LD-006

日本語blogページの自動収集と監視に基づくテキストマイニング
Automatically Collecting, Monitoring, and Mining Japanese Weblogs

奥村 学† 南野 朋之, 藤木稔明, 鈴木泰裕‡
Manabu OKUMURA Tomoyuki NANNO, Toshiaki FUJIKI, Yasuhiro SUZUKI

Abstract

We present a system that tries to automatically collect and monitor Japanese blog collections that include not only ones made with blog softwares but also ones written as normal web pages. Our approach is based on extraction of date expressions and analysis of HTML documents. Our system also extracts and mines useful information from the collected blog pages.

1. Introduction

The reputation of companies and products is now disseminated very quickly on the WWW, because everyone can send a message to the world easily and actively. Therefore, it is highly required to effectively utilize such a vast amount of information disseminated from many people around the world, for the purpose of finding out the reputation of a company or a product, and is so doing quickly coping with people's opinions as a kind of risk management, etc.

One of the information sources that have attracted much attention for these reasons is BBS(Bulletin Board System), and there have been some work that try to monitor it and extract and/or mine from it some useful information, and reflect the people's opinions in product development and company activity[5]. New Internet traffic watchers aim to elevate marketing to a science

Similarly, weblogs(blogs) are now considered as an attractive information source. Although the definition of blogs is not necessarily definite, it is generally understood that they are personal web pages authored by a single individual and made up of a sequence of dated entries of the author's thoughts, a sort of short-term journal, that are arranged chronologically. Blogs tend to be frequently updated and include links to others' blogs. The content and purposes of blogs varies greatly from links and commentary about other web sites, to news about a company/person, to diaries, photos, and so on[1].

It is said that blogs date back to 1996, but they exploded in popularity during 1999 with the emergence of blogger(`http://www.blogger.com`) and other easy-to-use publishing tools[4]. Most recently in 2002, a Newsweek article appeared estimating the number of weblogs to be half a million[4].

Recently, the study on analyzing the space of weblogs has become a hot topic, as the blogspace has began to exhibit explosive growth [4]. There are some sites that provide blog search services, by collecting and watching blogs, such as Technorati[2], Blogdex[3]. and Daypop[4].[5] These services are all based on the RSS(RDF Rich Site Summary) and/or the ping server for collecting and watching blogs. RSS is an extensible metadata description and syndication format, currently used for a number of applications, including news and other headline syndication, weblog syndication, etc [6]. The ping server is a mechanism to tell that a weblog has changed and is used for tracking updates to weblogs[7]. Therefore, collection and watching of blogs is considered to be relatively easy.

In Japan, however, since long before blog softwares became available, people have written 'diaries' on the web(called "web diaries"), that are quite similar to blogs in the content, and people still write them without any blog softwares. Therefore, (as we will show,) hand-edited collections of blogs are quite numerous in Japan, though most people now think of blogs as pages usually published using one of the variants of public-domain blog software, such as 'Movable Type'[7]. Therefore, it is quite difficult to exhaustively collect Japanense blogs, i.e., collect ones made with blog softwares and web diaries written as normal web pages.

With this as the motivation for our work, we present a system that tries to automatically collect and monitor Japanese blog collections of that include not only ones made with blog softwares but also ones written as normal web pages. Our approach is based on extraction of date expressions and analysis of HTML documents, to avoid having to depend on specific blog softwares, RSS, or the ping server. Furthermore, our system also extracts and mines useful information from the collected blog pages.

2. Architecture of Our System

The architecture of our system is shown in Figure 1. The 'Collection' module tries to collect candidate blog pages from the WWW. This is done 1) by crawling the

†東京工業大学 精密工学研究所
‡東京工業大学 総合理工学研究科 知能システム科学専攻

[1] `http://new.blogger.com/`

[2] `http://www.technorati.com/`

[3] `http://blogdex.media.mit.edu/`

[4] `http://www.daypop.com/`

[5] In `http://www.aripaparo.com/archive/000632.html` you can find a useful list of blog search engines.

[6] `http://www.oasis-open.org/cover/rss.html`

[7] `http://www.movabletype.org`

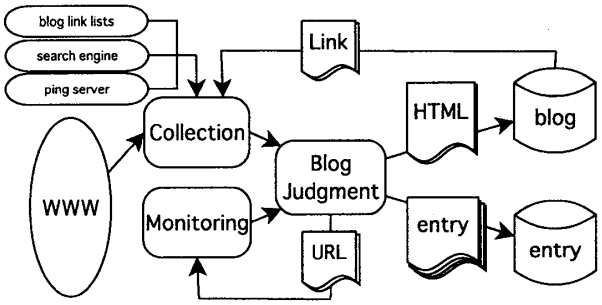

Figure 1: System Architecture

WWW, 2) by using the link lists of blog sites and the information from the ping server, and 3) from the links of the pages judged as blogs in the judgment module.

The 'Blog Judgment' module tries to select only the blog pages in the collection, by using the characteristics of the blog pages. A page is judged to be a blog if and only if a sequence of entries that are articles for a day can be extracted from the page. The entries should satisfy the following constraints:

1. Entries should contain a date expression, and it should be at the top of the entry.

2. The date expressions of the sequence of entries should be consistently formatted[8], and be arranged in ascending/descending order.

3. The tag sequence should be uniform for all the entries in the sequence.

The 'Monitoring' module tries to periodically watch the blog pages that the judgment module selected, and extract the updated entries in the pages.

The process of blog judgment is as follows. First, we automatically extract date expressions in web pages and annotate them with ⟨date⟩ and ⟨/date⟩ tags, by using the manually written regular expressions. If the date expressions lack year and/or month information, we try to supplement them by using heuristics, such as the nearest year(month) that appears before the expression.

Next, the sequence of entries for a day is tried to be extracted from a page. Even if such a sequence is extracted from a page, the page might not be a blog, because archives of BBSs, chats, and mailing lists, update description on a site page, and announcements of events can also include date expressions. To remove such non-blog pages, we apply heuristics to the collection, that take into account the following:

1. typical keywords included in non-blog pages, such as 'bbs,' 'chat,' 'reply,' and 're:'.

2. entries with future date expressions, which might indicate announcements.

3. whether the time between two adjacent entries is within a month[9].

4. whether at least one predicate(verb, adjective, etc.) is contained in an entry, since blogs and diaries should include descriptions on an event.

The collected blog entries can be retrieved by using a text search engine called GETA[2]. The entries can be ranked by the number of inbound links, interval of updates, freshness(recency), and their size.

We also try to mine the collected blog entries. Trend analysis is performed in the collection. Degree of 'burstiness' is computed for words, and a 'hot' keywords list that reflects the popular topics in a certain period is generated from the collection.

The 'burst of a word' is a phrase to indicate a sharp rise in frequency as the topic emerges[3], and originally in Kleinberg's work, targeting E-mail and research papers, his text mining algorithm tried to identify bursts in document streams in which messages arrive in temporal order. Simply put, a burst can be found by searching periods when the word tends to appear at shorter intervals than usual.

Unfortunately, as Kleinberg's original algorithm cannot be applied to blogs, we extended it[3] so it could discover dense periods of 'burstiness' in the blogs. The reason why the original algorithm cannot be applied to blogs is that as Figure 2 indicates, since the distribution of the blog entries is not uniform, the interval at which entries arrive tends to be shorter in the period when more blog entries arrive, and consequently, more bursts tend to be erroneously identified.

Subjective(evaluative) expressions in blog entries can also be automatically annotated and highlighted[6] in our system. The automatic annotation is done by using a naive pattern matching algorithm with a manually constructed dictionary of evaluative expressions.

3. Current Status and Evaluation

We have operated our system for two weeks since the end of December 2003, and obtained 39,272 blogs(pages) and 466,809 entries. Figure 2 shows the distribution of the dates for the collected entries. Our system can obtain rather old entries, whereas the collection methods based on RSS and/or the ping server can obtain only recent entries.

300 blogs randomly sampled from the collection were manually evaluated, and we found that 283 were actually blogs(94.3%) and 17(5.7%) were not, although in 24 of the correct ones extraction of entries was not completely correct.

[8]'2003/1/2' and '2-Jan-2003' are considered as inconsistent.

[9]Blogs and diaries will likely be updated more frequently.

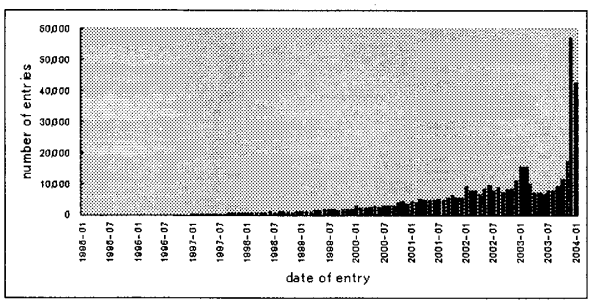

Figure 2: Distribution of the collected blog entries

By using the perl module that tries to detect what kind of blog creation tool was used for its creation[1](we patched it for Japanese), we found that of 39,272 pages, 32,219(82.0%) were judged as not using any blog tools and that the most popular blog tool was Movable Type(2,243;5.7%).

We now give two examples of the bursty phenomena obtained by our algorithm. For the word 'Olympics', we could find three periods when bursts occurred: February of 1998(at Nagano), September of 2000(at Sydney), and February of 2002(at Salt Lake City). As for the hot keywords list, for June of 2002 we could obtain a list that included 'World Cup,' 'Turkey,' 'England,' and 'Beckham,' reflecting that the World Cup soccer championships were held in Japan and Korea during that period.

At the conference site, we will show a demonstration of our latest system and discuss its latest status(the number of collected blogs, etc.).

References

[1] M. Ceglowski. Www::blog::identify - identify blogging tools based on url and content. `http://search.cpan.org/~mceglows/WWW-Blog-Identify-0.06/Identify.pm`, 2003.

[2] IPA(Information-technology Promotion Agency, Japan). Generic engine for transposable association: Geta. `http://geta.ex.nii.ac.jp/`, 2002.

[3] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1–25, 2002.

[4] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of the 12th International World Wide Web Conference*, pages 568–576, 2003.

[5] J. Wakefield. Catching a buzz. *Scientific American*, 2001. November.

[6] J. Wiebe, E. Breck, C. Buckley, and C. Cardie. Recognizing and organizing opinions expressed in the world press. In *Proc. of the 2003 AAAI Spring Symposium New Directions in Question Answering*, pages 12–19, 2003. Technical Report SS-03-07.

[7] D. Winer. Weblogs.com xml-rpc interface. `http://www.xmlrpc.com/weblogsCom`, 2001.