Emerging Topic Tracking System

5U-09

Khyou Bun KHOO Mitsuru ISHIZUKA
The University of Tokyo
e-mail: {kbkhoo, ishizuka}@miv.t.u-tokyo.ac.jp

1. Introduction

Information regardless of profession on the web is changing dynamically everyday. User or professional with interests in a particular field would like to be always updated with the latest hot topic concerning their profession. However, due to the changes happen in a random way, updating themselves by surfing some particular sites manually and regularly is both a difficult and time consuming job, yet there are no promises of new information changes have been taking place. Thus, here come the needs of a tool to track the changes in the information area and report it to the user regularly and automatically in a summarized format, as an emerging topic. And this is the goal of this research being carried on.

2. System Overview

Emerging Topic Tracking System or ETTS is implemented by using a collection of cooperative software agents. This agent community consists of 4 main components, which are Mediator, Area View System, Web Crawler and Changes Summarizer. Each of these agents has their own unique task to accomplish. They collaborate or interact with each other to achieve the goal of the whole ETTS system. Firstly, Mediator serves as the information gateway between user and ETTS system. It receives and studies the user's input query. Besides, it also acts as the presenter of the ETTS system, for delivering the summary of changes to the users. It presents the result in two forms, one is using Push Methodology by sending email to users, and the other is the Pull Methodology by constructing a Http Server. In the second stage, Area View System will take the responsibility to derive the most related or core web domains, which when combined together can represent the whole knowledge base of the user's field of interest. Area View System is instead a Meta-search engine, which direct the user's input query to commercial search engine like Yahoo, and further derive the core domains by analyzing the link structure of the hits. Then in the third stage, a Web Crawler is designed to collect all the html pages within the web domains derived by Area View System. Web Crawler adapts Breath-first search algorithm to scan through all the pages in all depth in the domains recursively. The collected html pages cluster from one domain virtually shapes an information cone with the main page at its tip [Fig 1]. And these information cones are believed to be having homogenous information and strong linking relationship between each other. These information cones realize an artificially structured information area in the decentralized Internet information space. Web Crawler will be dispatched to collect these information cones regularly, may be once in a month, depends on how dynamic the information area is. Then, with the old and newly collected information cones, in the final stage, we have this Changes Summarizer to compare the old and new html pages, in order to extract the changes, and then further generate a summary of the changes as an emerging topic to the user. This paper emphasize on the design of Changes Summarizer.

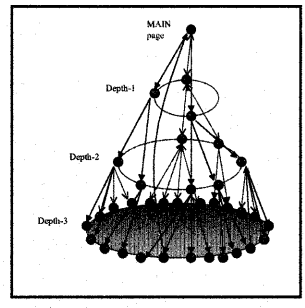

Figure 1.

3. Changes Summarizer

Changes Summarizer consists of two components. There are Changes Detector and Summary Generator. Changes Detector is designed to compare the old and new pages in order to derive the difference in sentence

units. But before comparing, Changes Summarizer has to analysis the html text, parse it to a tree structure, prune the unwanted branches, and then extract the sentences by using a sentence pattern matching method. Then, Summary Generator will generate a summary of changes as an emerging topic. Basically, this summary is formed by the sentences with highest average term weight. The term weight is calculated by utilizing a novel algorithm TF*PDF (Term Frequency * Proportional Document Frequency) [Eq.(1)] which evolve from the famous conventional term weight counting algorithm TF*IDF[1]. TF*PDF is imposed on all the changes from the domains under tracking. In calculating a term weight in a domain, we first calculate the normalized term frequency, and then we multiply it with PDF. PDF is defined as proportional document frequency. PDF concepts that terms that occur in more documents of changes are more valuable or weighted because more frequent these terms exist in different changes documents, the more these terms can imply the hot topic. Then the weight of the same term among the domains is summed together to get the total weight of the term. Significantly, term with high total term weight will imply the common hot topic among the domains or in the information area of user interest.

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

$$\left|\overrightarrow{F_j}\right| = \frac{F_j}{\sqrt{\sum_k (F_k)^2}}$$

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 occur; N_d=Number of document in domain d.

Table 1.

TERM	WEIGHT	TERM	WEIGHT
nuclear	29.002	united	4.762
weapons	11.598	missile	4.371
states	9.726	international	4.103
Treaty	8.315	peace	3.699
Conference	4.964	new	3.526

4. Experimental Model

A query of "nuclear weapon" was used and 25 domains were picked. Html files of these domains were collected on 2000 Apr 23 (261 Megabytes) and Apr 30 (263 Megabytes). Total changes (new sentences) from Apr 23 to Apr 30 were recorded as 3.61 Megabytes. Ten terms with highest weight is illustrated in Table 1. From the table, we can see that the term "nuclear" and "weapon" were gaining the highest weight since they are the keywords. And the following terms are the best terms to describe the keywords as the current most discussed topic. Two constraints were imposed: (1) Exclude the query terms during sentence weight counting (2) Only pick sentence with number of words > 15. As a result, the following two sentences in Table 2 were picked. The first sentence tells that American is about to deploy NMD system while the second sentence tells that Russian is again the deployment of this system. It is obvious that a good result of estimation of emerging topic in the information area of "nuclear weapon" was met.

Table 2.

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.
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.

5. Conclusion

ETTS is the post search engine Internet tool. ETTS is devoted to track the dynamic information in a particular area of interests, which is out of the capability of concurrent search engine's capability. TF*PDF was designed to complement the task of ETTS in deriving a summary of the changes or emerging topic in a particular information area.

Reference

[1] Salton, G. and Buckley, C.: Term-Weighting Approached in Automatic Text Retrieval, *Information Processing and Management*, Vol.14, No.5, 1998