Available online at www.sciencedirect.com
SCIENCE DIRECT
Knowledge-Based Systems 19 (2006) 164–171
Knowledge-Based SYSTEMS
www.elsevier.com/locate/knosys

Emerging topic tracking system in WWW

Khoo Khyou Bun *, Mitsuru Ishizuka

Department of Information and Communication Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Received 16 July 2003; accepted 3 November 2005
Available online 27 December 2005

Abstract

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 interest 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 trends of the WWW information space.

Keywords: WWW; Change tracking; Emerging topic; Text summarization; Online personalized Journal

1. Introduction and motivation

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 agent which can track and acknowledge 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. 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.

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. And the users would not know this until they go and look for the changes on the pages registered. Always, the users need to figure out themselves which part of the page has changed.

In general, the conventional page trackers can only tell that some pages have been updated or some pages are new. Users are left alone to figure out themselves what are the main topics behind the changes. At this point, we still lack of a tool that can track a particular information area of user's interest, collect the changes regularly, and generate a summary of the most discussed issues from the changes back to the user regularly. ETTS (Emerging Topic Tracking System) [9] presented in this paper is such a tool for the Web information space.

2. System architecture

Fig. 1 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. Then, Area View System will analyze 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

* Corresponding author.
E-mail addresses: kbkhoo@miv.t.u_tokyo.ac.jp (K.K. Bun), ishizuka@miv.t.u_tokyo.ac.jp (M. Ishizuka).

0950-7051/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.knosys.2005.11.008

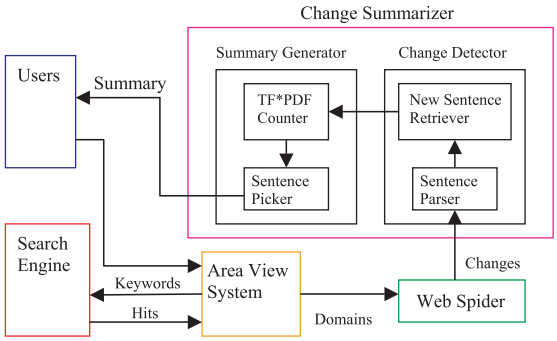

Fig. 1. ETTS system architecture.

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 novel TF×PDF (Term Frequency×Proportional Document Frequency) algorithm (Eq. 2) will be used to count the weight of the terms in the changes. This new 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.

2.1. Area View System

Area View System will first direct the user input keyword to a crawling type commercial search engine and collect all the returned URLs (hits). Each hit 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. Later from the hits, Area View System will extract 50 salient pages, which are the top pages linking to many sub pages as explained in the next paragraph. Domain URL of these salient pages would be the domain URL that occurs most frequently in the hits. Here, Area View System does not use the page rank information from the search engine for deriving these salient pages.

Salient page is the top page of a domain if the domain has its overall content relevant to the keyword. However, there are also some cases where only one sub-directory of a domain is devoted to the keyword. In these circumstances, the salient page will be the top page of the sub-directory. Area View System will determines this salient page as the top page of a domain or the top page of a sub-directory in the domain base on the shortest common path of its hits. If all the hits originated from a same domain have a shortest common path, the salient page is the top page of the sub-directory. Then, this salient page has a URL which would be the combination of the domain URL and the shortest common path. The principle on how Area View System would determine the salient page is illustrated in Fig. 2.

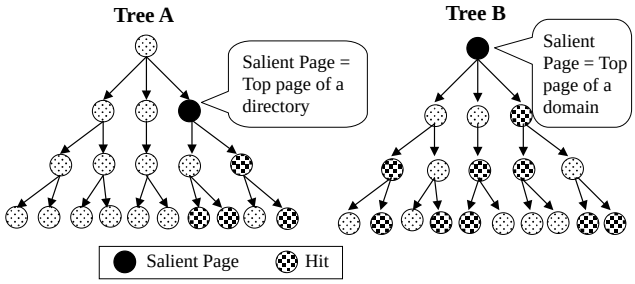

Fig. 2. Salient page determination.

Fig. 2 illustrates two different trees representing two domains. Each node represents a web page in the domain. In tree A, all the hits have a common path that is the 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 (Fig. 3). This cone provides a more comprehensive structure representation than a tree. Salient page is always at the tip of the information cone.

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 the Suitability Eq. 1 showed 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.

$$\begin{aligned}\text{Suitability} &= \text{File Ratio} + \text{Link Ratio} \\ &= \frac{\text{pages containing keywords}}{\text{total number of pages}} \\ &\quad + \frac{\text{links to other information cones}}{\text{total number of links}}\end{aligned} \tag{1}$$

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

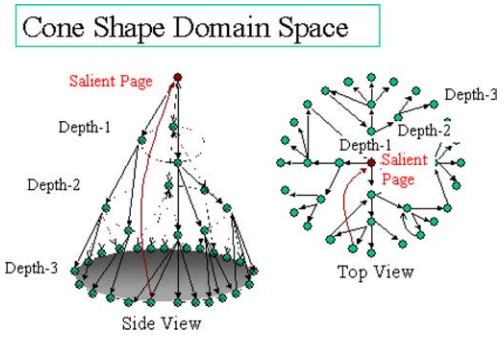

Fig. 3. Domain tree.

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. While the 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. The finally formed domain community would be the information area where we would perform tracking mechanism for changes extraction.

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 some new methodologies to derive the domain community highly devoted to the keywords.

2.2. Web Spider

Web Spider is an autonomous robot that dispatches to the Web regularly to scan all the selected information cones for new and updated HTML pages. Basically, Web Spider adapts the breath-first search algorithm to traverse all the pages in an information cone.

The flow chart in Fig. 4 illustrates the mechanism of Web Spider. First, it will analyze the salient page and all the pages linked from it. Then, it will examine each page to see whether it has changed. Similar to many other tracking tools, Web Spider uses the HTTP HEAD command to check the Last-Modified field of an HTML page for its last updated date. The page which has changed will be retrieved and saved in a database. Later on during next-depth checking, this newly changed page will also be analysed for other new internal links. In this recursive way, Web Spider will be able to retrieve all the new and changed HTML pages in an information cone.

2.3. Changes Summarizer

Changes Summarizer is designed to analyze the updated and new pages collected by the Web Spider, derive the changes and generate a summary of emerging topics from it. Changes Summarizer consists of two major components: Changes Detector and Summary Generator (Fig. 1).

Changes Detector is designed to derive the changes from the collected HTML pages. Changes is defined as a collection of text files containing all the sentences appear in the new pages but not in the old pages. Changes Detector will first wipe out all

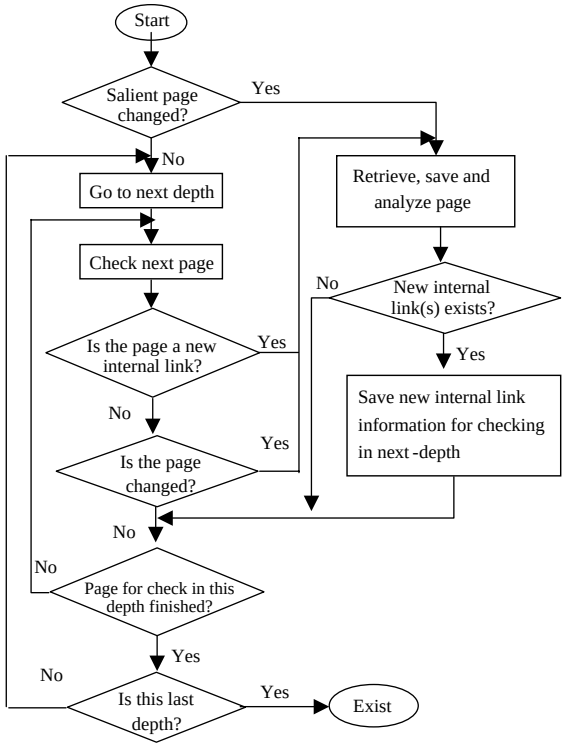

Fig. 4. Information cone traversing flow.

the HTML tags and parse the HTML pages to their sentences text files respectively. Then, it will compare the old and new versions of the sentences text file in order to derive the changes. Fig. 5 illustrates the mechanism of Changes Detector.

Summary Generator is designed to generate a summary from the changes. Summary Generator consists of two components: TF×PDF Counter and Sentence Picker. TF×PDF Counter will count the significance (weight) of the terms in the changes by using a novel TF×PDF algorithm. Terms are normally content words. Stop words like prepositions (i.e. in, from, to, out) and conjuctions (i.e. and, but, or) are eliminated via a general stop word list. Different from the conventional term weight counting algorithm TF×IDF [13], in TF×PDF algorithm, the weight of a term in a domain is linearly proportional to the term's within-domain frequency, and exponentially proportional to the ratio of document containing the term in the domain. The total weight of a term will be the summation of term's weight from each domain as follows.

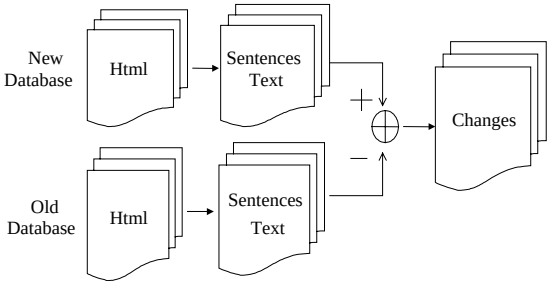

Fig. 5. Mechanism of changes detector.

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

$$|F_{jd}| = \frac{F_{jd}}{\sqrt{\sum_{k=1}^{k=K} F^2{}_{kd}}}, \quad (3)$$

where, W_j, Weight of term j; F_{jd}, Frequency of term j in the new sentences in domain d; n_{jd}, Number of changed pages in domain d where term j occurs; N_d, Total number of changed pages 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 significantly to the total weight of a term is the 'summation' of the term weight gained from each domain, provided that the term deems to explain the hot topic discussed widely in majority of the domains. This most discussed hot topic in the changes happened in majority of the domains is, in other words, an emerging topic in the information area. Therefore, the terms that deem to explain the emerging topic will be heavily weighted.

The second and third compositions are combined to give the weight of a term in a domain (within-domain term weight). The second composition is the normalized term frequency of a term in a domain ($|F_{jd}|$) as showed in Eq. 3. The term frequency needs to be normalized because the term from the domain with larger amount of changes has a proportionally higher probability of occurrence. We want, however, to give equal importance or equal weighting to the same term from each domain. Therefore, normalization ought to be carried out.

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

The total weight of a term (W_j) is equal to the sum of its term weight in all the domains. Reader may ask the reason why it is needed to distinguish the domains for calculating the term weight. The purpose of this move is to give equal importance to the changes from every domain, so that even though there is a domain containing large amount of changes with certain terms of high frequency, the results would not be deviated from detecting the terms that explain the broadly recognized main topics in majority of the domains.

In short, TF×PDF algorithm give significant weight to the terms that explain the 'hot' now topic in the changes from majority of the domains. Besides it should be noted that terms are content words, stop words like prepositions (i.e. in, from, to, out) or conjunctions (i.e. and, but, or) would be eliminated via a stop word list.

In the final stage, Sentences Picker will calculate the average weight of each sentence in the changes. The sentences with highest average weight will be used to construct the summary of changes. This summary of changes will be delivered to the users via email. This summary email is marked up with HTML tags, in such a way that the users can click on the particular summary sentence for accessing its HTML page from our Web server.

3. Sample runs

3.1. First experiment

A keyword of 'nuclear weapons' was used in the first experiment. Table 1 shows the 22 salient pages used for tracking. Third and fourth columns of Table 1 show the File Ratio and Link Ratio of each information cone respectively. Changes happened during the time interval between Apr. 23, 2000 and Apr. 30, 2000 was collected. Size of the changes (new sentences) was 3.61 Megabytes. The suitability of the cones ranges from 0.512 to 1.322.

Table 2 shows the top 30 most heavily weighted terms in the changes. Most of these 30 terms are very related to the keywords of 'nuclear weapons'. The resultant summary is showed in Table 3, where top three sentences are included. The highlighted terms in the sentences are the terms that appear in the list of top 30 most heavily weighted terms.

In the resultant summary, the first sentence contains nine terms (highlighted) that appear in the top 20 most heavily weighted terms. This sentence tells that The 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 against the ABM (Anti Ballistic Missile) treaty signed between USA and Russia 30 years ago. In the third sentence, there are dangling anaphors that make the sentence unclear because it do not 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 USA and Russia; whereas one of the emphasized potential enemy states is North Korea which is believed having the ability to penetrate a long range missile with nuclear warhead.

Thus, American people are in argument whether to build a national missile defense system which can intercept the incoming missiles. There are pros and cons in American people. However, consistent with t of our experiment, the CNN news article (appeared after our experiment on July 13, 2001) in Fig. 6 tells that The United State of America is speeding up

Table 1
First experiment salient pages (keywords: 'nuclear weapons')

Salient page	Name	Suitability	
		File Ratio	Link Ratio
http://www.acronym.org.uk/	The Acronym Institute	0.856	0.256
http://www.ananuclear.org/	Alliance for Nuclear Accountability	1.000	0.000
http://www.armscontrol.org/	The Arms Control Association — Homepage	0.767	0.018
http://www.basicint.org/	BASIC	0.825	0.120
http://www.bullatomsci.org/	Bulletin of the Atomic Scientists	0.982	0.056
http://www.ccnr.org/	The Canadian Coalition for Nuclear Responsibility	0.643	0.012.
http://www.ceip.org/programs/npp/	Carnegie Endowment — Non-Proliferation Project	0.520	0.073
http://www.cfcsc.dnd.ca/link/peace/	Peace, disarmament and arms control	0.457	0.086
http://www.clw.org/coalition/	Coalition to Reduce Nuclear Dangers — Working to Lower the Threat of Nuclear Weapons	0.902	0.018
http://www.cns.miis.edu/	Welcome to the CNS Website	0.605	0.088
http://www.dtra.mil/nuclear/	DTRA — Nuclear Support	0.909	0.000
http://www.fas.org/nuke/	Nuclear Resources	0.583	0.036
http://www.hookele.com/abolition2000/	Abolition 2000 — GLOBAL NETWORK TO ELIMINATE NUCLEAR WEAPONS	0.696	0.038
http://www.igc.org/disarm/	NGO Committee on Disarmament	0.973	0.084
http://www.ippnw.org/	IPPNW — International Physicians for the Prevention of Nuclear War	0.622	0.000
http://www.napf.org/	Home Page of Nuclear Age Peace Foundation	0.952	0.370
http://www.nci.org/	Nuclear Control Institute (NCI), Washington DC	0.817	0.001
http://www.nuclearfiles.org/	The Nuclear Files Experiencing ethical and political challenges of the nuclear age.	0.994	0.225
http://www.nukefix.org/	Nuclear weapon research on the Internet	0.816	0.042
http://www.stimson.org/policy/	The Committee on Nuclear Policy	0.988	0.063
http://www.un.org/Depts/dda/	United Nation— Disarmament	0.512	0.000
http://www.wagingpeace.org/	Home Page of Nuclear Age Peace Foundation and Abolition	0.952	0.186

Table 2
TF×PDF term weights (keywords: 'nuclear weapons')

Term	Weight	Term	Wt	Term	Wt	Term	Wt	Term	Wt
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	UN	2.680	War	2.237
United	4.762	Npt	2.572	Russian	2.216	Missile	4.371	US	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
Resultant summary for the keywords of 'nuclear weapons'

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

their missile defense tests, although it could violate the 1972 Anti-Ballistic Missile treaty.

3.2. Second experiment

A keyword of 'e-commerce' was used in the second experiment. There were 20 salient pages derived by Area View System. Changes happened during the two time intervals of Oct. 3. Nov. 3 and Dec. 4, 2000 were collected. Table 4 shows the salient pages which were used for tracking. Third and fourth columns of Table 4 show the File Ratio and Link Ratio of each information cone, respectively. Table 5 shows the top 30 most weighted TF×PDF terms in the changes happened between Oct. 3 2000 and Nov. 3 2000. Table 6 shows the top 30 most heavily weighted TF×PDF terms in the changes between Nov. 3 2000 and Dec. 4 2000.

From the data in Table 5 and Table 6 we can find that there are 16 terms remain in the top 30 most heavily weighted terms.

Fig. 6. CNN news July 13 2001.

They are 'Internet', 'online', 'information', 'click', 'Web', 'new', 'business', 'companies', 'customer', 'technology', 'e-commerce', 'use', 'customers', 'electronic', 'experience' and 'site'. Among them, the term 'Internet' gained and remained the term with the 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 can realize is that the term 'privacy' is not one of the terms in Table 5, but it appears as one of the top 10 most heavily weighted terms in Table 6. This shows that privacy had become one of the new important issues.

The resultant summary is constructed by the three sentences with highest average weight as in Table 7. The highlighted terms in Table 7 are among the top 30 most heavily 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 US government is unlikely to force electronic signatures implementation in Internet business transactions. And the third sentence concerns Web privacy practices. This third sentence tells that a Web site with good privacy practice should declare what the company will do with the users data.

In a CNN Web page (April 17, 2001) of Fig. 7, it was reported that more than 60 federal Web sites violates US privacy rules by using unauthorized software to track the browsing and buying habits of Internet users. While in Fig. 8, 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 hot issue or an emerging topic discussed widely.

4. Related work

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 [3], the HITS algorithm [4], bipartite subgraph identification [5] and focused crawling [6]. Besides, [1] describes a way to identify to topic relevant portions of a hierarchical space, while [2] 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 want to build an information space that will trap all the related new information. The precision is not highly important for us in the first state

Table 4
Second experiment salient pages (keyword: 'e-commerce')

Salient page	Name	Suitability	
		File Ratio	Link Ratio
http://ecommerce.internet.com/	Electronic commerce guide	0.894	0.022
http://www.ecommerce.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 weights (period between Oct. 3 and Nov. 3, 2000)

Term	Weight	Term	Wt	Term	Wt	Term	Wt	Term	Wt
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 weights (period between Nov. 3 and Dec. 4, 2000)

Term	Weight	Term	Wt	Term	Wt	Term	Wt	Term	Wt
Internet	2.927	Global	1.51	Electronic	1.122	Online	2.835	Technology	1.432
Said	1.077	Information	2224	Ecommerce	123	Policy	1045	Click	2139
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
Resultant summary for the keywords of ‘e-commence’

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 US 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

because later in the next stage, we will have our TF*PDF filter out unwanted information and present us the topic terms for generating a topics summary.

While using conventional commercial page tracker [10] [11] to track for HTML page changes, it can be annoying if the users are always pushed with acknowledge email although the changes is trivial. In order to solve this problem, WebBeholder [7] 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 changes is

Fig. 7. CNN News April 17 2001.

Fig. 8. CNN News May 25 2001.

not interesting to them. Vice versa, the users can miss some important changes.

Output from conventional tracking systems always show little or no information on how the pages have changed. Thus, the AT&T Internet Difference Engine (AIDE) [8] 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 pages.

Change detector [12] is a tool that is purposed to monitor the changes on entire web sites. It can tells if the structure of an organization has changed, instead of acknowledging some simple changes happen on certain web pages. It rely 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. Change detector can do more by informing the changes which is not easy to notice. 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.

5. Conclusion

The objective of this work is to design an intelligent Internet software application to derive the emerging topics (hot topics) from a particular information area on the World Wide Web. As the World Wide Web is open and dynamic, contents in any information area are changing dynamically. At any time, there will be some hot issues being discussed in any information area. Thus, 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 have developed, ETTS (Emerging Topic Tracking System), is to retrieve the changes in the information area of user interest, and further generate a summary of changes back to the users. We have shown some of the experimental results to illustrate its effectiveness.

To have this system reporting us the most updated topics related to our keywords regularly, we are 'all time aware' of the latest trends in the information area of our interest. In other words, we can live with our weekly, biweekly, monthly and bi-monthly personalized journals, which autonomously find emerging topics in the WWW space and present their summaries.

References

[1] G.M. Sacco, Dynamic taxonomies: a model for large information bases, IEEE Transactions on Knowledge and Data Engineering 12 (3) (2000) 468–479.
[2] L. Terveen, W. Hill, B. Amento, Constructing, organizing, and visualizing collections of topically related web resources, ACM Transactions on Computer–Human Interaction 6 (1) (1999) 67–94.
[3] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Proceedings of the Seventh International WWW Conference, 1998.
[4] Jon M. Kleinberg, Authoritative sources in a hyperlinked environment, Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 668–677.
[5] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Trawling the web for emerging cyber-communities, Proceedings of the eight International WWW Conference, 1999.
[6] S. Chakrabarti, M. van Der Berg, B. Dom, Focused crawling: a new approach to topic-specific web resource discovery, Proceedings of the eight International WWW Conference, 1999.
[7] S. Santi, I. Mitsuru, WebBeholder: a revolution in tracking and viewing changes on the web by agent community, in: Proceedings (CD-ROM) WebNet98, Three World Conference on WWW and Internet, Orlando, Florida, USA, 1998.
[8] Fred Douglis, Thomas Ball, Yih-Farn Chen, Eleftherios Koutsofios, The AT&T internet difference engine (AIDE): tracking and viewing changes on the web, World Wide Web 1 (1998) 27–44.
[9] K.B. Khoo, M. Ishizuka, Emerging topic tracking system, in: Proceedings of the three Int'l Workshop on Advanced Issues on E-Commerce and Web-Based Information Systems (IEEE Computer Soceity), San Jose, California, USA, 2001.
[10] ChangeDetect, http://www.changedetect.com/.
[11] WebSpector, http://www.illumix.com/.
[12] V. BoyaPati, K. Chevrier, A. Finkel, N. Glance, T. Pierce, R. Stokon, C. Whitmer, Change detector: a site-level monitoring tool for the WWW, in: Proceedings of the WWW Conference 2002, Hawaii, USA, 2002, pp. 570–579.
[13] G. Salton, C. Buckley, Term-weighting approached in automatic text retrieval, Information Processing and Management 14 (5) (1988) 513–523.