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. Due to its open characteristic, the Web is being posted with vast amount of new information dynamically. Consequently, at any time, there will be hot issues emerge in any information area which may interest the users. However, it is not practical for users to browse the Web all the time for the updates. 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 from the changes back to us from time to time. This summary of changes will be the latest most discussed issues and it may reveal an emerging topic.

1 System Architecture

Figure 1 illustrates the system architecture of ETTS. ETTS consists of three main components: Area View System (AVS), Web Spider and Changes Summarizer. After taking in a keyword from the user, AVS will direct the keyword to the commercial search engine Google [6]. 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 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.

1.1 Area View System

Area View System will direct the user input keyword to the search engine Google and collect up to 500 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

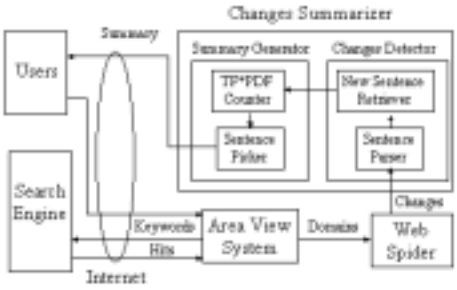

Fig. 1. ETTS System Architecture

of `http://www.cns.miis.edu/`, a path of research/ and a file name of nuclear.html. From the 500 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.

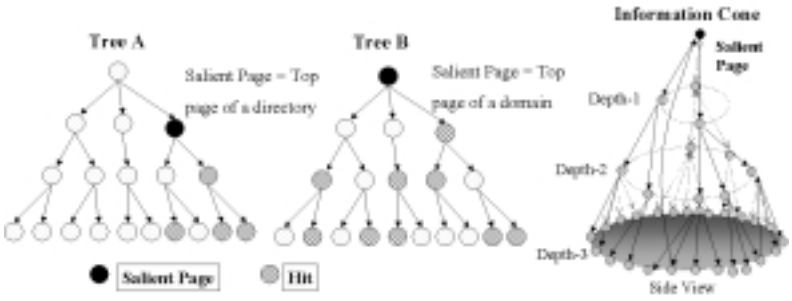

Fig. 2. Domain Tree and Information Cone

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 (Figure 2). This cone provides a more comprehensive struture represenation 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 determined by Equation 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.

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

1.2 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 [5] to traverse through the information cones.

1.3 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 (Figure 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 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 conjuctions (i.e. and, but, or) are eliminated via a general stop word list. Different from the famous TF*IDF [7] 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.

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

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

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.

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.

Table 1. First Experiment Salient Pages

Salient Page	**Suitability**	
	Content Page Ratio	**Outer Link Ratio**
http://www.acronym.org.uk/	0.856	0.256
http://www.ananuclear.org/	1.000	0.000
http://www.armscontrol.org/	0.767	0.018
http://www.basicint.org/	0.825	0.120
http://www.bullatomsci.org/	0.982	0.056
http://www.ccnr.org/	0.643	0.012
http://www.ceip.org/programs/npp/	0.520	0.073
http://www.cfcsc.dnd.ca/link/peace/	0.457	0.086
http://www.clw.org/coalition/	0.902	0.018
http://www.cns.miis.edu/	0.605	0.088
http://www.dtra.mil/nuclear/	0.909	0.000
http://www.fas.org/nuke/	0.583	0.036
http://www.hookele.com/abolition2000/	0.696	0.038
http://www.igc.org/disarm/	0.973	0.084
http://www.ippnw.org/	0.622	0.000
http://www.napf.org/	0.952	0.370
http://www.nci.org/	0.817	0.001
http://www.nuclearfiles.org/	0.994	0.225
http://www.nukefix.org/	0.816	0.042
http://www.stimson.org/policy/	0.988	0.063
http://www.un.org/Depts/dda/	0.512	0.000
http://www.wagingpeace.org/	0.952	0.186

2 Experimental Results

A keyword of "nuclear weapons" was used. In table 1, there were 22 information cones used for tracking. 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. TF*PDF Term Weight

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	u.n.	2.680	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 2 shows the 30 most weighted terms in the Changes. The result summary is showed 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 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 again 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 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 USA and Russia; whereas one of the emphasized potential enemy states is North Korea which is believed having the ability to penetrate long range missile with nuclear warhead. Thus, American peoples are in argument whether to build a national missile defense system that can counter attack incoming missile.

3 Related Work and Discussion

There are quite a number of commercial tracking tools [1] 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) [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. Other than tracking some specified URLs, some systems (Informant [3], Netmind [4]) are featured to detect the new pages containing the user input keywords.

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

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.

4 Conclusion

In this paper, we have proposed a novel system, ETTS, 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

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. http://www.netmind.com/
5. Stuart J. Russell and Peter Norvig (1995). Artificial Intelligence: A Modern Approach, Prentice Hall
6. http://www.google.com/
7. Salton, G. and Buckley, C.: Term-Weighting Approached in Automatic Text Retrieval, Information Processing and Management, Vol.14, No.5, 1998