

Information Area Tracking and Changes Summarizing System in WWW

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: Due to its open characteristic, the Web is being posted with vast amount of new information, or 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 Information Area Tracking and Changes Summarizing System (ATSS) 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.

1. Introduction and Related Work

User or professional would like to be always updated with the latest hot topic emerging in the particular information area of their interest. However, due to the information on 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 upon changes took place on the pages or information area of our interests.

There are quite a number of commercial tracking tools (Santi98) have become available for online services. Basically, when users want to track a particular html page on the Web, they need to register the URL of the page with the system. And upon any changes happen on the page, they will be acknowledged through email. However, 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) 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. Other than tracking some specified URLs, some systems, i.e. Informant (<http://informant.dartmouth.edu/>) and Netmind (<http://www.netmind.com/>) are featured to detect the new pages containing the user input keywords.

In general, the conventional page trackers 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.

2. System Architecture

Figure 1 illustrates the system architecture of ATSS. ATSS consists of three main components: Area View System (AVS), Web Spider and Changes Summarizer. After taking in an input keyword from the user, AVS will direct the keyword to the commercial search engine Google (<http://www.google.com/>). Then, AVS will analysis the returned hits and derive a number of *domains* that are most related to the keywords. These domains are grouped together to form an information area devoted to the keyword. Then, 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. Then, the Changes Summarizer will extract all the Changes (newly added sentences) from the collected html files by comparing the old and new database. Then, a new algorithm $TF*PDF$ (Term Frequency * Proportional Document Frequency) (Equation 1) 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.

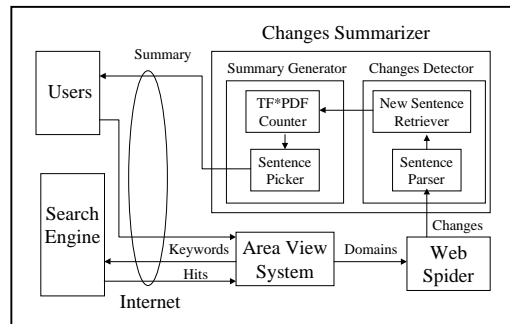


Figure 1: ATSS System Architecture

2.1 Area View System

Area View System will direct the user input keyword to the search engine Google and collect all the hits. Each hit has a unique URL that may consist 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`. From all the returned hits, AVS will further derive 50 salient pages with their domain URL occur most frequently. Salient page is the top page of a domain 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 will be the top page of the sub-directory. AVS 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 on how AVS can determine the salient page is illustrated in Figure 2.

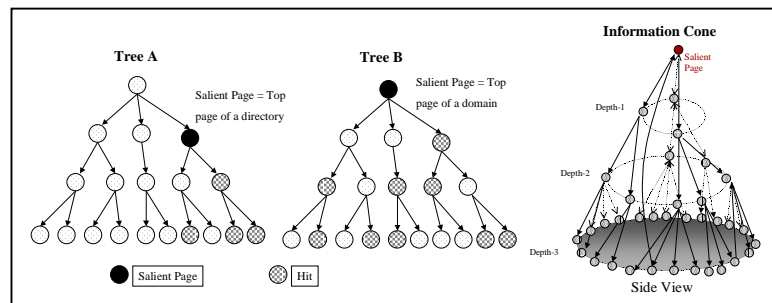


Figure 2: Domain Tree and Information Cone

Figure 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 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 as showed in Figure 2. This cone provides a more comprehensive structural representation than a tree. Salient page is always at the tip of the information cone.

However, by just analyzing the URL's frequency in determining the domains for tracking usage is insufficient. Hence, AVS will 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 Equation showed below. Suitability of an information cone is equal to its Link Ratio plus File Ratio. 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.

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

2.2 Web Spider

Web Spider is an autonomous robot that dispatches to the Web regularly to scan all the qualified information cones for new and updated html pages. Basically, Web Spider adapts Breath-first search algorithm (Russell95) to traverse through the information cones.

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 topic from the Changes. Changes Summarizer consists of two major components: Changes Detector and Summary Generator. 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 the html tags and parse the html pages in sentences text file. Then, it will compare the old and new version of sentences text file in order to derive the Changes. Then, Summary Generator will be used 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 the new TF*PDF algorithm. Terms are normally content words. Stop words like prepositions (i.e. in, from, to, out) and conjunctions (i.e. and, but, or) are eliminated via a general stop word list. Different from the famous TF*IDF (Salton98) algorithm, in TF*PDF, 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.

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 a summary.

$$W_j = \sum_{d=1}^D |F_{jd}| \exp\left(\frac{n_{jd}}{N_d}\right) \rightarrow (Eq. 1)$$

$$|F_{jd}| = \frac{F_j}{\sqrt{\sum_k (F_k)^2}} \rightarrow (Eq. 2)$$

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=Number of document in domain d;
 k=total number of terms in a domain;
 D=number of domains under tracking.

3. Experiment

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 in between Oct 3, 2000, Nov 3 and Dec 4, 2000 was collected. Table 1 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 File Ratio and Link Ratio of each information cone respectively.

Table 2 shows the top 30 most weighted TF*PDF terms in the Changes from Oct 3 2000 to Nov 3 2000. Table 3 shows the top 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, use, customers, electronic, experience and site. Among them, the term "Internet" gained and remained the term with highest term weight. This

concur with the fact that the Internet is the vital way in doing electronic commerce. From the data, another important point that we realized is that the term “privacy” was not one of the terms in Table 2, but it appeared as one of the top 10 most weighted terms in Table 3. This shows that privacy had become one of the new important issues. The resulted 3 sentences with highest average weight are as in Table 4.

The highlighted terms in Table 4 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.

Salient Page	Name	Suitability	
		File Ratio	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/e-commerce/	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/dcenter/e-commerce/	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 1: Experiment domains

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 2: TF*PDF Terms with highest weight (Oct 3 – Nov 3)

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 3: TF*PDF Terms with highest weight (Nov 3 – Dec 4)

<i>Top Sentences</i>	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

Table 4: Result Summary

4. Experiment Result Verification



Figure 3: CNN News April 17 2001



Figure 4: CNN News May 25 2001

In Figure 3, 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 4, 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.

5. Discussion

The objective of this work is to innovate an intelligent Internet software application to derive the emerging topic (hot topic) in a particular information area in the Web. Due to the Web is open and dynamic, contents in any information area is changing dynamically. Consequently, at any time, it is conceivable that there will be some hot issues being discussed and posted in any information area on the Web. The newly posted information is defined as Changes to that information area. And the system that we proposed, ATSS, is to retrieve this Changes and derive an emerging topic from it.

6. Conclusion

In this paper, we have proposed a novel system, ATSS, and evaluated it by putting a proper experiment in place. 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.

References:

- Santi Saeyor and Mitsuru Ishizuka (1998). 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*, Nov. 1998, Association for the Advancement of Computing in Education, Orlando, Florida, USA.
- Fred Douglass, Thomas Ball, Yih-Farn Chen and Eleftherios Koutsofios (1998). 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.
- Stuart J. Russell and Peter Norvig (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice-Hall Inc.
- Salton, G. and Buckley, C. (1998). Term-Weighting Approached in Automatic Text Retrieval. *Information Processing and Management*, Vol.14, No.5, 1998.