topic term ranking for TF and TF\*PDF in time series is used for comparison. Making use the TF ranks as a reference, we can see the effect of TF\*PDF of pushing up the topic term while suppressing the non-topic terms. Basically, there are 3 groups of terms we can classify. The first is the “constant top” that appear in the top of both TF\*PDF and TF list during the topic time frame. The second is the “drop out” terms, which are the non-topic terms come to the TF list. The third is the “push up” terms that ranked high because of the effect of TF\*PDF.

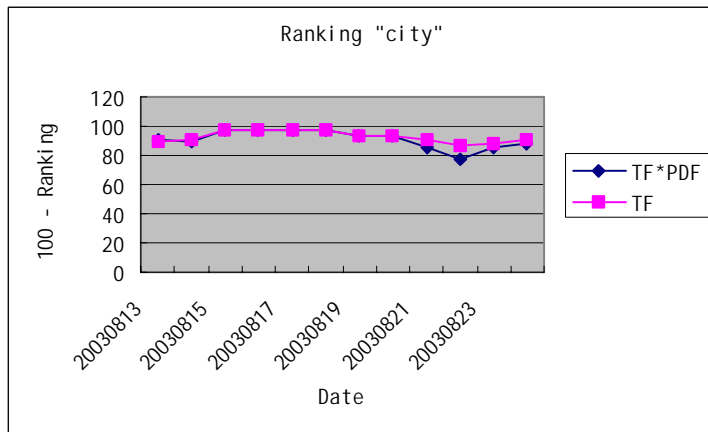
For the first type of topic terms group, Figure 31, 32 and 33 display the ranking plots of the “constant top” terms “blackout”, “power” and “city”. These three graphs show the same characteristic during the topic time frame, where terms’ rankings are constantly high. Since the scale on the y-axis is 100-Ranking, the highest value we can see in the graph is 99 (top or first place) while the lowest can go to negative few thousands depends on how many terms are there at the point. So, the top term should have a value of 99 in the graph while the second ranked at 98 and so on. When a term does not exist at a point, it is given a value of zero in the graph. For example, the first point of the term “blackout” is zero because it was not existing at the time, and then in the second point it has low ranking at the starting of the topic before it goes to constant high with value near to 99, where it is the highest ranking. Similar to the terms “power” and “city”, the ranking plots in both TF and TF\*PDF are at the top during the topic time frame. This is inline with our expectation that the “constant top” terms will have the ranking plots constantly high during the topic happening. In the following, we will see the characteristics of TF\*PDF capability of causing the “push up” and “drop out” of the topic and non-topic terms in the ranking.



**Figure 31: 100 – Ranking of “blackout” (constant top)**

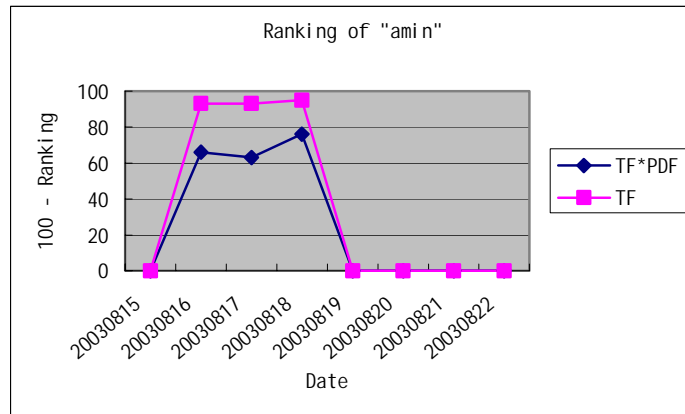


**Figure 32: 100 – Ranking of “power” (constant top)**

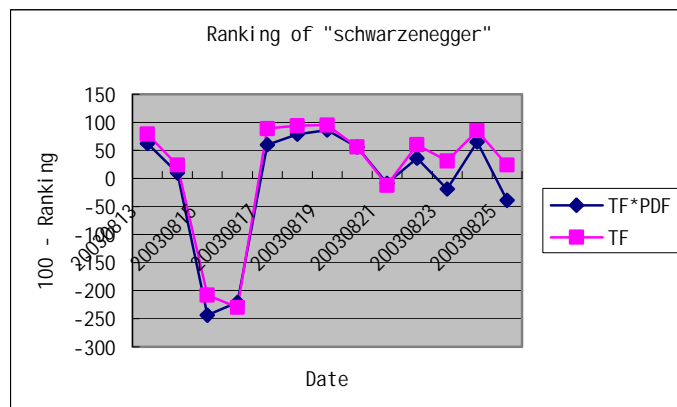


**Figure 33: 100 – Ranking of “city” (constant top)**

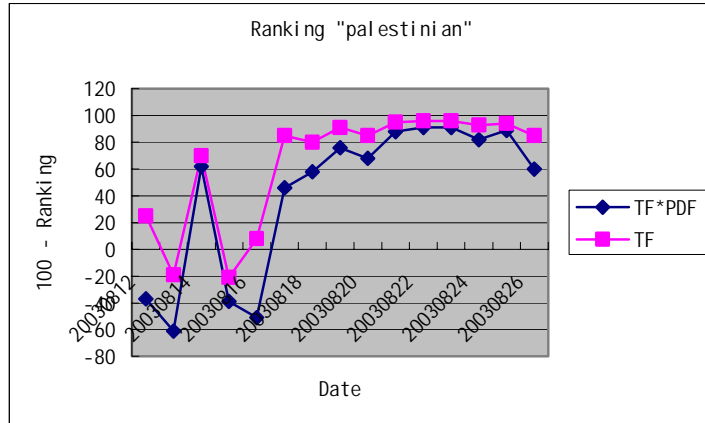
For the second group of “drop out” terms, we can see in the Figure 34, 35 and 36 that the non-topic terms “amin”, “schwarzenegger” and “palestinian” have a higher ranking plot for TF. This is because these terms are not topic terms and do not appear in the topic document burst, they automatically drop from the rank when the topic terms are pushed up by TF\*PDF algorithm in the TF\*PDF list. Especially we can see that the term “amin” ranked high in TF plot (low in TF\*PDF), because this non-topic term appears very frequently in one or few documents and gained high weight for TF, although this is not desirable. Thus, we have shown that the ranking of non-topic terms drop out from the TF\*PDF list but climb the TF list.



**Figure 34: 100 – Ranking of “amin” (drop out)**

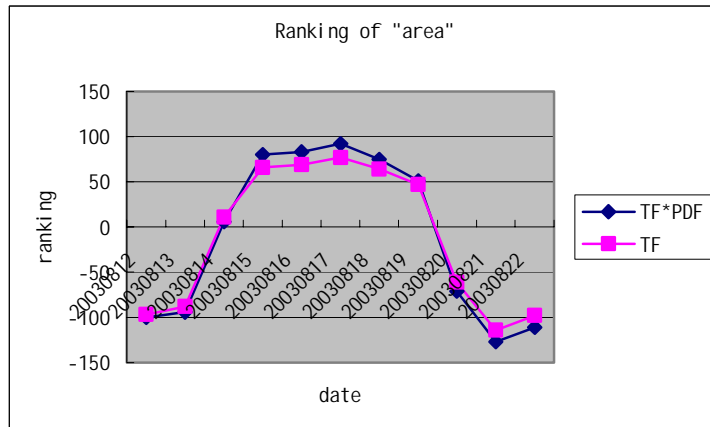


**Figure 35: 100 – Ranking of “schwarzenegger” (drop out)**



**Figure 36: 100 – Ranking of “palestinian” (drop out)**

For the third terms group of “push up”, we have here four ranking plots (Figure 37, 38, 39,40) to illustrate that the TF\*PDF algorithm automatically can push the topic terms up on the ranking list. We can see from the figures that the four “push up” terms, which are “area”, “night”, “failure” and “line” are forced high in the ranking for TF\*PDF during the topic time frame. Therefore, it can be concluded that the TF\*PDF takes advantages of the document burst effect to push the topic terms up in the ranking, and this will cause the non-topic terms to drop from the list automatically.



**Figure 37: 100 – Ranking of “area” (push up)**

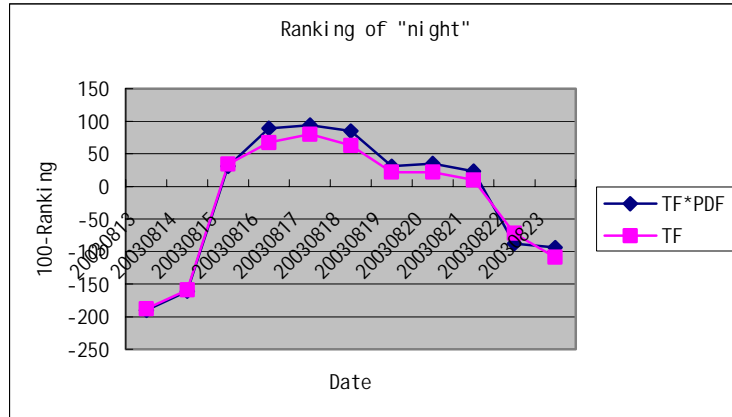


Figure 38: 100 – Ranking of “night” (push up)

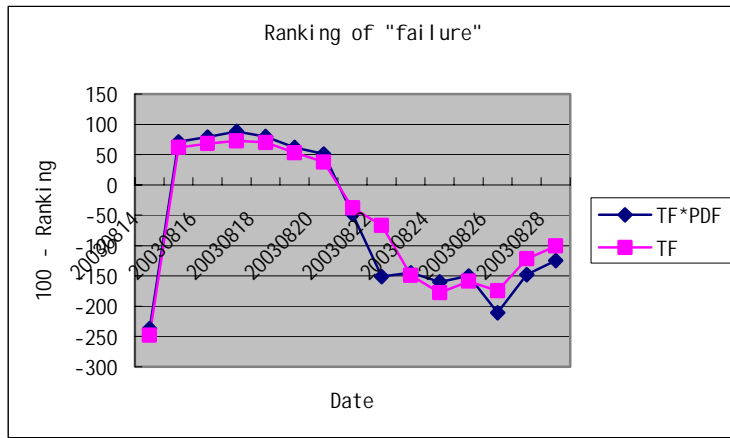


Figure 39: 100 – Ranking of “failure” (push up)

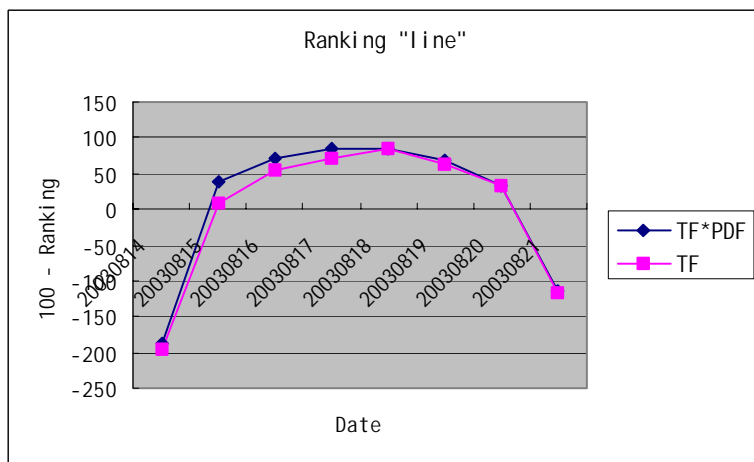


Figure 40: 100 – Ranking of “line” (push up)

### 5.4.7 Conclusion

An experiment done on the archive from July 20 to December 20, 2003 has been presented. A topic selection algorithm using a time window traversing along the TF\*PDF term weight distribution is proposed. Experimental results prove to us that the use of TF\*PDF would produce a higher value of  $W$  (Topic Weight Ratio), which in turn used to trigger the topic selection algorithm. The list below summarize the advantages and necessities for applying TF\*PDF algorithm instead of TF:

1. Put more weight and make the topic terms “outstanding” within the topic time frame
2. Weight strengthening process that makes the topic terms “outstanding” will possible the “discovery” and grouping of those topic terms with low weight value
3. Higher topic terms’ weight gives higher average weight of topic sentence, and thus makes the topic sentences extraction easier later on
4. Higher the value  $W$  (Topic Weight Ratio) works to trigger the topic selection from the archive well
5. At the same time pushing the topic terms up in the rank, TF\*PDF would filter out the non-topic terms that have high frequency in just one or few documents

## **5.5 Second Sample – Weekly News Topics Summarization**

This second experimental sample presents weekly news topics summarization. TF\*PDF algorithm is used to extract topic term words and then sentence clustering techniques is explored to generate weekly news topic summaries.

### **5.5.1 Corpus**

Similarly, experiments were run on the online news archive from 4 newswire sources: Associated Press (AP), The New York Times (NYT), Reuters and USATODAY. Section 5.5.2 presents the experiment done on a week news archive dated from May 13, 2002 to May 19, 2002. Next, Section 5.5.3 presents the experiment result done on the news archive dated from May 6 2002 to May 12, 2002. We have been trying to summarize weekly report in two consecutive weeks. The collected weekly news archives consist about 400-600 important news documents.

### **5.5.2 Experiment on Archive from May 13 to May 19**

Table 17 shows the top 30 most heavily weighted TF\*PDF terms. Table 18 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) in the sentences are highlighted in bold. For each respective sentence, its sentence unit vectors, status, date and source are stated in Table 19.

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 and 25. These sentences were then arranged chronologically to form a summary as shown in Section 5.5.2.1. 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 to 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 5.5.2.2.

**Table 17 : 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 18 : 25 Highest Weighted Sentences (May 13 to May 19)**

No.	Sentences	Weight
1	In <b>Washington</b> , a senior <b>Bush</b> administration <b>official</b> declined comment on the Likud resolution, but said President <b>Bush</b> remains committed to the establishment of a <b>Palestinian</b> state.	263.02
2	<b>White House</b> <b>officials</b> confirmed that <b>Bush</b> was told in a briefing a month before the <b>attacks</b> that bin <b>Laden's</b> al- <b>Qaeda</b> network had discussed hijacking <b>American</b> planes.	248.95
3	A <b>White House</b> <b>official</b> said on Saturday U.S. <b>intelligence</b> <b>officials</b> have detected "enhanced activity" that points to a potential new <b>attack</b> against the United States or <b>American</b> interests abroad.	233.07
4	<b>Palestinian</b> <b>officials</b> have been pondering elections and reform in the face of internal, international and <b>Israeli</b> demands for a restructuring of the <b>Palestinian</b> Authority and its <b>security</b> forces.	216.72
5	Meanwhile, 15 <b>Palestinian</b> Cabinet ministers offered to resign Saturday, <b>officials</b> said, a gesture to spur reforms in the <b>Palestinian</b> Authority , headed by <b>Arafat</b> .	206.36
6	Days after the summit concluded, <b>Israel</b> began <b>attacking</b> <b>Palestinian</b> territory in revenge for Palestinian suicide <b>bombings</b> on <b>Israeli</b> targets.	202.87
7	In the interview, the <b>official</b> described a plan to act against Mr. bin <b>Laden</b> that was developed in August, approved at the level of the "deputies" the No. 2 <b>officials</b> in several departments and then approved by the top cabinet <b>officials</b> on Sept. 4.	202.78
8	CHICAGO - President <b>Bush</b> will keep pushing for the creation of a <b>Palestinian</b> state despite a vote by <b>Israel's</b> ruling right-wing Likud party never to accept one, the <b>White House</b> said on Monday.	198.67
9	Wolfowitz repeated the <b>Bush</b> administration's view that an end to <b>Israeli</b> military occupation of <b>Palestinian</b> territory and a <b>Palestinian</b> state was key to solving the Arab- <b>Israeli</b> problem.	198.06
10	When <b>Israeli</b> troops moved into <b>Palestinian</b> towns last month, confining <b>Arafat</b> to his Ramallah office, many <b>Palestinians</b> gave him their backing, viewing the <b>Israeli</b> action as part of a larger <b>attack</b> on <b>Palestinian</b> aspirations for statehood.	194.94

11	The New York Times on Thursday quoted a senior <b>Israeli official</b> telling reporters in <b>Washington</b> on condition of anonymity that reforming the <b>Palestinian security</b> forces could not be accomplished while <b>Arafat</b> was in charge.	193.83
12	That plan, which was drawn up by high-ranking <b>officials</b> among several Cabinet departments, was awaiting President <b>Bush's</b> review when the World Trade Center and Pentagon were <b>attacked</b> .	192.28
13	<b>Israeli</b> forces have encircled <b>Palestinian</b> cities in the West Bank and set up checkpoints across <b>Palestinian</b> territories which they say are meant to prevent <b>attacks</b> on <b>Israelis</b> .	191.40
14	<b>Bush's</b> brother Jeb <b>Bush</b> , the Florida governor, faces re-election this year and also is depending on <b>Cuban Americans</b> , who vote heavily <b>Republican</b> .	188.67
15	On Wednesday night, the <b>White House</b> said President <b>Bush</b> was <b>warned</b> by <b>American intelligence</b> agencies in early August of Mr. bin <b>Laden's</b> desires to hijack airplanes.	187.19
16	<b>Carter</b> told Castro and leading <b>Cuban</b> scientists that he had asked <b>White House</b> , State Department and <b>intelligence officials</b> specifically if <b>Cuba</b> was transferring technology or other information that could be used in <b>terrorist</b> activities.	184.71
17	"It is the time for change and reform," <b>Arafat</b> told the <b>Palestinian</b> Legislative Council in a speech on the day <b>Palestinians</b> mark as the Nakba of the founding of <b>Israel</b> in 1948, which displaced hundreds of thousands of <b>Palestinians</b> .	184.10
18	United States <b>intelligence officials</b> said that they began to intercept communications among <b>Qaeda</b> operatives discussing a second major <b>attack</b> in October, and that they have detected recurring talk among them about another <b>attack</b> ever since.	183.17
19	<b>Arafat</b> responded to widespread pressure from ordinary <b>Palestinians</b> , <b>Israel</b> and foreign leaders by calling a speech to the <b>Palestinian</b> parliament on Wednesday for elections and reforms.	182.48
20	Indian <b>intelligence officials</b> say Dawood Ibrahim, a Bombay crime boss wanted in India, is back in <b>Pakistan</b> and plotting retaliatory <b>attacks</b> with Pakistani <b>intelligence officials</b> .	182.43
21	The <b>White House</b> also acknowledged on Friday that <b>security officials</b> had prepared a presidential order for a campaign to dismantle al <b>Qaeda</b> .	178.97
22	<b>Arafat</b> also condemned <b>Israel's</b> six-week incursion into West Bank cities to root out <b>Palestinian</b> militants that also destroyed much of the infrastructure of the <b>Palestinian security</b> forces.	174.92
23	WASHINGTON, May 17 The <b>White House</b> began an aggressive <b>attack</b> on <b>Democrats</b> in Congress today as President <b>Bush</b> tried to contain the political fury over a <b>warning</b> he received last August that Osama bin <b>Laden</b> might be planning a hijacking.	174.19
24	Key members of Congress have asked whether the government had information pointing to the <b>attacks</b> on America after the <b>White House</b> disclosed President <b>Bush</b> had had an <b>intelligence</b> briefing in early August that included concerns bin <b>Laden's</b> group might try to hijack a passenger plane.	172.13
25	In response to the uproar after the disclosure of the August <b>warning</b> to Mr. <b>Bush</b> , <b>White House officials</b> insisted that they had no serious evidence last summer that Al <b>Qaeda</b> was considering a suicide hijacking.	171.90

**Table 19 : 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	USA Today
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	USAToday
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

### 5.5.2.1 Summary1

*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.*

### 5.5.2.2 Summary2

*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.*

### 5.5.2.3 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 maximum acute rules to be classified into the cluster. 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 will 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 being evaluated 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 successfully. The related TF\*PDF terms in the lower part of Table 17 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 a document. The "all capital letter" words are excluded in weight counting.

#### 5.5.2.4 Page Extraction

The top three pages with highest average weight are as in Figures 41, 42 and 43 below. These three pages are ranked at the top (same sequence) when both counted by using either the top 30 or all the TF\*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 slightly higher at 132.34, 130.66 and 122.70.

Figure 41 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 18 with 248.95 average sentence weights. Figure 42 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 18 matches the first sentence in the Figure 43 page.



**Figure 41 : USATODAY, May 17**

**Top Stories - Reuters**

[Top Stories](#) | [AP](#) | [Reuters](#) | [The New York Times](#) | [USA TODAY](#)

## Arafat Links Elections to Israeli Pullout

*Fri May 17, 6:36 PM ET*

*By Michele Gerstberg*

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.

**Photos**



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

Figure 42 : Reuters, May 17

**Top Stories - Reuters**

[Top Stories](#) | [AP](#) | [Reuters](#) | [The New York Times](#) | [USA TODAY](#)

## 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.

**Photos**



"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 43 : Reuters, May 13

### 5.5.3 Experiment on archive dated from May 6 to May 12

Table 20 shows the top 30 most heavily weighted TF\*PDF terms extracted from May 6 to May 12. Figure 44, 45 and 46 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 20 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 cases. 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.

**Table 20 : Top TF\*PDF Terms (May 6 to May 12)**

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

**Top Stories - Reuters**

[Top Stories](#) | [AP](#) | [Reuters](#) | [The New York Times](#) | [USA TODAY](#)

## 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.

**Photos**



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 44 : Reuters, May 7

**Top Stories - USA TODAY**

[Top Stories](#) | [AP](#) | [Reuters](#) | [The New York Times](#) | [USA TODAY](#) | [NPR](#)

## 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.

**USATODAY.com** 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.

**Today in the Sky:**

Figure 45 : USATODAY, May 9



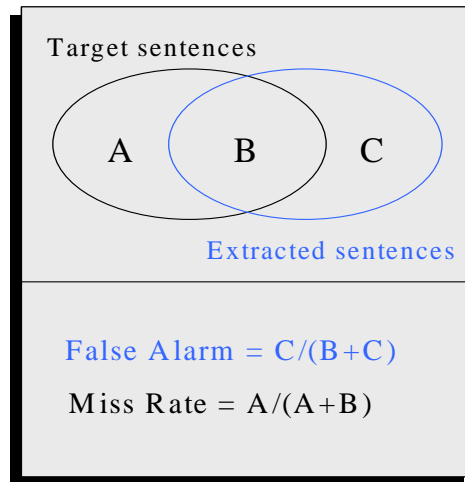
Figure 46 : AP, May 11

#### 5.5.4 About Evaluation

Many type of measures [jones97, tague92, salton89\_2, TREC] 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 [mani99\_2] was the initial work on developer-independent evaluation of automatic summarization system.

Precision/recall has been the most widely used methodology for doing evaluation (Section 3.3). If we are to use this method to do an evaluation and map our results according to the Figure 47 below, we would calculate the miss rate and false alarm as follow:

- 80% of all the 25 sentences were clustered successfully (area B in Figure 47)
- Two sentences in area A. These are target sentences, which were not clustered. Therefore, Miss Rate =  $2/(2+20) = 0.09$
- One sentence in C area. This is sentence grouped wrongly into the first summary cluster. Therefore, False Alarm =  $1/(1+20) = 0.05$



**Figure 47 : Precision/Recall Mapping**

### **5.6 Third Sample - Generating a Better-Coverage Summary of News Topics using Time Features and Sentence Clustering**

This part of our system is unique in using the event time frame and sentence clustering methodology to generate the better-coverage summary of a topic. The main idea is that topics, which are discussed in several channels concurrently are likely to be important and can be detected by our novel TF\*PDF algorithm. Next, by calculating the term weight acceleration, we can recognize the topic terms and time frame. Then, the sentences spanning in the event time frame are processed using clustering in order to generate a better-coverage summary of the topic. With the help of this system, we can receive a summary report of main topics periodically.

#### **5.6.1 A Better-Coverage Summary**

An upgraded capability of our system is presented here. Basically, the desired capability was to detect the trends and on-set of the event keywords by calculating their weight acceleration values, and generate a better-coverage summary for the event, which may consists of many stories spanning in the event time frame. This sample was run on a corpus of 850 documents from Jan 27 to Feb 15 2003.

These documents were collected from the four same news sources, which were Associate Press, Reuter, New York Time and USA Today.

Acceleration values are based on time features computed for each term. These time features enhance topic terms and time frame detection. [allan99, allan00, swan00] exploit time varying features for clustering in order to discover sets of related stories for a single event. Our approach uses similar distribution detection by comparing patterns of acceleration values of different terms, for detecting the event terms, which are presenting some common features in a particular time frame.

We calculated the TF\*PDF value of the terms from Jan 29 to Feb 15. The selected terms with highest weight are illustrated in Figure 48. The values (weight) of these selected terms are presented in Table 21. We used a three-day-wide corpus to calculate the TF\*PDF value for each day. For example, when calculating the TF\*PDF value for Jan 29, we used the documents from Jan 27,28,29, and for calculating the value for Jan 30, we used the documents from Jan 28,29,30.

We analyzed the terms' weight in Figure 48, and found out that they are characteristically divided into two groups. The first group contains the terms having sustaining weight, which are "Bush", "Iraq", "Saddam", "Korea", and "North". These terms are general words and related to some on-going important topics. The terms in the second group have their weight increases from almost zero to become one of the highest in a short period, and then decrease speedily until become low. These terms are "space", "shuttle" and "columbia". Their weight presents a "hill-shape" in the period from Feb 1 to Feb 9. This "hill-shape" characteristic of term weight contributes to the pattern for detecting the event terms, by calculating the term weight acceleration values with the logic presented in Table 23. The resulted term weight acceleration values are shown in Table 22.

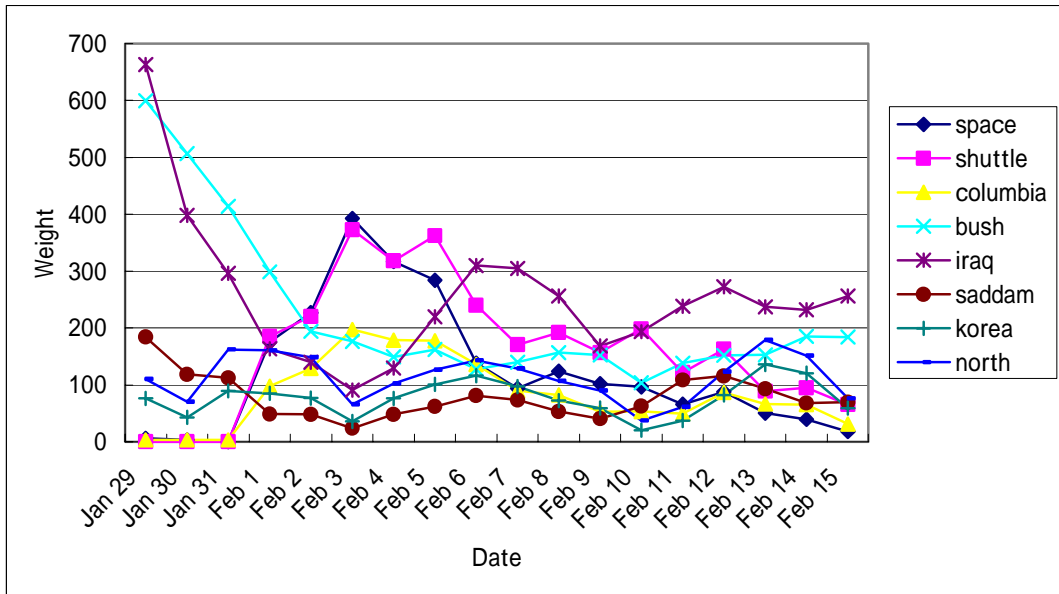


Figure 48 : TF\*PDF Weight of Selected Terms from Jan 29 to Feb 15

Table 21: TF\*PDF Weight of Selected Terms from Jan 29 to Feb 15

	space	shuttle	columbia	bush	iraq	saddam	korea	north
Jan 29	6.18	0	3.07	599.56	663.44	184.07	75.97	110.3
Jan 30	3.02	0	3.03	506.41	397.79	118.63	42.52	69.93
Jan 31	0	0	3.02	414.08	295.8	112.18	89.33	162.24
Feb 1	175.21	185.18	97.5	298.55	163.51	48.84	85.13	160.56
Feb 2	226.74	220.6	128.65	193.81	139.72	48.17	76.83	148.52
Feb 3	392.54	372.52	197.26	176.74	90.74	23.32	35.67	65.88
Feb 4	316.57	318.07	178.91	149.01	129.59	48.48	76.04	102.28
Feb 5	284.13	362.29	177.47	162.35	219.81	61.78	101.13	126.74
Feb 6	138.41	239.87	135.69	127.51	309.55	81.3	116.05	143.26
Feb 7	93.92	170.25	89.77	139.91	304.51	73.69	96.88	128.58
Feb 8	123.94	191.77	82.07	156.75	255.73	52.98	72.62	107.14
Feb 9	102.07	156.62	52.43	152.3	168.3	40.37	59.19	89.63
Feb 10	96.26	198.63	53.44	103.61	193.28	62.52	20.03	37.32
Feb 11	65.62	121.96	49.33	138.22	238.43	108.46	37.12	61.51
Feb 12	87.91	162.56	86.65	152.06	272.5	115.91	82.48	123.07
Feb 13	50.34	88.57	66.33	153.1	237.33	93.31	135.79	178.54
Feb 14	39.25	94.57	65.45	185.04	232.08	68.03	120.56	151.14
Feb 15	18.07	65.31	31.22	184.35	256.22	69.79	59.16	76.44

**Table 22 : Weight Acceleration of Selected Terms from Jan 31 to Feb 15**

	space	shuttle	columbia	bush	iraq	Saddam	korea	north
Jan 31	-3.02	0	-0.01	-92.33	-101.99	-6.45	46.81	92.31
Feb 1	175.21	185.18	94.48	-115.53	-132.29	-63.34	42.61	90.63
Feb 2	51.53	35.42	31.15	-104.74	-23.79	-0.67	-8.3	-12.04
Feb 3	165.8	151.92	68.61	-17.07	-48.98	-24.85	-41.16	-82.64
Feb 4	89.83	97.47	50.26	-27.73	-10.13	25.16	-0.79	-46.24
Feb 5	-32.44	44.22	-1.44	-14.39	90.22	13.3	25.09	24.46
Feb 6	-145.72	-122.42	-41.78	-34.84	89.74	19.52	14.92	16.52
Feb 7	-44.49	-69.62	-45.92	-22.44	84.7	11.91	-19.17	1.84
Feb 8	-14.47	-48.1	-7.7	16.84	-48.78	-20.71	-24.26	-21.44
Feb 9	-21.87	-35.15	-29.64	12.39	-87.43	-12.61	-13.43	-17.51
Feb 10	-5.81	42.01	-28.63	-48.69	-62.45	22.15	-39.16	-52.31
Feb 11	-30.64	-76.67	-4.11	-14.08	45.15	45.94	-22.07	-28.12
Feb 12	-8.35	-36.07	37.32	13.84	34.07	7.45	45.36	61.56
Feb 13	-37.57	-73.99	17	1.04	-35.17	-22.6	53.31	55.47
Feb 14	-11.09	-67.99	-0.88	31.94	-5.25	-25.28	38.08	28.07
Feb 15	-21.18	-29.26	-34.23	31.25	24.14	-23.52	-61.4	-74.7

### 5.6.2 Term Weight Acceleration

Table 22 shows the term weight acceleration value of the selected terms. These values are calculated by using the logic in Table 23. There are a rules set and three acceleration states being defined. A positive acceleration value leads to a +(positive) state; zero acceleration value gives a 0(neutral) state; and negative acceleration value (deceleration) results a -(negative) state. There is only one zero acceleration value appearing in Table 22. Zero acceleration value of a term happens when the rule number 5 takes effect, where the weight of the term on day T (WT) and day T-1 (WT-1) are the same. The rule number 1 will trigger the acceleration state into positive when WT is larger than both the WT-1 and WT-2; and the acceleration value (AT) will be calculated as WT minus WT-1. We execute the rule number 2 when the WT larger than WT-1, but smaller than WT-2. In this case, we calculate the acceleration value (AT) as WT minus WT-1 if the previous state is positive, or WT minus WT-2 if the previous one is negative. Also in this case, the new state will be the same with the previous state.

In the same way, we can explain the rules number 3 and 4. In addition, the rule number 5 is always evaluated before the rule number 6. Hence, none of the value in the Table 22 is resulted by rule number 6.

**Table 23 : Logic for Calculating Term Weight Acceleration**

	Conditions		Weight Acceleration ( $A_T$ )	New State
	Weight Comparison	Previous State		
1	$W_T > W_{T-1}$ $W_T > W_{T-2}$	X	$A_T = W_T - W_{T-1}$	+
2	$W_T > W_{T-1}$ $W_T < W_{T-2}$	+	$A_T = W_T - W_{T-1}$	+
		-	$A_T = W_T - W_{T-2}$	-
3	$W_T < W_{T-1}$ $W_T < W_{T-2}$	X	$A_T = W_T - W_{T-1}$	-
4	$W_T < W_{T-1}$ $W_T > W_{T-2}$	+	$A_T = W_T - W_{T-2}$	+
		-	$A_T = W_T - W_{T-1}$	-
5	$W_T = W_{T-1}$	X	0	0
6	$W_T = W_{T-2}$	X	0	Previous state

From either Table 21 and Figure 48 as well, we can see that the  $TF*PDF$  value of the terms “space”, “shuttle” and “Columbia” increase from nearly zero before Feb 1, peak on Feb 3, decrease speedily until Feb 9 and then gradually from Feb 9 onwards. Therefore, in the period from Feb 1 to Feb 9, these three terms in Table 22 displays a row of at least 4 acceleration values, followed by at least 4 deceleration values in a row. This pattern shows us that these 3 terms became a popular topic in the period. The other term that show similar pattern was “Iraq”, having 3 acceleration and 3 deceleration values in a row from Feb 5 to Feb 10. However, this term is the kind of sustaining term, together with terms like “Bush”, “Saddam”, “Korea” and etc, regarding the on-going topics. After surveying the archive, we realize that the weight of the term “Columbia” before Feb 1 was caused by the phrase “District of Columbia” in a number of documents.

By detecting the time frame, we can suggest the appearing period of the documents of an event. Consequently, we can calculate the TF\*PDF term values and do clustering with the sentences happened within this time frame, in order to perform a better-coverage summarization of the event. Therefore, we are able to provide a better-coverage summary of an event to our users. Also, we can tell our users which topic was hot in a particular period and which topic is still going on. In the following section, we will present the summary that we generate from the period from Feb 1 to Feb 9. In the future, we may “fine-tune” our acceleration calculating rules set in order to produce finer values for matching various keywords of events, basing on the fact that acceleration values of keywords from same kind of events may present common pattern and characteristics.

### 5.6.3 Topic Summary from Feb 1 to Feb 9

Table 18 shows the 30 terms with highest TF\*PDF value calculated using the archive from Feb 1 to Feb 9. After calculating the average weight of each sentence, we perform sentence clustering using these terms to produce the summary shown in the following subsection. We have successfully clustered 18 sentences after evaluating 26 sentences and arranged them in chronological order. However, some of the sentences have dangling anaphors that may sometimes make the summary difficult to understand. In the future, we would like to focus on more efficient sentences ordering using natural language processing techniques. Table 25 displays the unit vectors, compiling date and source of each sentence in the summary according to their position.

**Table 24 : Top 30 TF\*PDF Terms from Feb 1 to Feb 9**

<b>Term</b>	<b>Weight</b>	<b>Term</b>	<b>Weight</b>	<b>Term</b>	<b>Weight</b>
shuttle	952.72	american	254.33	Flight	184.85
space	781.08	council	248.19	Official	183.8
iraq	726.18	korea	228.06	Budget	182.17
nasa	634.86	center	222.03	Seven	180.44
bush	600.9	house	221.34	Secretary	176.53
columbia	468	nuclear	219.13	Iraqi	174.75

security	379.75	program	218.61	National	170.59
north	343.23	world	201.42	Saddam	161.87
Powell	325.32	white	196.56	International	161.26
washington	315.54	agency	189.39	crew	159.5

**Table 25 : Sentence's Unit Vector, Date and Source (Feb 1 to Feb 9)**

Sentence Number	Unit Vector	Date, Time (Feb)	News Sources
1	Bush seven space shuttle Columbia	1, 2:14 PM	Reuters
2	space shuttle Columbia seven	1, 3:23 PM	Reuters
3	space agency shuttle Columbia seven	1, 4:02 PM	Reuters
4	space shuttle Columbia NASA	1, 5:34 PM	Reuters
5	world space shuttle Columbia seven	1, 6:35 PM	AP
6	shuttle space program international	1, 6:39 PM	AP
7	NASA official shuttle Columbia program	1, 6:41 PM	Reuters
8	space seven shuttle Columbia	2,12:17 PM	Reuters
9	space center shuttle program	2, 3:05 PM	NYT
10	NASA budget space shuttle	3, 7:35 AM	USA TODAY
11	Shuttle program NASA budget	3, 5:38 PM	Reuters
12	NASA shuttle space	4, 7:38 AM	USA Today
13	Columbia NASA shuttle space	4, 7:40 AM	USA Today
14	space shuttle	4, 7:45 AM	USA Today
15	NASA shuttle flight space program	4, 7:45 AM	USA TODAY
16	shuttle program space agency	4, 9:03 AM	NYT
17	space agency shuttle	5, 9:02 AM	NYT
18	space shuttle	8, 3:00 PM	NYT

#### 5.6.4 Result Summary for the Time Frame from Feb 1 to Feb 9

*President Bush said on Saturday the nation mourned the death of seven astronauts aboard the doomed space shuttle Columbia. Immediate popular reaction in Baghdad on Saturday to the loss of the U.S. space shuttle Columbia and its seven-member crew -- including the first Israeli in space -- was that it was God's retribution. The U.S. space agency said on Saturday it could not yet determine what caused the destruction of the space shuttle Columbia in which seven astronauts were killed. The space shuttle Columbia was NASA 's oldest space plane and the first of its vaunted orbital fleet, flying its maiden mission in 1981. Government officials around the world expressed condolences Saturday for the loss of the U.S. space shuttle Columbia and its seven crew members, who included the first Israeli to fly in space. The horrific end of shuttle mission STS-107 was a devastating blow to the nation's space program; the Challenger explosion led to a 2 1/2-year moratorium on launches, and Saturday's accident could bring construction of the international space station to a standstill. NASA officials said they planned to keep the shuttle fleet grounded until they understand why the Columbia*

*was lost, and slow down manufacturing on the shuttle program. In the tiny space towns that dot Florida's Cape Canaveral, the mourning began in earnest on Sunday for the seven fallen astronauts of space shuttle Columbia. Inside the space center, where there is a contingency plan for everything, even this, two of the top administrators of the space shuttle program talked in precise language about failed sensors and damaged tiles on the left wing of the shuttle. Last March, NASA's own safety advisory panel reported that "budget projections for the space shuttle are insufficient to accommodate significant safety upgrades. These observers had raised concerns over the years about the safety of the shuttle program, citing reasons ranging from shuttle design to NASA budget constraints. A. NASA declines to speculate, but if the fleet were down to two shuttles, the scheduled 2004 completion of construction of the space station would almost certainly be set back. Because Columbia was almost 7,000 pounds heavier than NASA's other shuttles, it was more difficult to get it to the space station, which sits in a relatively distant orbit from the Earth. For the moment, investigators are interested in the possibility that damage to the space shuttle's heat-resistant tiles during liftoff on Jan. 16 might have caused the shuttle to come apart as it descended toward Earth on Saturday morning. Over 25 years, Dittmore climbed the ranks at NASA, from propulsion engineer, to shuttle flight director, to deputy assistant director of the space station program, and finally to manager of the entire shuttle program in 1999. But Ron D. Dittmore, the shuttle program manager, said today that the engineers might have been wrong and that the space agency was "redoing the analysis from scratch." The space agency was warned in 1990 that the protective tiles around the shuttle's wheel wells were particularly vulnerable to damage and failure, inviting catastrophe because those tiles protect both fuel tanks and the shuttle's hydraulic system. "There was uncontroverted testimony," he said, "that there were delays in the safety upgrades to the space shuttle and that the upgrades would make the space shuttle much more safe."*

## **5.7 The Uniqueness of Our Approach**

Our idea for topic detection is unique and does not require retrospective archive. Rather than applying document clustering techniques which requires more computational complexity, we detect the topic terms by using our TF\*PDF algorithm, which takes advantage that the main topics will be reported broadly in many documents in many news channels concurrently. Whereas, the existing news extraction and summarization systems discussed in previous chapter are seen tending to do daily document clustering for news topic summarization and present on their main page, the main topics that record a long thread of related documents.

Our system works on the basic concept that whenever there is a popular topic appearing, it will be discussed frequently in many news documents in majority newswire sources. Thus, instead of grouping the information from all sources into a "large" mixed corpus and calculating each

document's keyword with a standard TF\*IDF algorithm like in TDT, we rather give equal importance to news articles coming from each newswire source and channel it to our system in a parallel way. The terms that explain the popular topics appear frequently in many documents in every channel and should be weighted significantly. Whenever the majority channels contain the terms with high term weights concurrently, these are the terms that explain the main topics discussed recently. As a result, we can extract the topic terms in the news archive in a unique and more efficient way.

Next, by exploiting the changing characteristic of term weight, we measure the information "surprise" of a term of a particular time period by calculating its term weight acceleration value. Consequently, it is possible to recognize the up-rise and fall of a topic by matching its acceleration pattern of positive and negative sequences in a particular topic time frame. Later, we would generate a better-coverage summary by doing sentence clustering on the sentences appearing in this time frame. Topic summary is a way of output to allow users to understand the topic story lines better, compared to the conventional topic detection and tracking systems [swan00, havre02, patentminer97] that specialize in presenting the event features in their timelines interface.

Table 26 gives a comparison to the discussed systems. News Topics Summarizer is the part of our system presented in this chapter, while ETTS is the part of our system useful in detecting and tracking the static type of information that will be presented in the next chapter.

**Table 26 : A comparison between systems**

	<b>PatentMiner</b>	<b>TDT</b>	<b>ChangeDetect</b>	<b>News Topic Summarizer</b>	<b>ETTS</b>
Input Data	Patent DB	Tagged news corpus	Page URL	Raw news articles	Hypertext document
Predefined Query	Trend shape	None	Words or none	None	Topic keyword
Topic Weighting	Phrases frequency	Document frequency	Change	Term weight	Term weight
Computation algorithm	Sequential pattern matching	TF*IDF &	Clustering	Boolean logic	TF*PDF
Visualization	TF*PDF	Yes	None	None	None
Summarization	None	None	None	None	Yes

## 5.8 The Merits of Our Approach

**Table 27 : A comparison of TF\*IDF and TF\*PDF**

	<b>TF*IDF</b>	<b>TF*PDF</b>
Goal	Detect the uniqueness (keywords) in a document	Detect topic terms (keywords) in a collection of documents
Retrospective archive	Yes	No
Risk of losing topic tracking	Yes	No
Document Clustering	Yes	No (Sentence clustering)
Detection	Document frequency (cluster size)	TF*PDF terms and sentence clustering

Table 27 above give a comparison and outline the points that we have discussed earlier on the conventional TF\*IDF and our TF\*PDF algorithm. This table gives a clearer picture specifying the goals and characteristics of these two algorithms while doing news topics detection and tracking.

Finally, the table 28 below notes the merits of our system using TF\*PDF in News Topics Detection and Tracking.

**Table 28 : Merits of TF\*PDF in Topic Detection and Tracking**

Merits	Description
No retrospective corpus needed	<ul style="list-style-type: none"> <li>• TF*IDF needs a fairly large back corpus for calculating an effective (working) IDF</li> <li>• Therefore, IDF poses risk of topic drift/deviation depending on the back cases</li> </ul>
No risk of losing the tracks of popular topic	<ul style="list-style-type: none"> <li>• When more topic documents appear (document burst), lower will be the value of IDF, and thus lower the weight of the topic terms.</li> <li>• In reverse, TF*PDF give heavier weight to the topic terms while more topic documents appears, because high related document appearance means that the topic is popular</li> </ul>
Computational Complexity	<ul style="list-style-type: none"> <li>• Sentence clustering instead of document clustering performed</li> <li>• Objective of creating a unique and efficient detection algorithm achieved</li> </ul>
Flexibility	<ul style="list-style-type: none"> <li>• Our approach has more flexibility calculate and associate the topic terms with similar weight distribution</li> <li>• Topic weight variation reveals topic characteristics (such as start, accelerate, peak, sustain, end and etc)</li> </ul>

## 5.9 Conclusion

The rapid growth of online news archive has led to the need of an intelligent tool to assist users in detecting and digesting the important news topics. We address the first problem of topic detection by introducing a topic selection algorithm using TF\*PDF algorithm which makes the topic term weight “outstanding”, and analyze the topic term weight acceleration value. This TF\*PDF algorithm is designed in a way to give significant weight to the words that explain the “hot” topic in many documents in many news channels. Group of words from different topic may show different characteristic of weight at different time. Words explaining the temporal “hot” topic are likely to

present a row of positive acceleration values followed by a row of negative values in a certain time frame. Later after recognizing these topic words and happening time frame, we approach the second problem of topic digestion by doing sentence vector clustering on important sentences appearing in the topic time frame. In this way, we can produce a better-coverage summary on each topic and deliver a summary report of main topics to the users periodically. No retrospective corpus is used in the statistical calculation for detecting event's keywords. Furthermore, we show that TF\*PDF algorithm doesn't risk losing the detection and tracking of events in the midst of document burst, which is not the case if using TF\*IDF approach that basically do document clustering and then multi-documents summarization for a topic summary from each cluster. TF\*PDF algorithm for weighting the topic terms is novel. Compared to TF, TF\*PDF could make the topic terms weight "outstanding" during the topic happening time frame. TF\*PDF terms and sentence clustering approach is a new way for summarizing news topics.

## Chapter 6 Emerging Topic Tracking System (ETTS)

### 6.1 Introduction

In this chapter we present the part of our system that tracks the changes happened in a particular area of user's interests on the static Web and generate a summary of emerging topic to the user. This system consists of three main components: Area View System, Web Spider and Summary Generator. Area View System as a Meta-search engine will direct the user keyword to a commercial search engine, get the hits, do further analysis and derive a number of most relevance domain sites. Then, Web Spider will dispatch and scan all these domains at a certain time interval to collect all the modified and newly added html pages. Lastly, Summary Generator will first extract all the newly added sentences (or changes) from the collected html pages and then count the term weight in the changes by using our innovated TF\*PDF (Term Frequency \* Proportional Document Frequency) algorithm. Terms that deem to explain the emerging topic will be heavily weighted. Sentences with the highest average weight will be extracted to form a summary of emerging topic. We refer our system as ETTS (Emerging Topic Tracking System).

Due to its open characteristic, the Web is being posted with vast amount of new information changes continuously. Consequently, at any time, it is conceivable that there will be hot issues (emerging topics) being discussed in any information area on the Web. However, it is not practical for the user to browse the Web manually all the time for the changes. Thus, we need this Emerging Topic Tracking System (ETTS) as an information agent, to detect the changes in the information area of our interests and generate a summary of changes back to us regularly. This summary of changes will be telling the latest most discussed issues and thus revealing the emerging topics in the particular

information area. With this system, we will be “all time aware” of the latest information trends of our interests in the WWW information space.

## **6.2 Related Works and Motivations**

If we are stuck with the conventional view of the Internet, then we are in trouble because its contents are changing too quickly. Thus, users or professionals would like to be always updated with the latest hot topics emerging in the particular information area of their interest. However, due to the fact that the information in the Web is overwhelming and changing dynamically, updating ourselves by browsing through some particular Web sites of interest manually and regularly is both a difficult and time consuming job. Thus, we need a kind of information system, which can track and summary us the changes that appeared on the pages or information area of our interests.

Thus far, there have been quite a number of commercial tracking tools become available for services online. Basically, when users need the system to track a particular html page on the Internet for them, they need to register the URL of that particular html page with the system. Upon any changes happened to the page, the user will be acknowledged through e-mail. Usually, this kind of tracking tool can detect every detail of changes, but unfortunately because of this technology advancement, every trivial change that happened to the page would trigger the system to push user with acknowledgement e-mail. In order to solve this problem, some systems, i.e. WebBeholder [santi98], allow user to set a trigger level they prefer. Here, if and only if the total changes score is greater than the trigger level, the system will be triggered to send e-mail to the user. But there is always no appropriate trigger level can be defined accurately since there are many possible types of changes in html page (title, header, content character, color, text style and etc) with different score. So, the users might be fed with e-mail although the change(s) is not interesting to them and vice versa.

User can register multiple pages with a tracking system in order to keep watching in a wider area, but the users have to bear in mind that, in one single day, they may receive many emails of acknowledgement just because of some uninteresting changes. But the users yet to know this until they go and look for the changes on the pages registered. Always, the users need to scratch their head in order to figure out the part of the page that has changed. Output from concurrent tracking systems always show little or no information on how the pages have changed. Thus, the AT&T Internet Difference Engine (AIDE) [douglis98] and its successor TopBlend [chen00] have been contributing in solving this problem by automatically compares two html pages and creates a “merged” page to show the differences with special HTML markups. But if the difference is too substantial, the “merged” page can be very messy or even unreadable. Merging two pages into one page will raise the danger of creating syntactically or semantically incorrect HTML.

Besides providing service on tracking the URL(s) registered by the user, some systems also featured in detecting the new pages containing the input keyword from user. These systems even claimed to be the “best search monitoring tool” on the market. With them, user is allowed to input keyword of interests and select one of the commercial search engines for tracking purpose. Then, in a certain time interval, with the aids of the particularly selected search engines, these systems will detect the new pages related to the keyword and acknowledge the user. However, author found that the relevancy and the status of “new” of the results are not convincing after trying on them, because of the imperfectness of search engine. Users will always be presented with the links to a number of detected new pages. Each new page may contain the keyword but the keyword may not be describing the main topics of the page. Users are always left alone to figure out what are the main topics behind the changes happened to the area represented by the keyword. One of the main reasons causing inaccuracy in recognizing the correct new pages here can be the quality and the “up-to-date” status of the hits returned by search engine. [[cliff99]] quoted, “A particular search engine will run many

robots at the same time, in an attempt to keep its information current. However, the sheer size of the World Wide Web means that it will take some time (weeks) for a new page reference to appear in response to a user's search."

Additionally, some systems provide different columns to get input keyword from user for tracking new information in different category. For each category, these systems will visit some pre-determined web sites for new pages with the keyword appeared in it. For example, if a user want to track new information of a stock, these systems may constantly check for the appearing of the keyword on the web site, for example CBS or CNN Market pages. However, relying on only one or a few static web sites as the source(s) to gain new information from a wide general area will doubt the completeness of the changes happened to a small sub area can be reported.

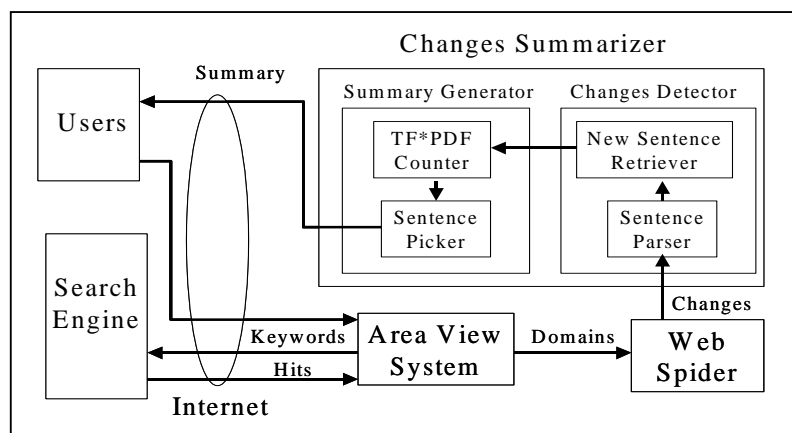
Thus far, we have studied a number of concurrent tracking tools and some of their deficiencies. Generally, in specific page tracking, user will be notified when the page is updated. While in keyword tracking, user will be presented with a chain of new pages containing the keyword. However, the user needs to go through every page in order to figure out what are the main topics behind the changes. To conclude, conventional page tracker only tells us that some pages have been updated or some pages are new. At this point, we still lack of a tool that can track a particular area of user's interests, collect the changes in a certain time interval, process and generate a summary of the most discussed issue in the changes to the user from time to time. We refer this most discussed issue as the emerging topic in that information area.

### **6.3 System Architecture**

Figure 49 illustrates the system architecture of ETTS. ETTS consists of three main components: Area View System, Web Spider and Changes Summarizer. After taking in a keyword from the user, Area View System will direct the keyword to a crawling type commercial search engine and get the output

hits. Then, Area View System will analysis the output URLs from the commercial search engine and derive a number of domains that are mostly related to the keywords. These domains are grouped together to form an information area devoted to the keyword. Next, the Web Spider will dispatch to the Web to scan all the HTML files in these domains regularly, in order to collect all the modified and newly added HTML pages.

Later, the Changes Summarizer will extract all the changes (newly added sentences) from the collected HTML files by comparing the old and new databases. Then, the TF\*PDF (Term Frequency \* Proportional Document Frequency) algorithm will be used to count the weight of the terms in the changes. This algorithm is innovated in a way to give more weight to the terms that deem to explain the most discussed issues in the changes. Lastly, sentences with the highest average weight will be extracted to construct a summary for the user.



**Figure 49 : ETTS System Architecture**

### 6.3.1 Area View System

Area View System is designed to draw the fraction of information space on the Web that can represent the input keyword from user. The derived fraction is basically a group of domain sites most related or devoted to the keyword. In other words, this group of domain sites can be the optima

information fraction to represent the full coverage to the information area related to the keyword. Whenever there are popular topics concerning the keyword become available on the Web, the possibility of appearance of related information in this fraction will be high. So, this group of domain sites is the most suitable fraction on the Web where we should keep on tracking for the changes, in order to derive the emerging topic devoted to the keyword from time to time.

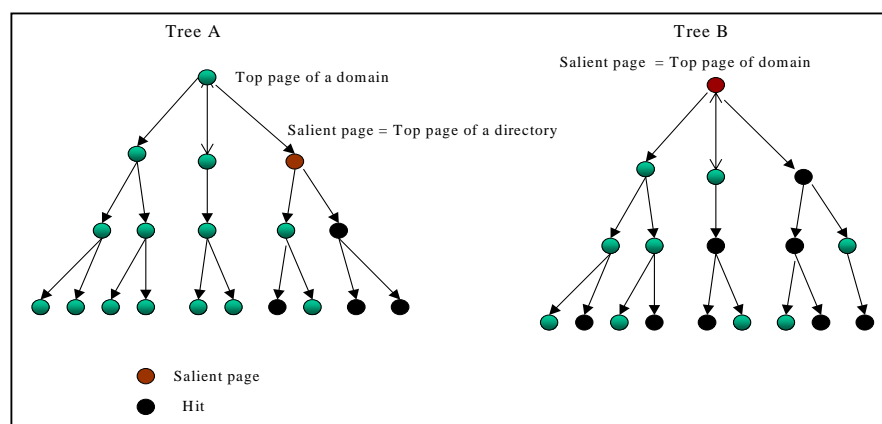
#### 6.3.1.1 Salient Domain Sites Retrieval

Although search engines are not suitable for use in tracking the changes to a particular information area, we need them to help us in addressing web pages with targeted information. Reader can accept this fact easily by looking at the number of hits returned by a search engine after keying in an arbitrary keyword. The huge hits number returned means that we have this lot of html pages containing the keyword, but this doesn't mean that the content of every page is relevant to the keyword. However, this is the most suitable source for us to recognize the salient domain sites devoted to the keyword.

#### 6.3.1.2 Salient Pages Derivation

We need the help of the search engine to identify the domain sites that are salient in representing a particular keyword. Firstly, Area View System will direct the keyword to a search engine and collect up to 500 hits or more. Each page of hits has a unique URL that consists of its domain URL, path, and its file name. For example, the page <http://www.cns.mii.edu/research/nuclear.html> has a domain URL of <http://www.cns.mii.edu/>, a path of [research/](#) and file name of [nuclear.html](#). Now, from the 500 hits, Area View System will further derive 50 salient pages with their domain URLs occur most frequently in the hits. Salient page is the top page of a domain site if the domain has its overall contents relevant to the keyword. But some of the domains have only a sub-directory devoted to the keyword. In this case, the salient page is the top page of the sub-directory. Area View System

determines this salient page either as the top page of a domain or the top page of a sub-directory in the domain by analyzing the shortest common path of the hits originated from the domain. If all the hits originated from a domain have a shortest common path, then the salient page is the top page of the sub-directory with the name of the path. The principles how Area View System determine the salient page is illustrated in Figure 50.



**Figure 50 : Salient Page Determination**

Figure 50 illustrates two different trees representing two domain sites. Each node represents a web page in the domain. In tree A, all the hits have a common path that is a top page of a sub-directory. In this case, the top page of the sub-directory is the salient page. While in Tree B, there is no shortest common path, so the salient page is the top page of the domain. Now, we can imagine that the combination of a salient page and all the pages under it shape an information cone (Figure 51) devoted to the keyword. Salient page is always at the tip of the information cone.

### 6.3.1.3 Salient Pages Verification

After determining the first 50 salient pages, Area View System will further do a more detail analysis on the information cones in order to identify the real information cones with high suitability. The

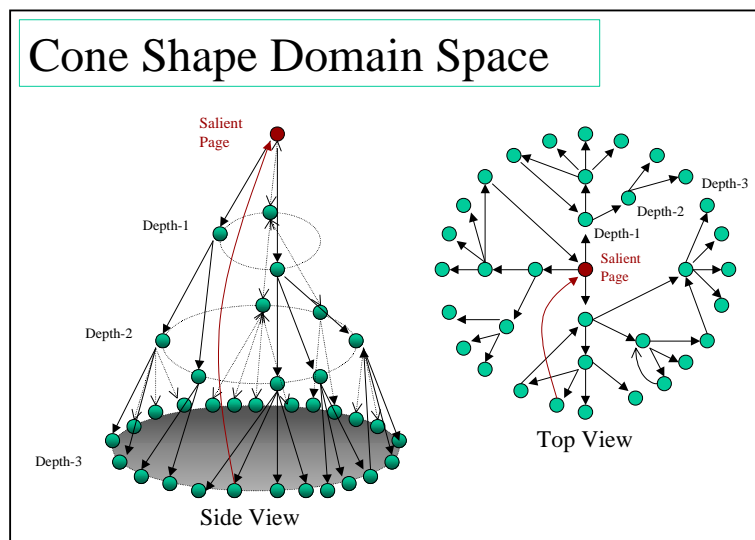
suitability of an information cone will be calculated by using the Suitability Equation shown below. Suitability of an information cone is equal to its File Ratio plus Link Ratio. The information cones with low suitability will be excluded from changes tracking.

$$\text{Suitability} = \frac{\text{total outer links pointing into other information cones}}{\text{total outer links}} + \frac{\text{total number of pages containing keyword}}{\text{total number of pages}}$$

All the information cones with suitability more than a certain trigger level will be added into the list of information cones used for tracking purpose. In other words, in the second stage of filtering, a suitability trigger level is used to reduce the 50 information cones to a number of information cones with high suitability level. These resulted information cones with high suitability level are the information cones with high percentage of their content related directly to the keywords, at the same time having strong linking relationship among each other. Thus, the suitability of an information cone is determined by two parameters, which are File Ratio and Link Ratio.

File Ratio of an information cone is equal to the ratio of number of file containing the keywords to the total number of file in the cone. Higher the value of this File Ratio, more likely will be the content of the information cone devoted to the keywords. On the other hand, Link Ratio of an information cone is equal to the ratio of number of link pointing into other information cones to the total number link in the cone. Higher the value of this Link Ratio, stronger will be the linkage of the information cone to the domain community devoted to the keywords. Finally, the formed domain community would be the information area where we would perform tracking mechanism for changes extraction.

With the salient page's information cones having suitability more than a certain trigger level, the information area of user interests is formed. This collection of information cones of real salient pages is the artificially structured fraction of the Web, which best for representing the keyword. This fraction of information space is believed to be homogenous and the cones are having strong linking relationship among each other.

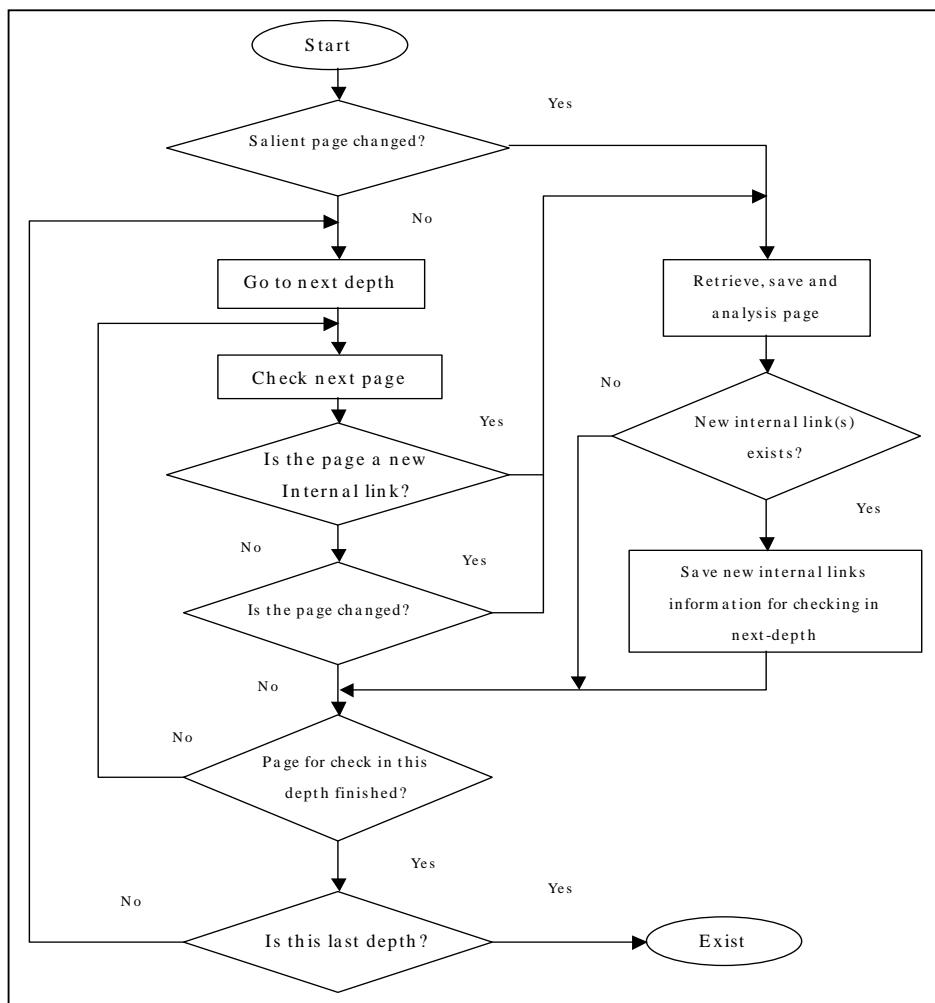


**Figure 51 : Domain Tree**

Search engine is just enough to provide us with a large number of hits related to the keywords. It can't tell which information cones are highly related to the keywords. Thus, Area View System is simply a search tool, which adapts its way for deriving the domain community highly devoted to the keywords.

### 6.3.2 Web Crawling

Web Spider is an autonomous robot that dispatches at a certain time interval to collect the desired html pages on the Web. The desired html pages are the updated and new pages in the information cones derived by Area View System. Basically, Web Spider adapts Breath-first search algorithm to traverse through the information cones for these pages. The traversing mechanism is illustrated in the flow chart in Figure 52.



**Figure 52 : Information Cone Traversing Flow Chart**

First, Web Spider will get the URL information of Salient page and all the pages in all depths under it. Next, Web Spider will detect the changed page one by one start from the salient page to the lowest depth level. Like most other tracking tools, Web Spider uses the HTTP HEAD command to check the Last-Modified field of a html page for changes. The changed page will be retrieved and saved in database. This retrieved page will be analyzed to check if there are new links pointing to new internal pages. New links' information will be saved in the next depth database. Later on during the next depth recursive checking, the new internal pages will also be retrieved, saved and analyzed to check if there are links pointing to other new internal pages. In this recursive way, Web Spider retrieves all the updated and new pages in a cone.

### **6.3.3 Change Summarizer**

Change Summarizer is designed to analyze the updated and new pages collected by the Web Spider, derive the changes and generate a summary of emerging topic from the changes. Change Summarizer consists of two major components: Change Detector and Summary Generator.

#### **6.3.3.1 Change Detector**

Change Detector is designed to retrieve the difference or change between the old pages and the newly collected pages. Change is defined as a collection of files with all the sentences exist in the new pages but not in the old pages. Change Detector has two component members to help it to accomplish its job. They are Sentence Parser and Sentence Retriever.

#### **6.3.3.2 Sentence Parser**

Basically, Hypertext Text Markup Language (HTML) file is a text file constructed with HTML marked-up tags and text content. HTML marked-up tags are generally used to format the outlook of the whole raw text and the style of some of the text. The raw text needs not to be containing only

complete sentences. It can be containing characters, terms and short phrases. Thus, html text file is merely a sequence of words and marked-up tags without proper structure. As a result, the html text file is chaotic and we are not able to retrieve the complete sentences from it directly by applying a sentence matching. Thus, before sentence retrieving, we need this Sentence Parser to do some preliminary jobs to wipe out all the tags and add a “new line” character at the appropriate points. So, the Html marked-up tags that format the outlook of the whole raw text or show the border of a sentence, for example <P>, <HR>, <H1>, </LI>, </TH>, </TR> and etc, will be substituted with a sentence delimiter. While all the text styling marked-up tags, for example <B>, <EM>, <A href=...>, <I> and etc will be substituted will a null character. The output will be a file of multiple lines of raw text. This raw text file will be outputted to Sentence Retriever.

#### 6.3.3.3 Sentence Retrieval

Raw text files of both the old and new version of updated pages will be inputted from Sentence Parser. Not every line of this raw text file will be containing full sentences. But at this point, every string of words in a line that starts with a capital letter and ended with a period sign (.) will be a correct sentence. So, Sentence Retriever will apply a sentence pattern matching line by line in order to extract all the complete sentences. All these sentences can be positioned in a table, list or paragraph in the original html text file. As a result, we get two sets of sentences from each pair of old and new version of updated page. Then, Sentence Retriever will compare these two sets of sentences and extract the new sentences that exist in the new version but not the old version. The extracted new sentences will be saved as the original file name respectively with its own extension. The collection of files of these new sentences is the “change” defined previously. All the files with sentences extracted from a new html file are automatically included in the change too. No sentences comparing needed to be done in this case. By the end, change is instead a collection of files of new sentences from all the domain information cones under tracking.

#### 6.3.3.4 Summary Generator

Summary Generator is designed to generate a summary of emerging topic from the change. Emerging topic will be the new topic that is discussed most frequently in almost all of the domains under tracking. Summary Generator consists of two components: TF\*PDF Counter and Sentence Picker.

#### 6.3.3.5 TF\*PDF Counter

TF\*PDF Counter is to count the weight of the terms in the change with the TF\*PDF (Term Frequency \* Proportional Document Frequency) algorithm. 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 domain, provided that the term deems to explain the hot topic discussed generally in the majority domains. This most discussed hot topic in the changes to majority domains is, in another words the emerging topic in the represented information area. Thus, the terms that deem to explain the emerging topic will gain a high weight. Also, larger the number of domains, more accurate will be this algorithm in recognizing the terms that explain the emerging topic. The second and third compositions combined to give the weight of a term in a domain. The second composition is the normalized term frequency of a term in a domain as showed in (Equation 2). The term frequency needs to be normalized because when different domain has a different size of changes, term from a domain with more changes has a proportionally higher probability that it will occur more frequently. But we want to give equal importance or equal weighting to the same term from each domain, so normalization should be done. The third composition is the proportional document frequency of a term in a domain. It is the exponential of the number of documents that contain the term to the total number of documents in a domain. 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 domain would be the term that

deems to explain the main topic behind the changes to a domain. To conclude, TF\*PDF algorithm give weight to the terms that explain the common hot topic or emerging topic in the changes to majority domains.

#### 6.3.3.6 Sentence Picker

In the final stage, Sentences Picker is to calculate the average weight of each sentence in the change and select the most suitable sentences to form a summary of emerging topic. Here, the most suitable sentences are basically the sentences with highest average weight containing highly weighted terms that deem to explain the emerging topic.

### 6.4 Samples Run

#### 6.4.1 First Sample (Oct 1 to Dec 15, 2003 on the keywords “Asia Economy”)

A keyword of “Asia economy” was used in this sample. A total of 6 information cones (Table 29) were evaluated to form the information area of “Asia economy”. Whenever new topics regarding Asia economy emerge on the Web, it is expected that the related information would appear in this portion of the Web too. A file ratio ranges from 0.46 to 0.99 of these information cones means that 46% to 99% of the files from these information cones contained either the keyword “Asia” or “economy”. Therefore, it is anticipated that whenever new information come into these cones, the probability for these new information to be related to “Asia economy” is almost more than 0.5.

New information in these cones has been gathered from October 1 to December 15, 2003 in a two weeks interval. Change has been processed on Oct 15, Oct 31, Nov 15, Nov 30 and Dec 15. Table 30 displays the top 20 TF\*PDF terms happened in the change in these time intervals. The high weights of some terms reveal to us that there were many topics regarding Apec meeting held in Thailand in Oct, China’s economy development, international trade issues, etc. Table 31, 32, 33, 34 and 35

present the top sentences (high average weight) extracted from the change in their time interval respectively.

Apec meeting is one of the most important economy events in Asia region. It summarizes the year's economic and international trade relation development, analyzes the short-term economy prospects and forecasts the future outlook. There was much information related to Apec in the change during the time interval from Oct 16 to Oct 30 (Table 32), which was the time Apec meeting was held. Thus, the term Apec gains its highest TF\*PDF weight in this period. Even for the weeks after the Apec meeting, much information reporting on the successes, findings, conclusions and future economic co-operation of Apec appeared on the Web. This is the reason why the term Apec gained high TF\*PDF weight continuously. Besides, lots of the time, China was at the center of the disputable topics, because its astonishing economy growth has created much trades frictions with countries such as US and Japan, at the same time emerged as a positive force for economic stability and progress in Asia and the world at large.

**Table 29 : Salient Pages in First Sample Experiment (keywords: "Asia economy")**

Salient Pages	Name	Suitability	
		File Ratio	Link Ratio
<a href="http://www.atimes.com/atimes/Asian_Economy.html">http://www.atimes.com/atimes/Asian_Economy.html</a>	Asia Times - Asian Economics	0.99	0.004
<a href="http://www.apecsec.org.sg/">http://www.apecsec.org.sg/</a>	Asia-Pacific Economic Cooperation	0.47	0.0001
<a href="http://www.business-in-asia.com/">http://www.business-in-asia.com/</a>	Business in Asia	0.46	0.003
<a href="http://www.asiabusinessstoday.org/">http://www.asiabusinessstoday.org/</a>	Asia Business Today	0.98	0.083
<a href="http://en.ce.cn/">http://en.ce.cn/</a>	China Economy Net	0.47	0.001
<a href="http://www.asiaobserver.com/economy.htm">http://www.asiaobserver.com/economy.htm</a>	Asia Observer - Economy	0.51	0.011

**Table 30: Top 20 TF\*PDF Terms in two weeks interval change from Oct 1 to Dec 15, 2003**

<b>Oct 15</b>	<b>Oct 31</b>	<b>Nov 15</b>	<b>Nov 30</b>	<b>Dec 15</b>
<b>Term Weight</b>	<b>Term Weight</b>	<b>Term Weight</b>	<b>Term Weight</b>	<b>Term Weight</b>
Apec 1.835	Apec 2.226	apec 1.409	apec 2.089	apec 1.816
china 1.372	asia 1.712	china 1.262	china 1.072	china 1.509
japan 1.015	china 1.429	economic 1.127	security 0.929	president 0.897
economic 1.005	economic 1.019	minister 1.007	asia 0.802	region 0.88
asia 0.935	world 0.787	asia 0.937	government 0.728	people 0.734
government 0.801	government 0.771	government 0.889	people 0.691	asia 0.715
trade 0.639	people 0.630	prime 0.762	economic 0.56	economic 0.708
year 0.565	trade 0.610	trade 0.731	year 0.513	trade 0.696
chinese 0.551	thailand 0.539	international 0.629	international 0.476	copyright 0.601
percent 0.529	market 0.532	year 0.601	thailand 0.470	south 0.556
international 0.526	year 0.514	political 0.548	development 0.433	government 0.522
development 0.515	development 0.505	japan 0.516	chinese 0.429	regional 0.496
world 0.507	country 0.476	people 0.507	trade 0.426	political 0.489
people 0.505	foreign 0.464	local 0.504	wake 0.426	million 0.473
social 0.474	president 0.461	market 0.479	company 0.424	year 0.472
population 0.429	return 0.456	chinese 0.476	stricken 0.423	house 0.469
country 0.410	chinese 0.456	report 0.453	idol 0.422	chinese 0.467
foreign 0.410	industry 0.455	thai 0.441	fancy 0.422	asian 0.455
security 0.408	minister 0.454	asean 0.435	swim 0.421	korean 0.444
public 0.388	international 0.413	region 0.417	smell 0.421	hyun 0.423

**Table 31: Sentences Extracted from change (Oct 1~Oct 15, 2003)**

<b>Top Sentences</b>	<b>Weight</b>
In the first nine months, China-Japan trade hit 95.97 billion US dollars, up 31.7 percent year-on-year; China-US trade, 90.98 billion US dollars, up 29.9 percent; and China-EU trade, 89.06 billion US dollars, up 41.5 percent.	0.706
Wu said that China and the U.S. have a complementary economic structure, and the obstacles that hindered Sino-US economic and trade relations had been removed after China entered the World Trade Organization.	0.6615
The 2003 APEC Economic Outlook summarises recent developments and analyses the short-term prospects of all 21 APEC Member Economies.	0.659
Asia's economic outlook would, over time, become less vulnerable to economic swings in industrial countries.	0.473

Japan and China together account for more than 50 percent of global foreign-currency reserves, and by some lights the yen and the yuan should be allowed to find their own level.	0.454
---	-------

**Table 32: Sentences Extracted from change (Oct 16~Oct 31, 2003)**

<b>Top Sentences</b>	<b>Weight</b>
The CSOM meeting is the first in a series of APEC Meetings in Bangkok, Thailand, in conjunction with the APEC Ministers Meeting on October 17-18 and the APEC Leaders' Meeting on October 20 -21.	1.067
At the meeting APEC Ministers reviewed the achievements of the 2003 APEC year hosted by the Kingdom of Thailand and agreed on initiatives for the 2004 APEC year to be hosted by the Republic of Chile.	1.037
Japan was feeling restless on the speed with which China was concluding FTA agreements China with the South Asian countries and it wanted to compete with China in this respect.	0.719
Trade has transformed China into a huge market of which the US is the biggest beneficiary because the US produces those goods that China needs most, such as aircraft.	0.615
The overwhelming emphasis on security matters at a gathering meant for trade marks the continuation of a trend that first occurred at the 2001 APEC forum hosted by China, followed by last year's meeting in Mexico.	0.613
**Do I need to export my Product?.Is my product or service needed in Asia? .Where should I start my efforts to export in Asia?.	0.561

Note: \*\* Is an example sentence without meaningful content but ranked high in weight

**Table 33: Sentences Extracted from change (Nov 1~Nov 15, 2003)**

<b>Top Sentences</b>	<b>Weight</b>
The tension over trade relations between China and the United States has ratcheted up this year as the latter's trade deficit with China hit a new record.	0.658
The first chapter of the 2003 APEC Economic Outlook reviews and analyzes recent developments in, and future prospects of, APEC economies and the global economy.	0.610
China's government leaders have to keep the financial system growing fast to take care of hundreds of millions of unemployed, but doing so is creating huge economic distortions.	0.532
However, one fact those critics of China's trade policy in the United States have so far overlooked, if not ignored, is that more than half of China's exports come from foreign-funded firms today.	0.528
Meanwhile, China continues to construct the so-called rings of "political friendliness", "economic cooperation" and "military exchange" around its periphery.	0.511

**Table 34: Sentences Extracted from change (Nov 16~Nov 30, 2003)**

<b>Top Sentences</b>	<b>Weight</b>
One of the great achievements of the 2003 APEC Year has been the signing of a Memorandum of Understanding, or MOU, on Cooperation among APEC Financial Institutions dealing with SMEs and micro-enterprises.	0.649
We feel this contribution from APEC is important because after all, APEC economies make up 47 percent share of world trade.	0.607
The report has been prepared by the APEC International Centre for Sustainable Tourism in collaboration with universities, governments and tourism operators in the APEC Region.	0.608
China's central government has attached great importance to revitalizing northeast China's economy since last year.	0.425
China's gigantic trade surplus with the US thus masks the fact that overall, China's imports and exports were nearly in balance.	0.397

**Table 35: Sentences Extracted from change (Dec 1~Dec 15, 2003)**

<b>Top Sentences</b>	<b>Weight</b>
APEC economies will also focus on WTO negotiations, security in the region and strengthening APEC to make it more efficient and responsive to all stakeholders in 2004.	0.738
The Chinese premier said accelerated economic growth in China would provide new opportunities and give further impetus to the growth of China-US relations.	0.580
At the APEC meeting, Chinese President Hu Jintao encourage countries to maintain the stability of the region and to counter efforts by the US to increase economic ties with non-Asian members of APEC such as Australia.	0.558
Losing Face: South Korean President Roh Moo Hyun vowed to clean up politics, but his own house is getting covered in grime.	0.502
At the same time, however, China's economic value to Japan is clear - China's imports of Japanese products have in no small way helped the country's lackluster economy.	0.481

#### 6.4.2 Second Experiment Sample

A keyword of “nuclear weapons” was used. There were 22 salient pages and thus 22 information cones derived to represent the information area of “nuclear weapons” on the Web. Change happened during the time interval between Apr 23, 2000 and Apr 30, 2000 was collected. The size of change (new sentences) was calculated at 3.61 Megabytes. Table 36 shows the experiment model. In the first column are the URLs of the salient pages, and the names of the respective domains are recorded in the second column. Third column shows the size of each information cone on Apr 23, 2000 while

the forth column shows the changes happened to each domain respectively. Fifth and sixth columns show the content page ratio and outer link ratio of each information cone respectively. The suitability of cones, sum of file ratio and external link ratio, ranges from the minimum 0.512 to maximum 1.332. This is a relatively high suitability value. This reveals that these 22 domains grouped together to form a strongly related information area of “nuclear weapon”.

The weight of the terms in the changes was counted by TF\*PDF algorithm. Table 37 shows the 30 most weighted terms in the changes. Result Summary consists of three sentences with highest average weight extracted by Sentence Picker as in Table 38. The highlighted terms in the sentences are the terms that appear in top position in the list of 30 most weighted terms.

In the resultant summary (Table 38), the first sentence contains nine terms (highlighted) from the top 20 most weighted terms. This sentence has the highest average weight of 3.151 units. This sentence tells that United States of America is about to deploy a national missile defense system. The second sentence tells that Russia objects to this deployment since it is again the ABM (Anti Ballistic Missile) treaty signed between United States of American and Russia 30 years ago. In the third sentence, there are dangling anaphor that make the sentence unclear because it don't tell who are the two nuclear weapon states and potential enemy states. But if we are aware of the international military movements, we should know that the two largest nuclear weapon states are United States of American and Russia; whereas one of the emphasized potential enemy states is North Korea. This is because North Korea has the capability to penetrate long ranges missile with nuclear warhead to United States of American.

**Table 36 : Second Experiment Salient Pages (keywords: “nuclear weapon”)**

Salient Page	Name	Size*	Changes**	Suitability	
				File Ratio	Link Ratio
<a href="http://www.acronym.org.uk/">http://www.acronym.org.uk/</a>	The Acronym Institute	14.8M	38.7K	0.856	0.256
<a href="http://www.ananuclear.org/">http://www.ananuclear.org/</a>	Alliance for Nuclear Accountability	203K	0	1.000	0.000
<a href="http://www.armscontrol.org/">http://www.armscontrol.org/</a>	The Arms Control Association – Homepage	8.43M	0	0.767	0.018
<a href="http://www.basicint.org/">http://www.basicint.org/</a>	BASIC	12.1M	419K	0.825	0.120
<a href="http://www.bullatomsci.org/">http://www.bullatomsci.org/</a>	Bulletin of the Atomic Scientists	22.6M	157K	0.982	0.056
<a href="http://www.ccnr.org/">http://www.ccnr.org/</a>	The Canadian Coalition for Nuclear Responsibility	10.0M	1.21M	0.643	0.012
<a href="http://www.ceip.org/programs/npp/">http://www.ceip.org/programs/npp/</a>	Carnegie Endowment - Non- Proliferation Project	9.39M	338K	0.520	0.073
<a href="http://www.cfsc.dnd.ca/link/peace/">http://www.cfsc.dnd.ca/link/peace/</a>	Peace, disarmament and arms control	217K	0	0.457	0.086
<a href="http://www.clw.org/coalition/">http://www.clw.org/coalition/</a>	Coalition to Reduce Nuclear Dangers - Working to Lower the Threat of Nuclear Weapons	10.9M	219K	0.902	0.018
<a href="http://www.cns.mii.edu/">http://www.cns.mii.edu/</a>	Welcome to the CNS Website	16.5M	226K	0.605	0.088
<a href="http://www.dtra.mil/nuclear/">http://www.dtra.mil/nuclear/</a>	DTRA - Nuclear Support	203K	0	0.909	0.000
<a href="http://www.fas.org/nuke/">http://www.fas.org/nuke/</a>	Nuclear Resources	69.6M	911K	0.583	0.036
<a href="http://www.hookele.com/abolition2000/">http://www.hookele.com/abolition2000/</a>	Abolition 2000 – GLOBAL NETWORK TO ELIMINATE NUCLEAR WEAPONS	1.71M	187 byte	0.696	0.038
<a href="http://www.igc.org/disarm/">http://www.igc.org/disarm/</a>	NGO Committee on Disarmament	4.06M	1.68K	0.973	0.084
<a href="http://www.ippnw.org/">http://www.ippnw.org/</a>	IPPNW – International Physicians for the Prevention of Nuclear War	834K	0	0.622	0.000
<a href="http://www.napf.org/">http://www.napf.org/</a>	Home Page of Nuclear Age Peace Foundation	8.64M	1.74K	0.952	0.370
<a href="http://www.nci.org/">http://www.nci.org/</a>	Nuclear Control Institute (NCI), Washington D.C.	8.26M	0	0.817	0.001
<a href="http://www.nuclearfiles.org/">http://www.nuclearfiles.org/</a>	The Nuclear Files Experiencing ethical and political challenges of the nuclear age.	26.0M	52.4K	0.994	0.225
<a href="http://www.nukefix.org/">http://www.nukefix.org/</a>	Nuclear weapon research on the Internet	1.62M	0	0.816	0.042
<a href="http://www.stimson.org/policy/">http://www.stimson.org/policy/</a>	The Committee on Nuclear Policy	1.22M	0	0.988	0.063
<a href="http://www.un.org/Depts/dda/">http://www.un.org/Depts/dda/</a>	United Nation- Disarmament	1.86M	224 bytes	0.512	0.000
<a href="http://www.wagingpeace.org/">http://www.wagingpeace.org/</a>	Home Page of Nuclear Age Peace Foundation and Abolition 2000	8.50M	31.8K	0.952	0.186

\*Size on Apr 23, 2000 \*\* Changes from Apr 23, 2000 to Apr 30, 2000

**Table 37 : TF\*PDF Term Weights (keywords: “nuclear weapon”)**

TERM	WEIGHT	TERM	WEIGHT	TERM	WEIGHT
nuclear	29.002	disarmament	3.364	world	2.400
weapons	11.598	2000	3.356	national	2.351
states	9.726	defense	2.919	power	2.349
treaty	8.315	review	2.735	like	2.288
conference	4.964	u.n.	2.68	war	2.237
united	4.762	npt	2.572	russian	2.216
missile	4.371	u.s.	2.559	plutonium	2.114
international	4.103	arms	2.518	use	1.959
peace	3.699	security	2.494	Fuel	1.938
new	3.526	russia	2.411	global	1.911

Resultant summary reveals to us that American peoples were planning to build a national missile defense system that can intercept the incoming missiles. There are pros and cons in American people. However, consistent with the result of our experiment, the CNN news article (appeared after our experiments) in Figure 53 and 54 stated that administration of U.S. says they will go ahead with missile program without Russian approval, and later says they will speed up the missile tests.

**Table 38 : Resultant Summary for the keywords of “nuclear weapon”**

As <b>world</b> leaders gather for the <b>2000</b> Non-Proliferation <b>Treaty</b> Review <b>Conference</b> at the <b>United Nations</b> , the <b>United States</b> is on the verge of deploying a National <b>Missile Defense</b> system.
If <b>Russia</b> objects to the <b>United States</b> defending itself against the offensive efforts of other <b>states</b> that were not even conceivable threats when the <b>ABM Treaty</b> was signed nearly 30 years ago, then the <b>United States</b> must make it clear that it is no longer bound by the <b>ABM Treaty</b> .
Leaders of both the nuclear weapon <b>states</b> and potential enemy <b>states</b> know these facts and know that the <b>United States</b> , in response to a <b>missile</b> attack, could wipe out their regimes, if not their countries.

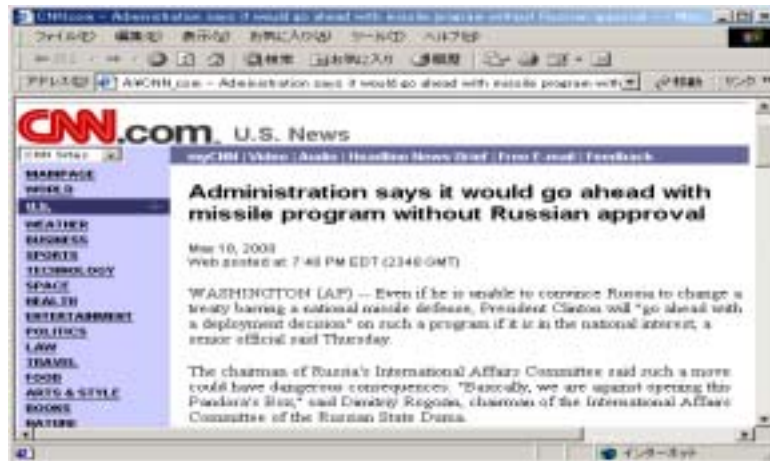


Figure 53 : CNN News May 18, 2000



Figure 54 : CNN News July 13, 2001

### 6.4.3 Third Experiment Sample

A keyword of “e-commerce” was used. There were 20 salient pages derived by Area View System with the help of the commercial search engine Google. Changes happened during the time interval between Oct 3, 2000, Nov 3 and Dec 4, 2000 was collected. Table 39 shows the experiment data. In the first column are the URLs of the salient pages, and the names of the respective domains are

recorded in the second column. Third and fourth columns show the Content Page Ratio and Outer Link Ratio of each information cone respectively.

**Table 39 : Third Experiment Salient Page (keyword: “e-commerce”)**

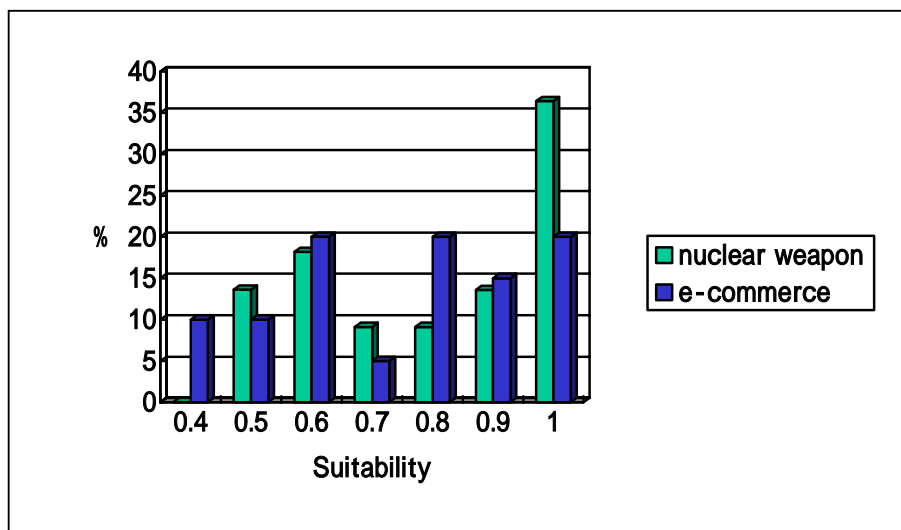
Salient Page	Name	Suitability	
		File Ratio	Link Ratio
<a href="http://ecommerce.internet.com/">http://ecommerce.internet.com/</a>	Electronic Commerce Guide	0.894	0.022
<a href="http://www.commerce.net/">http://www.commerce.net/</a>	Commerce Net	0.643	0.006
<a href="http://www.ecominfocenter.com/">http://www.ecominfocenter.com/</a>	eCommerce Info Center – One Stop for eCommerce Info, Services, products and technologies	0.796	0.005
<a href="http://www.goodexperience.com/">http://www.goodexperience.com/</a>	Goodexperience.com	0.666	0.003
<a href="http://www.anu.edu.au/people/Roger.Clarke/EC/">http://www.anu.edu.au/people/Roger.Clarke/EC/</a>	Roger Clarke’s Electronic Commerce	1	0.013
<a href="http://www.emarketer.com/">http://www.emarketer.com/</a>	eMarketer – the world’s leading provider of internet statistics	0.996	0.004
<a href="http://cism.bus.utexas.edu/">http://cism.bus.utexas.edu/</a>	Center for Research in Electronic Commerce, UT Austin	0.575	0.003
<a href="http://ec.fed.gov/">http://ec.fed.gov/</a>	Electronic Commerce Home Page	0.475	0.002
<a href="http://special.northernlight.com/e-commerce/">http://special.northernlight.com/e-commerce/</a>	Northern Light Special Edition : Electronic Commerce	1	0.041

<a href="http://ecom.das.state.or.us/">http://ecom.das.state.or.us/</a>	Oregon Center for Electronic Commerce & Government	1	0.013
<a href="http://www.becrc.org/">http://www.becrc.org/</a>	Electronic Commerce Resource Center (ECRC), Bremerton WA	0.801	0.020
<a href="http://www.ecommercetimes.com/">http://www.ecommercetimes.com/</a>	E-Commerce Times: the E-Business and Technology Super Site	0.997	0.002
<a href="http://www.cio.com/forums/ec/">http://www.cio.com/forums/ec/</a>	E-Business Research Center - Electronic Commerce Research Center	0.5	0.008
<a href="http://www.cptech.org/ecom/">http://www.cptech.org/ecom/</a>	CPT's Page on Electronic Commerce	0.681	0.016
<a href="http://www.diffuse.org/">http://www.diffuse.org/</a>	Diffuse – Home Page	0.993	0.003
<a href="http://www.ec2.edu/dccenter/ecommerce/">http://www.ec2.edu/dccenter/ecommerce/</a>	EC2@USC - Digital Commerce Center – Electronic Center	0.723	0.017
<a href="http://www.ecommercecommission.org/">http://www.ecommercecommission.org/</a>	Advisory Commission on Electronic Commerce	0.827	0.001
<a href="http://www.ecomworld.com/">http://www.ecomworld.com/</a>	Electronic Commerce World	0.605	0.001
<a href="http://www.ecrc.uofs.edu/">http://www.ecrc.uofs.edu/</a>	Scraton ECRC	0.431	0.002
<a href="http://www.epic.org/">http://www.epic.org/</a>	Electronic Privacy Information Center	0.883	0.010

The suitability of used information cones range from 0.433 to 1.041. The percentage of information cones with a certain suitability level is illustrated in Figure 55. Five percent, the lowest percentage of the information cones have a suitability ranges from 0.700 to 0.799.

The average suitability level is relatively lower than the average suitability from the second experiment sample with the keyword “nuclear weapon”, but the difference is not large. Thus, the combination of these 20 information cones can represent the information area of “e-commerce” on the Web well.

Table 40 shows the 30 highest TF\*PDF terms in the change from Oct 3 2000 to Nov 3 2000. Table 41 shows the 30 highest TF\*PDF terms in the Changes from Nov 3 2000 to Dec 4 2000. From the data, we found that there are 16 terms remaining in the top 30 most weighted terms. These terms are “Internet”, “online”, “information”, “click”, “Web”, “new”, “business”, “companies”, “customer”, “technology”, “e-commerce”, “use”, “customers”, “electronic”, “experience” and “site”. Among these terms, “Internet” gained and remained the term with highest term weight. This concurs with the fact that the Internet is the vital way in doing electronic commerce. From the data, another important point that we noticed is that the term “privacy” was not one of the terms in Table 40, but it appeared as one of the 10 most weighted terms in Table 41. This shows that privacy became one of the new important issues. The resulted 3 sentences with highest average weight are as in Table 42.



**Figure 55 : Percentage of Information Cones Vs Suitability**

**Table 40 : TF\*PDF Term Weights (period between Oct 3 and Nov 3, 2000)**

<b>TERM</b>	<b>WEIGHT</b>	<b>TERM</b>	<b>WEIGHT</b>	<b>TERM</b>	<b>WEIGHT</b>
Internet	2.859	business	1.212	looking	0.888
Web	2.093	Click	1.185	b2b	0.885
information	1.818	Topic	1.151	type	0.881
Online	1.73	customers	1.001	electronic	0.881
New	1.524	Terms	0.994	just	0.864
companies	1.493	logistics	0.94	word	0.85
e-commerce	1.42	XML	0.909	2000	0.835
Search	1.398	definition	0.905	letter	0.833
customer	1.238	Use	0.894	experience	0.824
Glossary	1.23	technology	0.891	site	0.804

**Table 41 : TF\*PDF Term Weights (period between Nov 3 and Dec 4, 2000)**

<b>TERM</b>	<b>WEIGHT</b>	<b>TERM</b>	<b>WEIGHT</b>	<b>TERM</b>	<b>WEIGHT</b>
Internet	2.927	Global	1.51	electronic	1.122
Online	2.835	technology	1.432	said	1.077
information	2.224	ecommerce	1.23	policy	1.045
Click	2.139	Services	1.197	users	1.033
Web	2	e-commerce	1.184	experience	1.015
New	1.782	Company	1.184	local	0.974
Business	1.772	Use	1.161	site	0.971
companies	1.583	customers	1.15	licensing	0.922
Privacy	1.568	Service	1.145	notices	0.912
Customer	1.52	Legal	1.132	permissions	0.9

**Table 42 :Resultant Summary for the keyword of “e-commerce”**

Top Sentences	Average Weight
Regardless of what your <b>company</b> is doing <b>online</b> -- <b>information technology</b> , content or <b>e-commerce</b> -- as the <b>Internet</b> changes so does your <b>business</b> .	1.136
No one, including the U.S. government, seems to believe that the government should force <b>Internet companies</b> to use <b>electronic</b> signatures for <b>Internet</b> transactions.	0.958
One of the leading <b>Web privacy</b> practices is the use of a <b>Web site privacy policy</b> to explain what a <b>company</b> does with personal <b>information</b> gathered on the <b>site</b> .	0.957

From the result summary, we can see that the sentence average weight is relatively lower than the sentence average weight in second experiment. The highlighted terms in Table 42 are among the 30 most weighted terms. In the first sentence, it tells that the Internet changes any kind of business doing online, which is electronic commerce. In the second sentence, it tells that U.S. government is unlikely to force electronic signatures implementation in Internet business transactions. And the third sentence concerns Web privacy practices.

In a CNN web page (April 17, 2001) of Figure 56, it was reported that more than 60 federal Web sites violates U.S. privacy rules by using unauthorized software to track the browsing and buying habits of Internet users. While in Figure 57, it tells that because of under pressure to protect privacy better, advertising industry has set up two new Web sites that let computer users refuse to have their personal data collected and profiled when they visit popular commercial Web sites. These two figures with the news emerged few months after the experiment done, agree with the experiment results that privacy would be a popular issue or an emerging topic discussed widely.



Figure 56 : CNN News April 17 2001



Figure 57 : CNN News May 25 2001

## 6.5 Comparison to Related Works

We use a set of information cones to form the information area of a keyword. Our keyword information area can be interpreted as web community in other link-based research for identifying collections of related pages such as the PageRank algorithm [brin98], the HITS algorithm [kleinberg98], bipartite sub-graph identification [kumar99] and focused crawling [chakra99]. Besides, [sacco00] describes a way to identify to topic relevant portions of a hierarchical space, while [terveen

99] gives a methodology to derive the sites that pertain to a given topic. Therefore, our information area is unique in a way that it is a set of information cones that would accommodate all the new information related to the keyword into it. Instead of a community of members with high precision but small like HITS, we need to build an information space that will trap all the related new information. The precision in the sense of information relevancy at the first stage is not the highest priority because later in the next stage, we will filter out the unwanted information and present the topic terms for generating a topic summary.

While using conventional commercial page tracker [changedetect, webspector] to track for changes on HTML pages, it can be annoying if the users are always pushed with acknowledge emails although the changes is trivial. In order to solve this problem, WebBeholder [santi98] allows user to set a trigger threshold they prefer. Here, if and only if the total changes score is greater than the trigger level, the system will be triggered to send e-mail to the user. Yet there is always no appropriate trigger threshold can be defined accurately since there are many possible combinations of changes in an HTML page (title, header, content character, color, text style and etc.) with different score. As a result, the users might still be pushed with e-mail although the change is not interesting to them. Vice versa, the users can miss some important changes. On the other hand, AT&T Internet Difference Engine (AIDE) [dougkis98] and TopBlend [chen00] have been trying to create a “merged” page to show the differences with special HTML markups. However, the “merged” page can be very messy and raise the danger of creating syntactically or semantically incorrect HTML pages.

ChangeDetector [changedetector] is a tool that is purposed to monitor the changes on entire web sites. It can tell if the structure of an organization has changed, instead of acknowledging some simple changes happen on certain web pages. It relies on machine learning techniques and intelligent crawling in collecting pages in some huge web sites. Its prototype system could monitor more than 2000 web sites in a week.

Thus far, we have gone through a number of concurrent tracking tools and some of their goods and deficiencies. In general, these systems will only inform us with the URL of the new and changed pages. ChangeDetector can do more by informing the changes embedded in a web site. However, we still lack of a kind of system like ETTS, which can automatically process the changes and conclude the main topics (emerging topics) in a particular information area on the Web.

## **6.6 Conclusion**

This chapter has presented ETTS, which is an intelligent software application for detecting and tracking the emerging topic (hot topic) from a particular information area on the Web. Since the Web is open and dynamic, contents in any information area is changing dynamically and it could be anticipated that there will be some hot issues being discussed in any information area at any time. Thus, we can take the assumption that web pages or articles regarding the hot issues will be posted dynamically on the information areas on the Web. All these newly added information are defined as changes to the information area. The system that we propose, ETTS (Emerging Topic Tracking System), is to retrieve the changes happening in the information area of user interest, and further summarize an emerging topic from the changes. ETTS consists of three main components. They are Area View System, Web Spider and Changes Summarizer.

For each user's input keyword of interested information area, Area View System will derive and analysis a group of web domains, in order to gain a set of web domains that can represent that information area in perfect. Area View System calculates the suitability of the qualified domains by analyzing the content page ratio and the outer ink ratio of each domain. From the experiment results, we found that the approach adapted by Area View System is satisfactory and useful in recognizing or matching a specific information area on World Wide Web. After recognizing an information area on the Web with Area View System, Web Spider functions as an autonomous software robot that

dispatches regularly at a fix time intervals to collect information change from that information area. Web Spider scans through all the domains using Breath First Search algorithm. Newly posted documents or html pages will be collected and saved in the database of ETTS. Therefore, Web Spider is the new information (change) collector and important in ETTS. It makes sure all the change happened to an information area will be retrieved for further analyzing by Changes Summarizer.

Changes Summarizer applies TF\*PDF to judge the terms that reveal an emerging topic, from a few domain cones that representing the topic information area on the Web. The experiments done have proved to us that TF\*PDF algorithm works out the way we expected and produces good results. Lastly, after developing ETTS and evaluating some performance issues, we conclude that this system has explored a totally new and innovated approaches, it can derive a topic information area on the Web and extract the emerging topics from the topic information area.

## Chapter 7 Conclusion

The objective of this research is to invent the state-of-art technology capable of capturing and analyzing the changes on the Web. In other words, the goal of this research is to address the problem of detecting and tracking the popular topics in the changes on the information-wealth and dynamically changing Web.

The Web has emerged as the most important channel for information dissemination, exchanges, sharing and gathering, and thus make the Web itself to have become the biggest networked information storage in the world. Nevertheless, the high growth of Web doesn't stop and the contained information keeps on proliferating. Hence, huge amount of new information or changes are being posted on the Web dynamically. In order to stay competent in this new information age, these new information is vital and needed to be delivered to us in time. However, browsing the Web manually for changes is inefficient and unrealistic. Thus, intelligent information system is essential for gathering, analyzing and delivering the useful information in the change to us. Therefore, our research is to provide a framework for detecting and tracking the change on the Web. This system is proposed as an approach towards the automatic online journalism of new information (change) on the Web. These information changes are classified into two types: "Flow" and "Stock".

"Flow" type information (i.e. news) come to the Web constantly and regularly, at a rather fast pace. This type of information puts up a big amount of textual data online daily and provides a challenge for the research of topic detection and tracking. This task is responsible to discover the evolving features and concepts of interesting topics from the mass of textual data by investigating its content, structure and distribution over time. Basically it takes as input a collection of temporal textual data and recognizes the topic trend in time series. Recent research has showed much advancement and attempts have been tested by using the technologies such as linguistic, statistical, learning, clustering

and etc. However, most of the efforts done works on precisely tagged (manually) corpus, which is not raw text. Besides, most of these works merely display some features (words, n-grams, document frequency) associated with each detected topic, which show little details about the story lines of the topic. Therefore, a summary of each topic will be more useful for the users to understand the flow of each topic. Lastly, most of them apply clustering techniques to aggregate topic documents and use the document cluster size to measure the significance of a topic, resulted in need of some computational complexities. Therefore, we are motivated to invent an efficient algorithm to do the topic detection and features extraction, and further generate a summary for each popular topic.

Therefore, for “Flow” type information, our goal is to address the measures for detecting and summarizing the important topics in news archive, given a range of news channels. Our topic detection algorithm TF\*PDF has a unique concept and efficient. It works on the idea that the topics being discussed in several channels concurrently are likely to be “hot” and important. Thus, this algorithm is designed in a way to give significant weight to the terms that explain the important topics in many documents in many news channels concurrently. Later, by exploiting the weight variances of these topic terms, we are getting to know the topic time frame by measuring the information “surprise” in the term weight. The terms of a popular topic should present an acceleration value during the rise of the topic and a deceleration value before the topic fades. Later, we will do sentences clustering on the important sentences appearing in the topic time frame, based on their sentence vector by using the extracted topic term as unit vector. The topic cluster will be the prime cluster generated. In this way, we can generate a better-coverage summary, which may cover the flows of the topic emerged in a certain time frame. Another way in reverse where we want to summarize all the main topics existing in a certain period (for example the pass week or month), similarly we will use our TF\*PDF algorithm to extract the topic terms for the week and then do clustering on the important sentences in the week. In this scenario, we will produce a range of sentences clusters each devoted to

a topic. Sentences in each cluster will be arranged chronologically to form a topic summary respectively. As a result, we could automatically generate a summary report of main topics to the users from time to time.

“Stock” type information, mainly the static web pages, change unpredictably doesn’t know at when and in what form. Therefore, there is a need for monitoring systems to check the pages or the part of the Web we are interested in frequently for reporting the changes to us in time. Realizing that the currently available changes monitoring systems are just doing a little more than reporting the URL of the page which has changed, regardless of the changes’ importance or trivialness, we propose ETTS (Emerging Topic Tracking System), which is an intelligent software application for detecting and tracking the emerging topic (hot topic) from a particular information area on the Web. Since the Web is open and dynamic, contents in any “information area” is changing dynamically; pages or articles regarding the “hot” issues will be posted dynamically in it. All the newly added information is defined as “changes” to the information area. The goal of ETTS is to retrieve the changes in the information area of user interest, and further summarize an emerging topic from the changes. ETTS consists of three main components: Area View System, Web Spider and Changes Summarizer. For each user’s input keyword of interested information area, Area View System will derive a group of web domains representing the information area. Then, Web Spider functions as an autonomous software robot will dispatch regularly to collect changes from this information area. Web Spider as a new information (changes) collector will makes sure all the changes happened to an information area will be retrieved and delivered to Changes Summarizer for further analyses. Later, Changes Summarizer will apply TF\*PDF algorithm to judge the terms that reveal an emerging topic and generate a summary of emerging topic from the important sentences containing the topic terms.

To conclude, our system aims to innovate the technology and use a new TF\*PDF algorithm to detect the prominent topics in the changes. In the framework and domain of problem addressed, this

algorithm is more superior than the conventional TF\*IDF algorithm in a way that it doesn't need retrospective corpus, besides posing minimal risk of losing the tracks of detection and tracking of popular topics. Also, our system requires less computational complexity while offering more flexibility. It crawls the Web, collects the changes and journalizes a summary of popular topics to the user. It does more than the conventional web tracking systems that just acknowledges the URLs of changed pages. It will become our personalized e-journalist on the Web and periodically provide us with the collection and e-publication of current new events. Our approach using TF\*PDF algorithm for weighting the topic terms and sentence clustering for topic summarizing is totally new, different from TF\*IDF approach that basically do document clustering and then multi-documents summarization for a topic summary. ETTS as an intelligent agent on the Web can detect the changes in an information area of our interests and generate a summary of changes back to us regularly. This summary of changes will be telling the latest most discussed issues and thus revealing the emerging topics in the particular information area. With this system, we will be "all time aware" of the latest trend in information space of your interest on the WWW.

## Bibliography

- [TBLee01] T. Berners-Lee, J. Hendler, O. Lassila: The Semantic Web, *Scientific American*, 284, 34-43 (2001)
- [boyapati02] Boyapati V., Chevrier K., Finkel A., Glance N., Pierce T., Stokton R., and Whitmer C. : ChangeDetector: A Site-Level Monitoring Tool for the WWW, In WWW2002, pp. 570-579 (2002)
- [chen00] Chen Y., Douglis F., Huang H. and Vo K.-P.: TopBlend: An Efficient Implementation of HtmlDiff in Java, In WebNet2000 (2000)
- [douglis98] Fred Douglis, Thomas Ball, Yih-Farn Chen and Eleftherios Koutsofios. The AT&T Internet Difference Engine (AIDE): Tracking and Viewing Changes on the Web, *World Wide Web Volume 1 Issue 1*, 1998. page 27-44.
- [hirschberg77] D.S. Hirschberg. Algorithms for the longest common subsequence problem, *Journal of the ACM*, 24(4):664--675, October 1977.
- [jacobson92] G. Jacobson and Kiem-Phong Vo. Heaviest Increasing/Common Subsequence Problems, *Proceedings of the 3rd Annual Symp. of Combinatorial Pattern Matching*, Vol. 64, Springer-Verlag, pp. 52-65, 1992.
- [changedetect] ChangeDetect, <http://changedetect.com>
- [chakra99] Soumen Chakrabarti and Martin van den Berg and Byron Dom, Focused crawling: a new approach to topic-specific {Web} resource discovery. *Computer Networks*, volume 31 NO.11-16 pp1623-1640 1999
- [michelan00] Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles and Marco Gori, Focused Crawling using Context Graphs, *26th International Conference on Very Large Databases, VLDB 2000, Cairo, Egypt 2000*
- [chakra02] S. Chakrabarti, K. Punera, M. Subramanyam, Accelerated focused crawling through online relevance feedback, *World Wide Web (ACM)*, Hawaii 2002
- [edward00] J. Edwards, K. McCurley and J. Tomlin, An Adaptive Model for Optimizing Performance of an Incremental Web Crawler, *Proceedings of the 10th International World Wide Web Conference*, Hong Kong, May 2001.
- [kosala00] R. Kosala and H. Blockeel. Web Mining Research: A survey. *ACM SIGKDD Explorations*, Vol. 1, No. 2, pp. 1-15, 2000.
- [brin98] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Proc. Of 7<sup>th</sup> World Wide Web Conference*, 1998
- [cowie96] J. Cowie and W. Lehnert. Information Extraction. *Communications of the ACM*, vol. 39(1): pp. 80-91, 1996
- [touchgraph] TouchGraph. <http://www.touchgraph.com>

[delphion] <http://www.delphion.com>

[martin97] A. Martin, T.K.G. Doddington, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance, Proceedings of EuroSpeech97, vol. 4, pp. 1895-1898, 1997.

[havre02] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing Thematic Changes in Large Document Collections, IEEE Transactions on Visualization and Computer Graphics, 8(1), Jan-Mar 2002

[swan00] R. Swan, D. Jensen. TimeMines: Constructing Timelines with Statistical Models of Word Usage, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, August 20-23, 2000

[fisher95] D. Fisher, S. Soderland, J. McCarty, F. Feng, and W. Lehnert. Description of the UMASS Systems as Used for MUC-6, Proc. of The 6th Message Understanding Conf., pp. 127-140, Nov 1995.

[pottenger01] W.M. Pottenger, Y. Kim, and D.D. Meling. HDDI: Hierarchical Distributed Dynamic Indexing, Data Mining for Scientific and Engineering Applications, Jul 2001

[tarjan72] R.E. Tarjan. Depth First Search and Linear Graph Algorithms, SIAM J. Computing, 1:146-160 1972

[brian97] Brian Lent and Rakesh Agrawal and Ramakrishanan Srikant. Discovering Trends in Text Databases, 3rd Int'l Conf. on Knowledge Discovery and Data Mining, California USA, 1997.

[srikant96] R. Srikant and R. Agrawal. Mining Sequential Patterns: Generalizations and performance improvements, 5th Int'l Conf. on Extending Database Technology (EDBT), Avignon, France, 1996.

[agrawal95] R. Agrawal, G. Psaila, E.L. Wimmers, and M. Zait. Querying shapes of histories, Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, 1995

[fien03] Fien De Meulder and Walter Daelemans. Memory-Based Named Entity Recognition using Unannotated Data, Proceedings of CoNLL-2003 Edmonton, Canada, pp. 208-211 2003

[borthwick98] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman . NYU: Description of the MENE Named Entity System as used in MUC-7, Proceedings from the Seventh Message Understanding Conference (MUC-7), Fairfax, VA. April 29 - May 1, 1998.

[jain88] A. Jain and R. Dubes. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, 1988

[kaufman90] L. Kaufman and P. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley and Sons, New York, 1990

[hartigan75] J. Hartigan. Clustering Algorithms, John Wiley and Sons, New York 1975

[barzilay02] Regina Barzilay, Noemie Elhadad and Kathy McKeown. Inferring Strategies for Sentence Ordering in Multidocument News Summarization, *Journal of Artificial Intelligence Research (JAIR)*, 2002, Vol. 17, pp 35-55.

[barzilay01] R. Barzilay, N. Eohadad, and K.R. McKeown. Sentence Ordering in Multidocument Summarization, *Proceedings of the 1st Human Language Technology Conference*, San Diego, California, 2001

[barzilay99] R. Barzilay and K. McKeown and M. Elhadad. Information Ffusion In The Context of Multi-Document Summarization, *Proceeding of ACL'99*, Maryland, June 1999

[vasileios00] H. Vasileios, G. Luis and M. Ankinedu. An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering, *Proc. of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*, 2000

[yang98] Yiming Yang, T. Pierce, and J. Carbonell. A study on retrospective and on-line event detection, *Proc. of SIGIR '98: 21st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, ACM Press, pp. 28-36, 1998

[allan01] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics, *Proceedings of SIGIR*, 2001

[essence] <http://www.newsinessence.com/>

[newsblaster] <http://www1.cs.columbia.edu/nlp/newsblaster/>

[radev98] Radev D.R., and McKeown K.R. Generating Natural Language Summaries from Multiple On-line Sources, *Computational Linguistics* 24(3): 469-500, 1998

[radev00] Radev D.R., Jing H. and Budzikowska M., Summarization of Multiple-Document: Clustering, Sentence Extraction and Evaluation, *Proceedings of The ANLP/NAACL Workshop on Summarization*, Seattle, WA., 2000

[radev00\_2] D. Radev and H. Jing and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies, *Proceedings of The ANLP/NAACL Workshop on Summarization*, Seattle, WA., 2000

[goldstein98] J. Carbonell and J. Goldstein. The use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, *Poster Session, SIGIR'98*, Melbourne, Australia 1998

[mani99\_3] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1--2):35-- 67, 1999

[allan98] J. Allan, R. Papka, and V. Lavrenko, Online New Event Detection and Tracking, *Proc. SIGIR 1998: 21<sup>st</sup> Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, ACM Press, New York, 1998, pp. 37-45

[yang98] Yiming Yang, T. Pierce, and J. Carbonell., A study on retrospective and on-line event detection. In proceedings of SIGIR '98: 21<sup>st</sup> Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, ACM Press, pp. 28-36, 1998

[yang99] Yiming Yang, Jaime G. Carbonell, et al., Learning Approaches for Detecting and Tracking News Events, IEEE Intelligent Systems, 1999, pp. 32-43.

[kbkhoo02] K.B. Khoo and M. Ishizuka. Topic Extraction from News Archive Using TF\*PDF Algorithm, Proceedings of 3rd Int'l Conference on Web Information Systems Engineering (WISE 2002) ,IEEE Computer Soc., pp.73-82, Singapore, Dec. 2002

[kbkhoo01] K.B. Khoo and M. Ishizuka, Emerging Topic Tracking System, Proc. of Web Intelligence (WI01), LNAI 2198 (Springer), pp. 125-130, Maebashi, Japan. 2001

[kbkhoo01\_2] K.B. Khoo and M. Ishizuka, Information Area Tracking and Changes Summarizing in WWW, Proc. of WebNet 2001, International Conf. on WWW and Internet, pp. 680-685, Orlando, Florida. 2001

[kbkhoo 01\_3] K.B. Khoo and M. Ishizuka, Emerging Topic Tracking System (pilot version), Proc. of 3rd Int'l Workshop on Advanced Issues on E-Commerce and Web-Based Information Systems (IEEE Computer Soc.), San Jose, California, USA, 2001.

[hayes97] P. Hayes, L. Knecht, and M. Cellio. A News Story Categorization System. Morgan Kaufmann Publishing, San Francisco, pp. 518-526, 1997. Originally appeared in Proceedings of the 2<sup>nd</sup> Conference on Applied Natural Language Processing 1988.

[masland92] B. Masland, G. Linoff, and D. Waltz, Classifying news stories using memory based reasoning, Proceedings of SIGIR '92, pp. 59-65, 1992

[dhara00] S. Dharanipragada, M. Franz, J.S. McCarley, K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Statistical Models for Topic Segmentation, SIGIR 2000

[lafferty99] J. Lafferty, D. Beeferman and A. Berger, Statistical Models for Text Segmentation, Machine Learning, special issue on Natural Language Learning, C. Cardie and R. Mooney eds., 34(1-3), pp. 177-210, 1999

[salton89] G. Salton and C. Buckley : Term-Weighting Approached in Automatic Text Retrieval, Information Processing and Management, Vol. 4, No. 5, pp. 513-523 (1989)

[salton89\_2] G. Salton. Automatic Text Processing. Addison Wesley Publishing Co., Massachusetts, 1989.

[mckeown95] R. K. McKeown and Dragomir R. Radev, Generating Summaries of Multiple News Articles, Proceedings 18th Annual International pp. 74-82 ACM SIGIR, Seattle, Washington 95

[salton97] G. Salton, A. Singhal, M. Mitra & C. Buckley, Automatic Text Structuring and Summarization, Information Processing and Management, Elsevier Science, 33, 193-207. 1997

[mani99] I. Mani and M. T. Maybury (eds): *Advances in Automatic Text Summarization*, MIT Press, 1999.

[jones97] K.S. Jones and P. Willett, *Readings in Information Retrieval*, Morgan Kaufmann Publishing, San Francisco, 1997. Chapter 4, pp. 167-256

[tague92] J. Tague-Sutcliffe, Measuring the informativeness of a retrieval process, *Proc. of SIGIR '92*, pp. 23-36, 1992.

[trec] <http://trec.nist.gov/>

[mani99\_2] Mani, I., T., House, D. Klein, G., Sundheim, B., and Hirschman, L., The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of EACL-99*, Bergen, Norway, June 1999

[allan99] Russell Swan and James Allan, Extracting Significant Time Varying Features from Text, *Proceeding of Eight Int'l Conference on Information Knowledge Management, CIKM'99*, pages 38-45, Kansas City, Missouri, November 1999. ACM

[allan00] James Allan, Victor Lavrenko, Daniela Malin, and Russell Swan, Detections, Bounds, and Timelines: UMass and TDT-3, *Topic Detection and Tracking Workshop (TDT-3)*, Vienna, Virginia, February 2000

[santi98] Santi Saeyor and Mitsuru Ishizuka: WebBeholder: A Revolution in Tracking and Viewing Changes on the Web by Agent Community, in *proceedings of WebNet98, 3<sup>rd</sup> World Conference on WWW and Internet*, Orlando, Florida, USA, Nov. 1998.

[cliff99] Cliff Pratt, "Searching the Web Using a 3-D Model", *Webnet Journal* April-June 1999, Vol.1, No.2.

[kleinberg98] Jon M. Kleinberg, Authoritative Sources in a Hyperlinked Environment, *Proc. of 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668-677, 1998.

[kumar99] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for Emerging Cyber-Communities, *Proc. 8th International WWW Conference*, 1999.

[sacco00] G.M. Sacco. Dynamic Taxonomies: A Model for Large Information Bases, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 12(3): pages 468-479, May/June 2000.

[terveen 99] L. Terveen, W. Hill, and B. Amento. Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources, *ACM Transactions on Computer-Human Interaction*, Vol. 6(1): pages 67-94, March 1999.

[webspector] WebSpector. <http://www.illumix.com/>

[salton71] G. Salton, *The SMART Retrieval System -- Experiments in automatic document processing*, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

## Publication Lists

### Published International Conference Papers:

1. K.B. Khoo and M. Ishizuka: "Topic Extraction from News Archive Using TF\*PDF Algorithm", Proc. of 3rd International Conference on Web Information Systems Engineering (WISE 2002), IEEE CS Press, pp. 73-82, Singapore. 2002
2. K.B. Khoo and M. Ishizuka: "Emerging Topic Tracking System", Proc. of Web Intelligent (WI 2001), LNAI 2198 (Springer), pp. 125-130, Maebashi, Japan. 2001
3. K.B. Khoo and Mitsuru Ishizuka: "Information Area Tracking and Changes Summarizing in WWW", Proc. World Conf. on WWW and Internet (WebNet 2001), pp.680-685, Orlando, Florida, USA. 2001
4. K.B. Khoo and M. Ishizuka: "Emerging Topic Tracking System (pilot version)", Proc. 3rd Int'l Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS-2001), IEEE CS Press, pp.2-11, San Jose, California, USA. 2001
5. Adam Jatowt, Khoo Khyou Bun and Mitsuru Ishizuka: "Query-based Discovering of Popular Changes in WWW" , Proc. IADIS Int'l Conf. on WWW/Internet (IADIS 2003), Algarve, Portugal, Vol.1, pp.477-484 (2003.11)

### Submitted International Journal Papers:

1. K.B. Khoo and M. Ishizuka, "Emerging Topic Tracking System in WWW", Knowledge-Based Systems, by ELSEVIER Publisher (submitted)
2. K.B Khoo, Adam Jatowt and M. Ishizuka, "Automatic News Digest: Detecting and Summarizing News Topic", International Journal of Computational Intelligence and Applications (IJCIA), by World Scientific Publisher (submitted)
3. K.B Khoo, Adam Jatowt and M. Ishizuka, "Generating a Better-Coverage Summary of News Topics using Time Features and Sentence Clustering", International Journal of Web Engineering and Technology (IJWET), by iEL (Inderscience Publisher) (submitted)

### Domestic Conference Papers:

1. K.B. Khoo, H. Dohi and M. Ishizuka, "Topic Extraction and Summarization in News Archive using TF\*PDF Algorithm and Sentence Vector Clustering", Forum on Information Technology (FIT 2002), pp. 303-304, Tokyo, Japan. 2002
2. K. B. Khoo and Mitsuru Ishizuka, "Emerging Topic Tracking System in World Wide Web", Proceedings of 61th Conference of Information Processing Society of Japan (IPSJ), Japan, October, 2000