Emerging Topic Tracking System

Khoo Khyou Bun+ Mitsuru Ishizuka
Dept. of Information and Communication Engineering
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
{kbkhoo,ishizuka}@miv.t.u-tokyo.ac.jp

Abstract

*We designed a system that track the changes to a particular area of user's interests on the Web and generate a summary of emerging topic back to the user. This system consists of three main components, which are 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 adapting a newly innovated algorithm TF*PDF (Term Frequency * Proportional Document Frequency). 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).*

1. Introduction and Related Work

If we are stuck with the conventional view of the Internet, then we are in trouble because its contents are changing too quickly. But user or professional would like to be always updated with the latest hot topic concerning the particular area of their interest on the Internet. However, due to the changes happen in a random way, updating ourselves by browsing through some particular Web sites of interest manually and regularly is both a difficult and time consuming job, yet there are no promises of any information changes have been taking place. Thus, here come the needs of this kind of intermediate system which can track and acknowledge us upon changes took place 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. And 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 [1] allows 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 which part of the page 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) [2] has 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 page.

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. Informant [3] claimed to be the ”best search monitoring tool” on the market. At there, 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, Informant 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 it. 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. [4] cited, ”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.”

Very similar to Informant, Netmind [5] is another broadly utilized tracking tool. Netmind provides different columns to get input keyword from user for tracking new information in different category. For each category, Netmind 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, Netmind will constantly check for the appearing of the keyword on the web site CBS Marketwatch. However, relying on only one or a few web sites as the source(s) to gain new information from a wide general area, for example health and medical, will doubt the completeness of the changes happened to a small sub area can be reported.

Thus far, we have gone through a number of concurrent tracking tools and some of their deficiencies. In general, 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 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.

2. Area View System

Area View System is designed to draw the fraction of information space on the Internet that can represent the particular 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 of the area related to the keyword. Whenever there is new information concerning the keyword become available on the Internet, the possibility that the information appear in this fraction will be very 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.

2.1. Search Engine as Salient Domain Sites Finder

Because of the complexity and heterogeneity of WWW, we need search engines to help us in finding Web pages with targeted information. But search engine is not suitable for use in tracking the changes to a particular information area. 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 hits number means that we have this lot of html pages containing the keyword but this doesn’t mean that 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.

2.2. Salient Pages Derivation Approach

Thus, we need the help of the search engine to identify the domain sites that are salient in representing a particular keyword. At first, Area View System will direct the keyword to a search engine and collect up to 500 hits. Each page of hits has a unique URL that may consists of a domain URL, a path, and a file name together. For example, the page http://www.cns.miis.edu/research/nuclear.html has a domain URL of http://www.cns.miis.edu/, a path of research/ and a 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. 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 as whether 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 1.

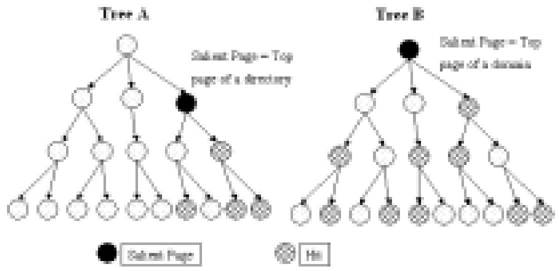

Tree A
Tree B
Salient Page = Top page of a directory
Salient Page = Top page of a domain
Salient Page
Hit

Figure 1.

Figure 1 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 2.] devoted to the keyword. Salient page is always at the tip of the information cone.

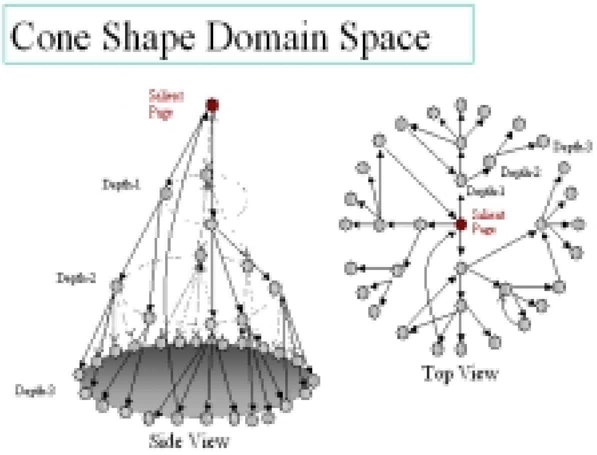


Figure 2.

2.3. Salient Pages Verification

By just analyzing their unique URL, the 50 salient pages derived above are still not fully reliable. Hence, Area View System need to do more detail analysis on the salient pages' information cone in order to recognize the real information cones with high suitability. Basically, Area View System will study the pages located inside the salient page's information cone. The suitability [Equation 1] of an information cone for tracking will be determined by two parameters: outer link ratio and content page ratio. Outer link ratio is the ratio of the number of outer links in the information cone pointing into the other 49 information cones to the total outer links in the information cone. While content page ratio is the ratio of the number of pages in the information cone containing the keyword to the total number of pages in the information cone.

All the salient page's information cones with suitability more than a certain trigger level will be added into the list of information cones used for tracking purpose. This collection of information cones of real salient pages is the artificially structured fraction on Internet representing the keyword the best. This fraction of information space is believed to be homogenous and the cones are having strong linking relationship among each other.

3. Web Spider

Web Spider is an autonomous robot that dispatches at a certain time interval to collect the desired html pages on the Internet. 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 [6] to traverse through the information cones for the desired pages. The mechanism is illustrated in the flow chart in Figure 3.

3.1. Mechanism

Firstly, Web Spider will get the URL information of Salient page and all the pages in all depths under it. Then, 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 link 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 will retrieve all the updated and new pages in an information cone.

4. Changes Summarizer

Changes Summarizer is designed to analyze the updated and new pages collected by the Web Spider, derive the

$$Suitability = \frac{number\ of\ outer\ links\ pointing\ into\ other\ information\ cones}{total\ outer\ links} + \frac{number\ of\ pages\ containing\ keyword}{total\ number\ of\ pages} \tag{1}$$

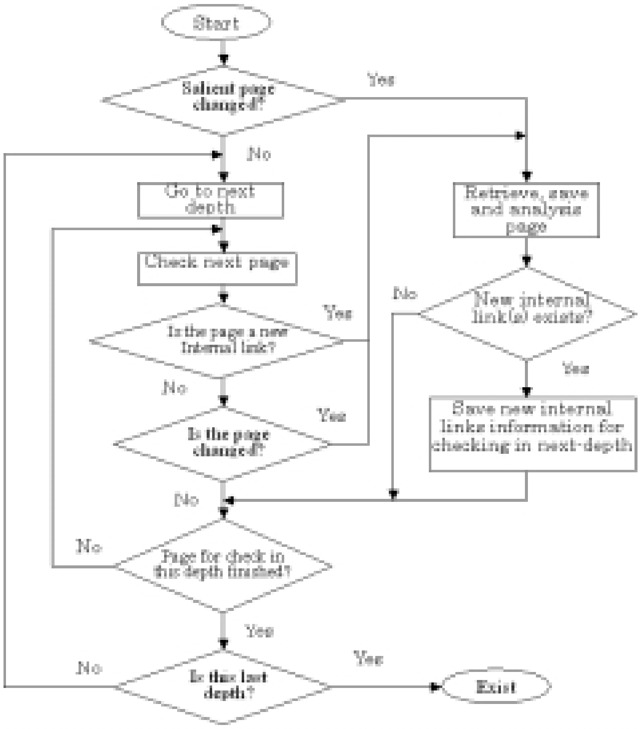

Start
Salient page changed?
Yes
No
Go to next depth
Retrieve, save and analysis page
Check next page
Is the page a new Internal link?
Yes
No
New internal link(s) exists?
No
Yes
Save new internal links information for checking in next-depth
Is the page changed?
Yes
No
Page for check in this depth finished?
No
Yes
Is this last depth?
No
Yes
Exist

Figure 3.

changes and generate a summary of emerging topic from the changes. Changes Summarizer consists of two major components. They are Changes Detector and Summary Generator. Figure 4 illustrates the combination of Changes Summarizer.

4.1. Changes Detector

Changes Detector is designed to retrieve the difference or Changes between the old pages and the newly collected pages. Changes is defined as a collection of files with all the sentences exist in the new pages but not in the old pages. Changes Detector has two component members to help it to accomplish its job. They are Sentence Parser and Sentence Retriever. Figure 5 illustrates the mechanism of Changes Detector.

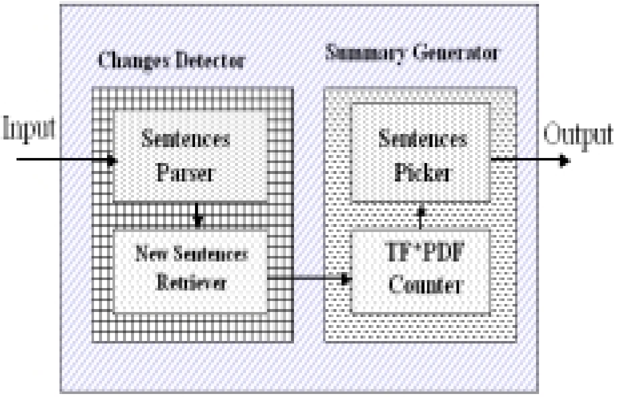


Figure 4. Changes Summarizer

4.1.1 Sentence Parser

Basically, Hypertext Markup Language (HTML) file is a text file consists of 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 any characters, terms or 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. Hence, 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.

4.1.2 Sentence Retriever

Raw text files of Html pages will be inputted from Sentence Parser. Not every line in a raw text file will be containing

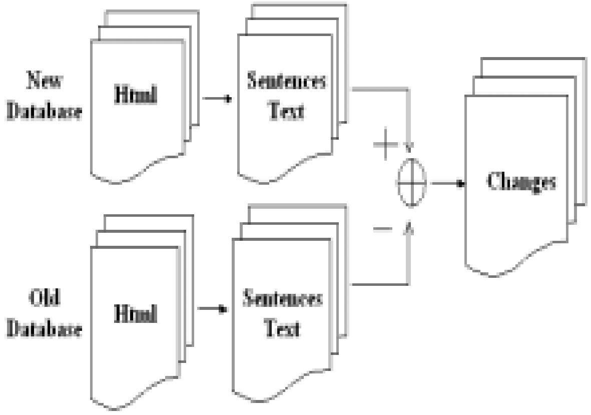


Figure 5. Mechanism of Changes Detector

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. Thus, 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 will gain two sets of sentences from each pair of old and new version of Html pages. 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 file extension. The collection of files of these new sentences is the Changes defined previously. All the files with sentences extracted from a new html file are automatically included in Changes too. No sentences comparing needed to be done in this case. In the end, Changes is instead a collection of files of new sentences from all the domain information cones under tracking.

4.2. Summary Generator

Summary Generator is designed to generate a summary of emerging topic from Changes. 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, they are TF*PDF Counter and Sentence Picker.

4.2.1 TF*PDF Counter

TF*PDF Counter is to count the weight of the terms in the Changes with a new statistical algorithm TF*PDF (Term Frequency * Proportional Document Frequency) (Equation 2).

$$W_j = \sum_{d=1}^{d=D} |F_{jd}| \exp(\frac{n_{jd}}{N_d}) \tag{2}$$

$$|F_j| = \frac{F_j}{\sqrt{\sum_{k=1}^{k=K} {F_k}^2}} \tag{3}$$

W_j=Weight of term j; F_{jd}=Frequency of term j in domain d; n_{jd}=Number of document in domain d where term j occurs; N_d=Total number of document in domain d; K=Total number of terms in a domain; D=number of domains under tracking.

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 majority of the domains. This most discussed hot topic in the changes happened to the majority of the 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 3). 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.

4.2.2 Sentences Picker

In the final stage, Sentences Picker is to calculate the average weight of each sentence in the Changes 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.

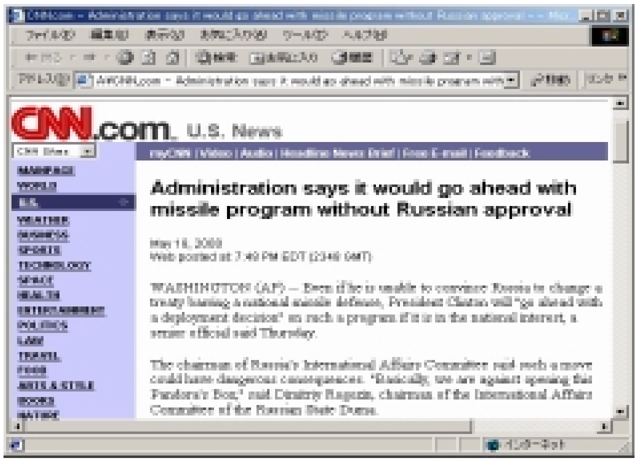
5. First Experiment Model

A keyword of ”nuclear weapons” was used. There were 22 salient pages derived by Area View System with the help of the commercial search engine Google [7]. Changes happened during the time interval between Apr 23, 2000 and Apr 30, 2000 was collected. Total changes (new sentences) were noted at 3.61 Megabytes. Table 1 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 weapons”.

The weight of the terms in the changes was counted by TF*PDF algorithm. Table 2 shows the 30 most weighted terms in the changes. Resulted summary consists of three sentences with highest average weight extracted by Sentence Picker as in Table 3. The highlighted terms in the sentences are the terms that appear in the list of 30 most weighted terms. The first sentence contains nine terms (highlighted) that appear in 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 anaphors 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 be able to 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, because it is believed that North Korea already has the capability to penetrate long ranges missile with nuclear warhead to United States of American. Thus, American peoples were in argument whether to build a national missile defense system that can counter attack the incoming missile.

5.1. Result Verification

Two weeks later, in CNN news on May 18, 2000 (illustrated in Figure 6), administration of U.S. says it will go ahead with missile program without Russian approval.



Administration says it would go ahead with missile program without Russian approval

May 18, 2000
Web posted at: 7:49 PM EDT (2349 GMT)

WASHINGTON (AP) -- Even if he is unable to convince Russia to change a treaty barring a national missile defense, President Clinton will "go ahead with a deployment decision" on such a program if it is in the national interest, a senior official said Thursday.

The chairman of Russia's International Affairs Committee said such a move could have dangerous consequences. "Basically, we are against opening this Pandora's Box," said Dmitry Rogozin, chairman of the International Affairs Committee of the Russian State Duma.

Figure 6.

6. Second Experiment Model

A keyword of ”e-commerce” was used. There were 20 salient pages derived by Area View System with the assist 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 4 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 and fourth columns show the Content Page Ratio and Outer Link Ratio of each information cone respectively.

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 7. The Suitability with lowest percentage is from 0.700 to 0.799, only five percent.

The average Suitability level is relatively lower than the average Suitability from the first experiment with the keyword ”nuclear weapons”, but the different is not large. The combination of these 20 information cones can represent the information area of ”e-commerce” on the Web well.

Table 5 shows the 30 most weighted TF*PDF terms in the Changes from Oct 3 2000 to Nov 3 2000. Table 6 shows the 30 most weighted TF*PDF terms in the Changes from Nov 3 2000 to Dec 4 2000. From the data, we found that there are 16 terms remain in the top 30 most weighted terms. There are Internet, online, information, click, Web, new, business, companies, customer, technology, e-commerce,

Table 1. First Experiment Model

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

Table 2. TF*PDF Term Weight

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

Table 3. Result Summary

Top Sentences	Average Weight
As **world** leaders gather for the **2000** Non-Proliferation **Treaty** Review **Conference** at the **United** Nations , the **United States** is on the verge of deploying a National **Missile Defense** system.	3.151
If **Russia** objects to the **United States** defending itself against the offensive efforts of other **states** that were not even conceivable threats when the ABM **Treaty** was signed nearly 30 years ago, then the **United States** must make it clear that it is no longer bound by the ABM **Treaty**.	2.630
Leaders of both the nuclear weapon **states** and potential enemy **states** know these facts and know that the **United States**, in response to a **missile** attack, could wipe out their regimes, if not their countries.	2.588

Table 4. Second Experiment Model

Salient Page	Name	Suitability	
		Content Page Ratio	Outer Link Ratio
http://ecommerce.internet.com/	Electronic Commerce Guide	0.894	0.022
http://www.commerce.net/	Commerce Net	0.643	0.006
http://www.ecominfocenter.com/	eCommerce Info Center - One Stop for eCommerce Info, Services, products and technologies	0.796	0.005
http://www.goodexperience.com/	Goodexperience.com	0.666	0.003
http://www.anu.edu.au/people/Roger.Clarke/EC/	Roger Clarke's Electronic Commerce	1	0.013
http://www.emarketer.com/	eMarketer - the world's leading provider of internet statistics	0.996	0.004
http://cism.bus.utexas.edu/	Center for Research in Electronic Commerce, UT Austin	0.575	0.003
http://ec.fed.gov/	Electronic Commerce Home Page	0.475	0.002
http://special.northernlight.com/ecommerce/	Northern Light Special Edition : Electronic Commerce	1	0.041
http://ecom.das.state.or.us/	Oregon Center for Electronic Commerce & Government	1	0.013
http://www.becrc.org/	Electronic Commerce Resource Center (ECRC), Bremerton WA	0.801	0.020
http://www.ecommercetimes.com/	E-Commerce Times: the E-Business and Technology Super Site	0.997	0.002
http://www.cio.com/forums/ec/	E-Business Research Center - Electronic Commerce Research Center	0.5	0.008
http://www.cptech.org/ecom/	CPT's Page on Electronic Commerce	0.681	0.016
http://www.diffuse.org/	Diffuse - Home Page	0.993	0.003
http://www.ec2.edu/dccenter/ecommerce/	EC2@USC - Digital Commerce Center - Electronic Center	0.723	0.017
http://www.ecommercecommission.org/	Advisory Commission on Electronic Commerce	0.827	0.001
http://www.ecomworld.com/	Electronic Commerce World	0.605	0.001
http://www.ecrc.uofs.edu/	Scraton ECRC	0.431	0.002
http://www.epic.org/	Electronic Privacy Information Center	0.883	0.010

Table 5. TF*PDF Term Weight (Oct 3 - Nov 3)

Term	Weight	Term	Weight	Term	Weight
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 6. TF*PDF Term Weight (Nov 3 - Dec 4)

Term	Weight	Term	Weight	Term	Weight
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 7. Result Summary

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

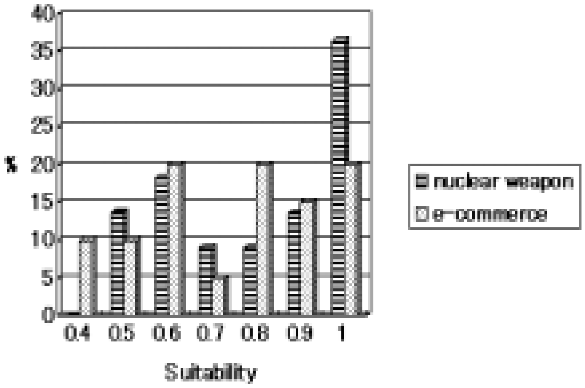


Figure 7.

use, customers, electronic, experience and site. Among them, the term ”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 5, but it appeared as one of the top 10 most weighted terms in Table 6. This shows that privacy became one of the new important issues. The resulted summary of 3 sentences with highest average weight are as in Table 7.

From the resulted summary, we can see that the sentence average weight is relatively lower than the sentence average weight in first experiment. This is because the average weight of the 30 most weighted terms are relatively low. The highlighted terms in Table 7 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.

7. Conclusion

The objective of this work is to design an intelligent Internet software application to derive the emerging topic (hot issue) from a particular information area on World Wide Web. Due to the World Wide Web is open and dynamic, contents in any information area is changing dynamically. At any time, there will be some hot issues being discussed in any information area. And it is a good assumption for us that web pages or articles regarding that hot issues will be dynamically posted on that information area on the Web. All these newly added information are defined as Changes to that information area. The system that we propose, ETTS (Emerging Topic Tracking System), is to retrieve the Changes to the information area of user interest, and further summarize an emerging topic from the Changes.

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 Link Ratio of each domain. From the experiments done, we found that the approach adapted by Area View System is appealing. So, Area View System is excellent in recognizing or matching a specific information area on World Wide Web. After recognizing an information area on World Wide Web with Area View System, Web Spider functions as an autonomous software robot that dispatches regularly at a fix time interval to collect Changes 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. Web Spider is one the major critical component in ETTS. It makes sure all the Changes happened to an information area will be retrieved for further analyzing by Changes Summarizer.

Changes Summarizer uses a novel algorithm TF*PDF (Term Frequency Proportional Document Frequency) to judge the terms that reveal an emerging topic. The experiments done show us that TF*PDF algorithm works out the way we want it be. By building ETTS and evaluating some performance issues, we conclude that this system is satisfying in fulfilling our research objectives.

References

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

[2] 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.

[3] http://informant.dartmouth.edu/

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

[5] http://www.netmind.com/

[6] Stuart J. Russell and Peter Norvig (1995). Artificial Intelligence: A Modern Approach, Prentice Hall

[7] http://www.google.com/