

Key words:
Weblog, social concerns, analysis tool

Tomohiro FUKUHARA* and Toshihiro MURAYAMA*

An analysis tool for understanding social concerns using Weblog articles

An analysis tool for understanding social concerns using Weblog articles is proposed. Weblog is a popular personal publishing style on the Internet. By analyzing Weblog articles, we can find social concerns and public opinions at this time. On the other hand, collecting and analyzing a large number of Weblog articles are not easy. We propose an analysis tool of Weblog articles for assisting users' understanding of social concerns. The tool collects Weblog articles automatically. By using the tool, users can understand social concerns easily. We describe typical patterns of social concerns found by the tool.

1. INTRODUCTION

The Internet becomes a large space of information. Although contents creation and management were not easy in the early stage of the WWW (World Wide Web), a content management system (CMS) changed the situation. CMS assists creation and maintenance of Web pages. One can easily create and maintain Web pages by using CMS. In addition to CMS, an online publishing style called *Weblog* becomes popular. Many people publish Weblog articles using a CMS instantly. The number of Weblog articles is increasing rapidly.

Weblog articles contain various topics such as on personal activities, technology, politics, international problems, and so on. By browsing Weblog articles, we can find frank and up-to-date opinions on various topics such as computer software, restaurants, social problems, and so on.

*Research Institute of Science and Technology for Society (RISTEX),
2-5-1 Atago, Minato-ku, Tokyo, JAPAN (E-mail: fukuhara@ristex.jst.go.jp)

In our laboratory, we investigate methods for solving problems in the socio-technical domain such as nuclear power plant, traffic safety, food safety, medical safety, prevention of disasters, information security, and so on. By analyzing Weblog articles, we can find points in which people feel anxious and doubts. If we can find those points, we can find solutions of the problem. Thus, understanding social concerns is important in socio-technical domain.

In this paper, we propose an analysis tool for understanding social concerns using Weblog articles. Proposed tool collects Weblog and news articles automatically, and provides a graph that shows recent trend of social concerns, and related news articles. Users can find an overview of social concerns.

This paper consists of the following sections. In the section 2, requirements of an analysis tool for understanding social concerns is described. In section 3, the architecture and functions of the implemented system are described. In section 4, typical patterns of social concerns found by the tool are described. In section 5, we compare with related works. In section 6, we conclude.

2. UNDERSTANDING SOCIAL CONCERNS

In this section, we describe (1) an approach for understanding social concerns using Weblog articles, and (2) requirement of an analysis tool.

2.1 SOCIAL CONCERNS IN WEBLOG ARTICLES

Understanding social concerns is important for decision making. Weblog articles are good resource for understanding social concerns because opinions in those articles reflect both of the real and the virtual worlds. People write an article when events occurred in the real or virtual world. From this article, we can find how people think about social problems. On the other hand, it is difficult for collecting and analyzing articles because large number of articles is published every day. It is quite difficult to collect and analyze these articles manually.

Our aim is to understand social concerns effectively. Toward this aim, we collect Weblog articles and analyze them. An analysis tool that facilitates our understanding of social concerns is needed.

2.2 REQUIREMENTS

Followings are requirements of an analysis tool for understanding social concerns.

1. Collecting Weblog articles automatically
2. Analyzing articles quantitatively and qualitatively

3. Visualizing analysis results

The first is to collect Weblog articles automatically. The system should collect and index articles so that users can retrieve a number of articles effectively. Because a large number of articles are posted every day, indexing and retrieving articles effectively is important issue. The system should allow users to retrieve up-to-date articles.

The second is to analyze Weblog articles quantitatively and qualitatively. It is important for understand social concerns from both of quantitative and qualitative aspects. In case of quantitative aspect, various data related to social concerns should be provided with users. In case of the qualitative aspect, the tool should extract information from articles, and integrates the extracted information into an analysis result.

The third is to visualize analysis results. Facilitating users' understanding of analysis results is important. The system should visualize analysis results by converting text and numerical data into graphs so that users can easily understand the results.

3. PROTOTYPE SYSTEM

We created a prototype system of the analysis tool according to the requirements. In this section, we describe the architecture and the functions of the system.

3.1 SYSTEM ARCHITECTURE

Fig. 1 shows an overview of the system. The system consists of a database, an httpd server, and several Perl scripts for collecting and retrieving Weblog articles. The system collects RSS (RDF Site Summary¹) files of Weblog sites. An RSS file contains title, summary, date of publish, author, and category of articles. Because an RSS file is described in XML, it is easy to extract information. The system collects RSS files every 10 minutes. We started to collect articles from 18 March 2004, and acquired 1,012,251 articles at this moment².

RSS files are collected from (1) personal Weblog sites, (2) news sites, and (3) governmental Web sites. In case of the first and the second type of information, the system collects RSS files from Japanese Weblog ping servers such as Myblog Japan³ directly. In case of the third type of information, the system acquires Web pages from governmental Web sites, converts them into RSS files, and acquires information from those files.

¹ <http://web.resource.org/rss/1.0/spec>

² 2004/05/20 19:38:00

³ <http://www.myblog.jp/>

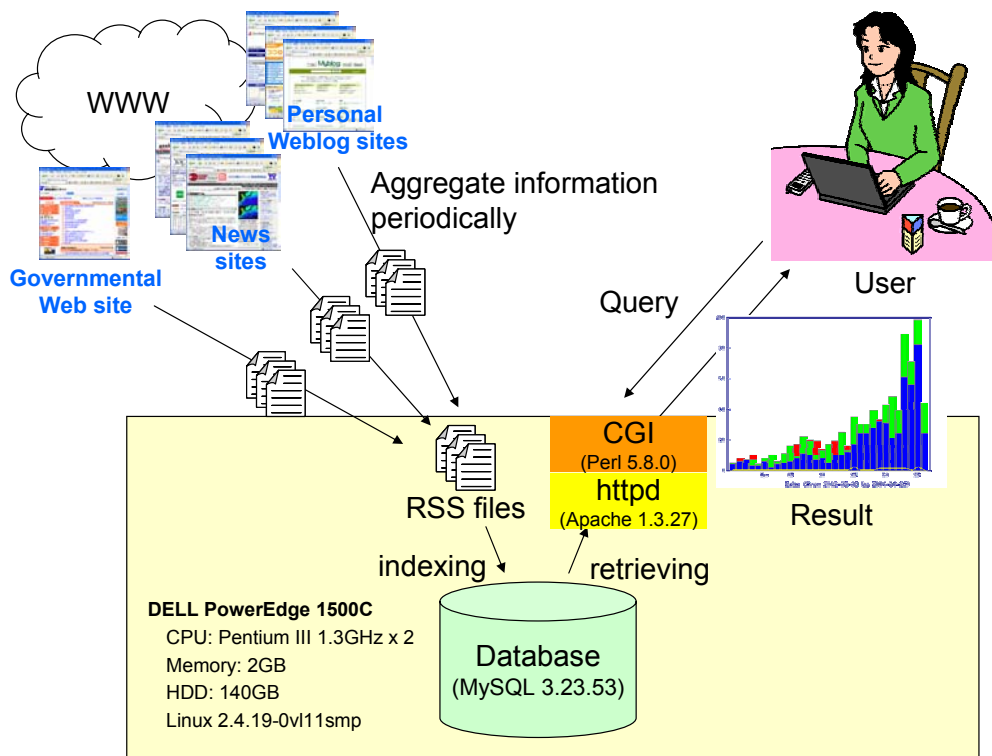


Fig. 1. Overview of the system

The system accepts user's query, which is a set of keywords, and returns a Web page as an analysis result.

3.2 FUNCTIONS

The system has following functions.

1. Retrieving articles
2. Extracting keywords of the day
3. Finding relevant news articles of the day

First, the system has the retrieving function. Users can retrieve Weblog articles by specifying a set of keywords. We use the Boolean model [1]. Fig. 2 shows screen shots of search results. The left figure contains a graph of a daily trend of articles, and a list of articles. Users can find trend of the keyword(s). The right figure is another search result which contains images in articles. Because the system collects articles every 10 minutes, users can acquire up-to-date results.

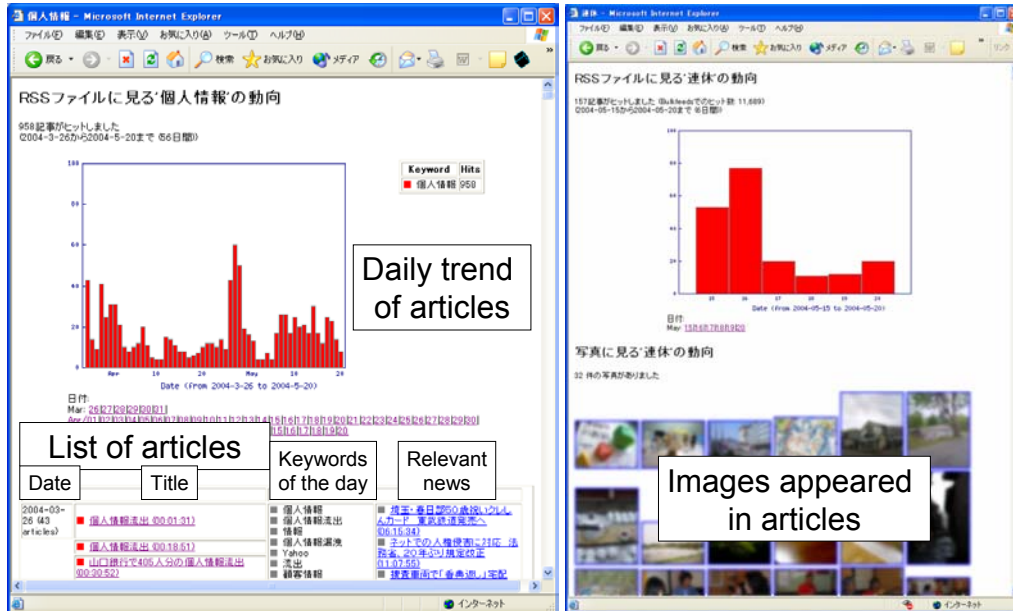


Fig. 2. Screen shots of search results

Second, the system extracts keywords from articles of the day, and displays them. We use a term extracting tool developed by Nakagawa et al.[2]. By browsing extracted keywords, users can find topics appeared in articles of that day.

Third, the system finds relevant news articles and displays them. News articles are retrieved by (1) keywords specified by a user, and (2) keywords extracted from articles.

4. PATTERNS OF SOCIAL CONCERNS

In this section, we describe typical five patterns of social concerns. Fig. 3 shows a list of those patterns.

4.1 PERIODIC PATTERN

In this pattern, several peaks appear periodically. This pattern appears when there occur social events, which are watched with keen interest, periodically. The graph of periodic pattern shown in Fig. 3 is the result of a keyword "Winter Sonata" that is the name of Korean TV drama. Because this drama attracts public attention recently in Japan, several periodic peaks clearly appeared in the graph.

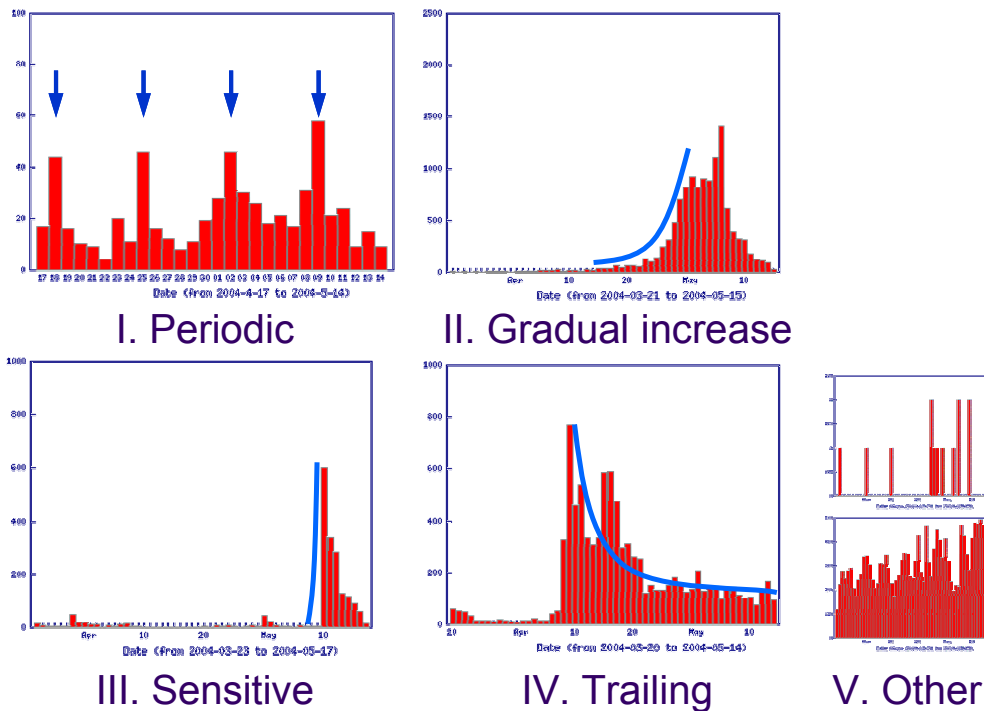


Fig. 3. Patterns of social concerns

4.2 GRADUAL INCREASE PATTERN

In this pattern, a peak appears gradually. This pattern appears when people know the social event beforehand, and they are talking about the event on the Weblog. The graph of gradual increase pattern shown in Fig. 3 is the result of a keyword “Golden Week” which is the name of Japanese national holidays. As shown in the graph, people had strong interest in these holidays beforehand.

4.3 SENSITIVE PATTERN

This pattern has a keen peak. This pattern appears when a serious matter, which has a heavy impact on the society, is broadcasted. The graph of sensitive pattern shown in the Fig. 3 is the result of a keyword “Winny” which is the name of pear-to-pear

software. Articles on “Winny” are posted intensively on May 10 when the programmer of this software was arrested⁴.

4.4 TRAILING PATTERN

In this pattern, social concerns last after one or several matters occur. The graph of trailing pattern shown in the Fig.3 is the result of a keyword “Iraq”. The period of the graph begins from 20 March to 14 May 2004 in which several Japanese are kidnapped in Iraq and released. After kidnapping, various opinions this issue have been appeared constantly on the Weblog.

4.5 THE OTHER

This pattern is the otherwise case of the former four patterns. Graphs of the other pattern shown in Fig. 3 are the result of a keyword “Company” (lower graph) and a keyword “Nuclear power plant” (upper graph). This pattern appears when (1) the keyword is general term such as “Weblog” and “Diary”, and (2) the keyword is not a hot topic at this time. Note that the latter type of keywords shows sensitive and gradual increase patterns when those keywords are paid attention by mass media or influential Web site.

5. DISCUSSION

In this paper, we propose a prototype system of an analysis tool for understanding social concerns using Weblog articles. As related works, there are several related projects such as blogPulse³[3] and blogWatcher⁶[4]. The point of this work is to facilitate users to understand social concerns by combining related resources. Thus proposed system finds and provides related news articles with users. News articles are important information for understand social concerns especially in socio-technical domain. We are planning to combine various statistical data available on the Internet so that users can resolve the cause of peaks in daily trend of articles.

⁴ See <http://en.wikipedia.org/wiki/Winny>

⁵ <http://www.blogpulse.com/>

⁶ <http://www.lr.pi.titech.ac.jp/blogwatcher/> (in Japanese)

6. CONCLUSION

We proposed an analysis tool for understanding social concerns using Weblog articles. Understanding social concerns is important for resolving problems in socio-technical domain. We describe requirements of an analysis tool, and implemented. We found five patterns of social concerns. From those social concerns, we can find the importance of a topic. Our future work is to (1) continue to collect articles so that the range of a term over one year, (2) analyze articles qualitatively, and (3) provide related statistical data on the Internet in the analysis result.

REFERENCES

- [1] Baeza-Yates R. & Ribeiro-Neto B. (editors), *Modern Information Retrieval*, Addison & Wesley Longman, 1999.
- [2] Nakagawa H., Mori T., Automatic Term Recognition based on Statistics of Compound Nouns and their Components, *Terminology*, Vol.9 No.2, pp. 201-209, 2003.
- [3] Glance N.S., Hurst M., and Tomikiyo T., BlogPulse: Automated Trend Discovery for Weblogs, *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, NY, U.S., 2004. (available from <http://www.blogpulse.com/www2004-workshop.html>, accessed 2004-05-19)
- [4] Nanno T., Suzuki Y., Fujiki T., and Okumura M., Automatic Collection and Monitoring of Japanese Weblogs, *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, NY, U.S., 2004. (available from <http://www.blogpulse.com/www2004-workshop.html>, accessed 2004-05-19)