

An Efficient Method Predicting Update Probability on Blogs

BUMSUK LEE and BYUNG-YEON HWANG*

Department of Computer Science and Engineering

The Catholic University of Korea

43-1 Yeokgok 2-dong, Wonmi-gu, Bucheon-si, Gyeonggi-do 420-743

REPUBLIC OF KOREA

{bslee, byhwang}@catholic.ac.kr

Abstract: - In this study, we assumed that there is a specific pattern to when a blogger is actively posting, in terms of days of the week and, more specifically, hours of the day. We analyzed 15,119 blogs to determine a blogger's posting preference. This paper proposes a method to predict the update probability based on a blogger's posting history and preferred days of the week. We applied this method to 12,115 blogs to check the precision of our predictions. The evaluation shows that the model has a precision of 0.5 for over 93.06% of the blogs examined.

Key-words: - Web 2.0, Blog, RSS, Predicting Method, Update Probability, Analysis

1. Introduction

Web 2.0 technologies have emerged on the Internet over the last several years, and important issues have come to the fore, such as growing social networks and personal media [1]. Social networks, represented by such successful sites as Facebook.com, not only connect people directly, but also provide new services that use the network as their platform. Many of these sites offer web services such as file sharing and instant messaging, or even network-connected desktop applications.

Personal media has grown rapidly following the appearance of professional blogging tools. Until a few years ago, the maintenance of a personal webpage required basic knowledge of HTML, but today such easily available websites as WordPress.com or Blogger.com allow anyone to author a personal blog, and no specialized computer knowledge is required. It is quite easy to create articles on a personal blog and share them via searching services such as Google.com or Technorati.com. Originally, bloggers discussed their personal lives or lifestyle issues, but today many blogs specialize in professional issues such as the economy or politics. Many people believe that blogs are as influential as the traditional media.

According to Technorati's "state of the blogosphere 2008," 94.1 million people, or 50% of the Internet users in the United States, read blogs in May 2008, and 22.6 million people (12%) have their

own blog. It is estimated that 77% of active Internet users read blogs [2].

RSS plays a key role in blog services [3,4] and is one of the most successful XML services ever [5]. RSS is a technique for notifying subscribers that new content has been posted, and the subscriber does not need to visit the website or blog. RSS reader applications operate on the user's desktop, or are offered as Web services. Users add the URL of the RSS to the application, which periodically checks for updates and notifies the readers.

Currently, new services are emerging that are offering RSS reader tools like a portal service, referred to as a meta-blog. A meta-blog gathers RSS from blogs by operating crawlers or by inducing people to add RSSs to their own blogs. The meta-blog periodically checks the collected blogs for new content, and then indexes the contents so that it can respond to user queries [6].

The development of an efficient update manager is urgently required, because the contents of an RSS feed are continuously changing [7,8]. Meta-blogs can check for updates every ten minutes, every hour, or at any specified static interval. Some meta-blogs classify blogs according to the update frequency, and they can check frequently updated blogs at a different time interval than those updated less frequently. Checking for updates too frequently results in unnecessary overheads, so new methods for predicting updates to an RSS feed are necessary. Blog postings have a particular pattern unique to each

blogger's activities, and we expect that it is possible to predict a blog update by analyzing the days of the week and hours of the day that the blogger actively posts new content.

The aims of this paper are two-fold: (1) to analyze blog postings to determine specific posting patterns and (2) to evaluate a heuristic method that predicts the update probability based on the blogger's posting pattern and history. The main contributions of this paper are the analysis results and evaluation. We found that most bloggers prefer certain days of the week and certain hours of the day. The proposed update prediction method has a high precision, as demonstrated by the evaluation of real-world blogs.

The rest of this paper is composed as follows. Section 2 of this paper provides a brief overview of existing studies that analyze the characteristics of RSS feeds and that discuss the problem of blog update checking. Section 3 describes the method for analyzing of blog postings that we used to find the preferred days of the week and hours of the day for blog updates. Section 4 reports the results of applying this method to predict the update probability by using real-world data. Finally, Section 5 contains the conclusions and a discussion of future studies.

2. Related works

2.1 Characteristics of an RSS feed

Liu, et al [9] analyzed client behavior and feed characteristics of RSS. They collected snapshots of RSS content by actively polling every hour 99,714 feeds listed in the feed directory syndic8.com, and used them to analyze updates in terms of update rate and amount of change. There were two notable results; one was that the feed update rates consisted of two extremes: either very frequent or very rare. More than 55% of feeds were updated in the first hour, while 25% of feeds were not updated during the entire polling period. The second result shows the average update time. In total, 57% of the feeds had an average update interval of less than two hours, while 25% of the feeds remained the same over three days. These results indicate that the polling periods for RSS readers should depend on the feeds, and that a meta-blog needs an update manager that checks each blog at a different time interval. In other words, checking each blog for updates every ten minutes or once an hour might be unnecessary.

2.2 Problems of checking for an update to an RSS feed

An RSS feed is continuously updated and meta-blogs need to reflect newly updated contents in their search results. Consequently, meta-blogs check for updates to all of the stored blog RSS feeds at a particular time interval. A shorter interval is better for collecting new content of a critical nature. Rapid collection means that a meta-blog can reflect rapidly changing issues in its search results, which improves user satisfaction. However, checking for updates too frequently results in unnecessary overheads. For example, it is not necessary to check for updates more than once an hour when a blog has only one posting a day. If there were millions of blogs in the checking list, the inefficiency in system resource usage would be significant. Accordingly, a meta-blog requires an adequate time interval for checking for updates.

Every blogger could have a unique posting pattern, as explained in Section 2.1. Therefore, a meta-blog can reduce checking overheads by applying an adaptive checking method based on the blogger's update patterns. The meta-blog contains an update-checking module, and some meta-blogs group blogs based on the frequency of updates. Most meta-blogs have a multi-thread type of update-checking module in a distributed environment, which is used to reduce overheads. High system overhead sometimes results in delayed responses to user requests. To resolve this problem, this paper proposes a heuristic method that predicts updated content on a blog.

3. Analysis of day and time preferences

In this section, we assumed that each blogger has a unique update pattern. We therefore analyzed the days of the week and the hours of the day that the blogs are most likely to be updated.

3.1 Dataset

We obtained a list of 28,881 RSS feeds by implementing an RSS crawler, and selected 15,119 feeds that conformed to RSS 2.0 specifications. The dataset was gathered over the course of four weeks, but only a two-week dataset with a stable collection was analyzed in the experiment.

3.2 Preferences: Days of the week and hours of the day

Fig. 1 and Fig. 2 describe the number of posts during the two weeks, and the number of posts during each hour of the day, respectively. Fig. 1 shows a very similar graph with a statistics chart of Internet usage according to days of the week [10]. The Monday to Wednesday period had twice the number of posts as the Friday to Sunday period.

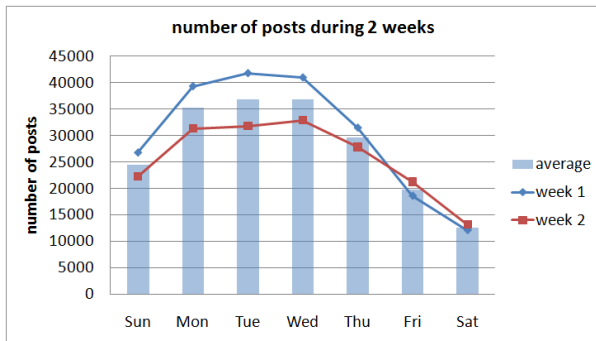


Fig. 1 Number of posts during two weeks

Fig. 2 shows the number of posts over 24 hours, and the post count corresponds to the daily cycles of modern activities. The number of posts decreases from 1 am to 9 am, and then increases from 9 am, when people are generally using their computers at work. Similarly, it increases by about ten thousand immediately after 2 pm and by about five thousand immediately after 5 pm. This indicates that bloggers tend to write posts either after lunch or after returning home. The increase in the number of posts grows until 1am, when Internet usage peaks. The most postings in a day are from midnight to 1am.

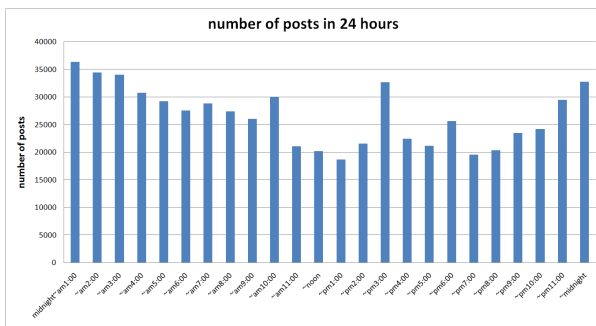


Fig. 2 Number of posts in 24 hours

An adaptive update manager that searches for updates based on the day of the week and on the time of day might operate effectively in real-world conditions. The application of the proposed method to each blog differs according to its update status, so it is necessary to analyze each blog. The rest of this

section is dedicated to the analysis of the top-five blogs, based on the number of articles, and four mid-level groups.

Fig. 3-1 describes the analysis of the top-five blogs. The average of these five blogs is similar to Fig. 1, but each blog showed a more explicit preference for certain days of the week. We composed four mid-level groups based on the number of posts, and selected five blogs in each group: blogs with over 100 posts (post-100), and blogs with 50 (post-50), 20 (post-20) and 10 (post-10). The analysis results are shown in Fig. 3-2. These groups also show a preference for certain days of the week, especially a blog of the post-100 group that had a strong preference for Sunday. This RSS feed was a podcast of a German Internet broadcasting service.

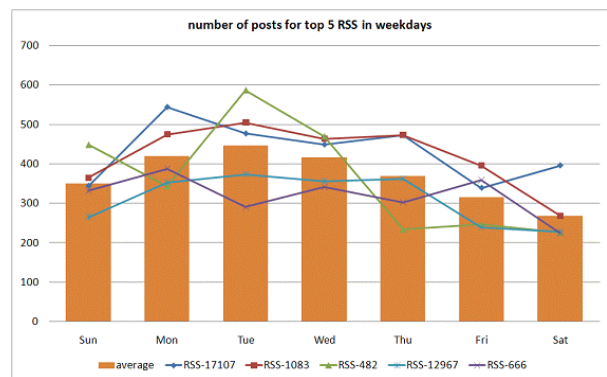


Fig. 3-1 Number of posts for top-5 blogs in weekdays

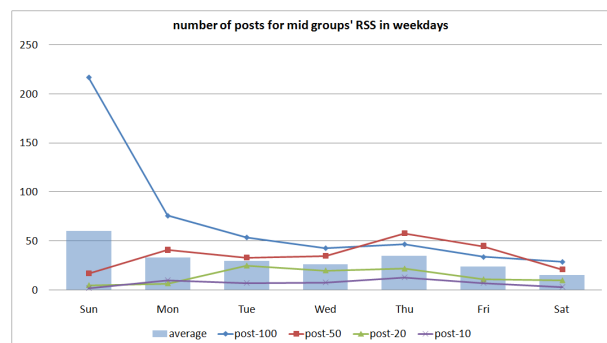


Fig. 3-2 Number of posts for mid-level blogs in weekdays

Fig. 4-1 and Fig. 4-2 show the number of postings in 24 hours for the aforementioned blogs. Each blog had a strong time preference that was higher than the average. RSS-482 in Fig. 4-1, which included movie information, showed a highly explicit preference: there were no postings at all from 6 am to 11 am. Fig.

4-2 shows that the time of day preferences in the mid-level group were also specific.

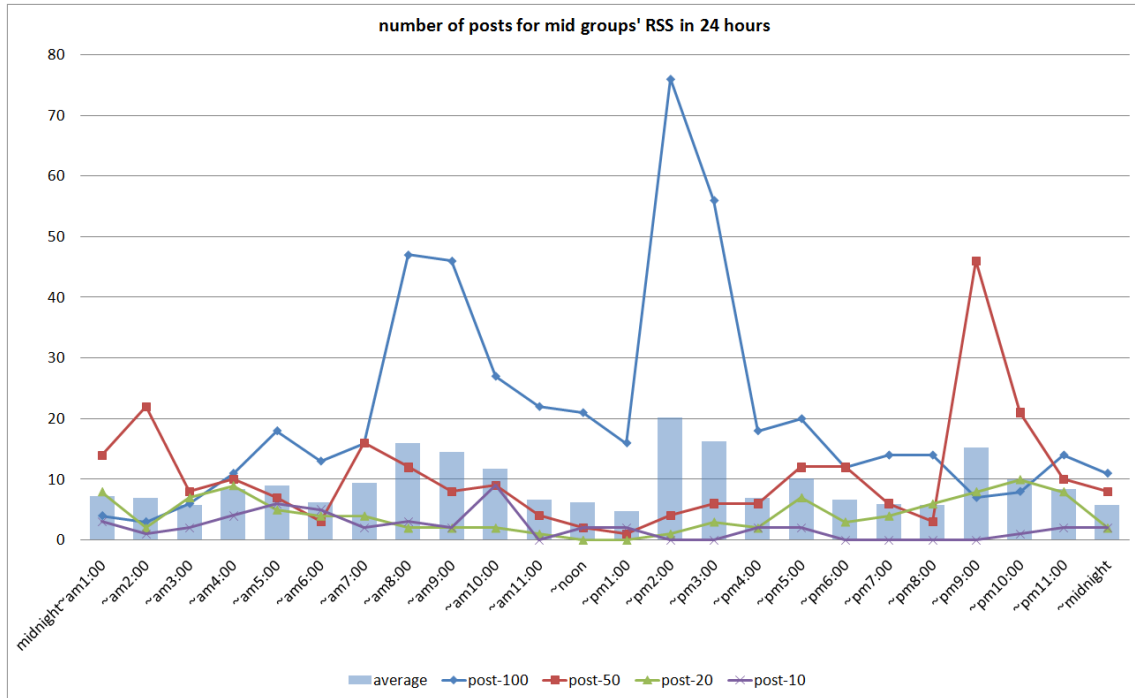


Fig. 4-1 Number of posts for top-5 blogs according to time

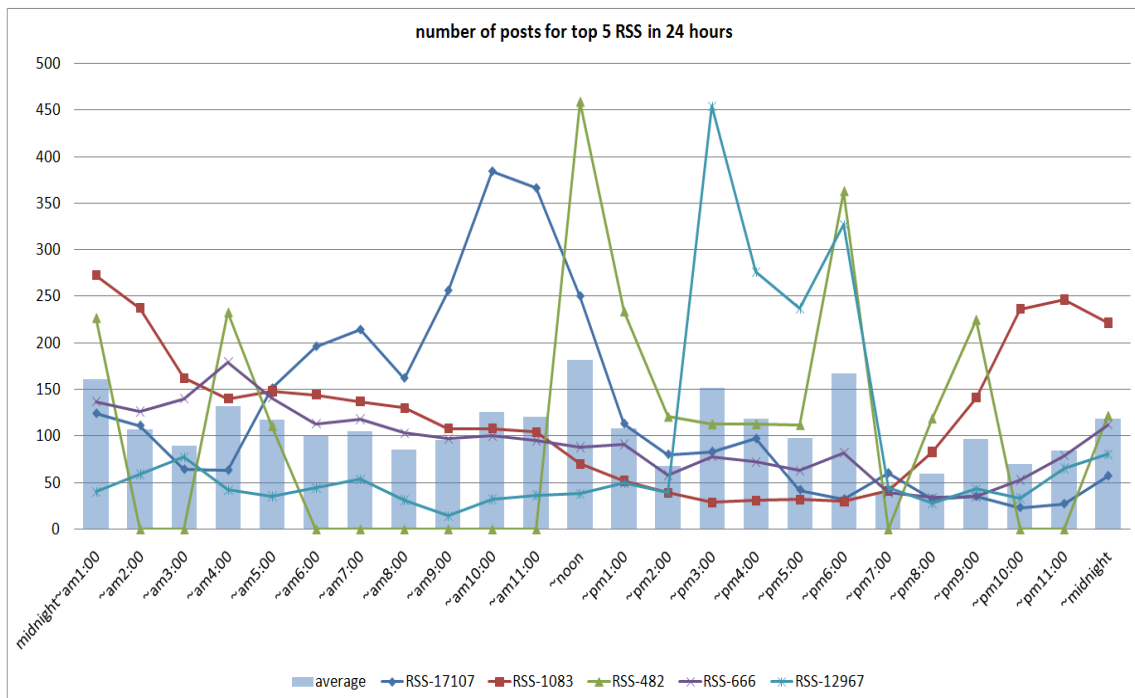


Fig. 4-2 Number of posts for mid-level blogs according to time

4. Update prediction and evaluation

4.1 Update prediction

This paper proposes a method to predict the update probability for the following day by analyzing the gathered RSS feeds. We define the update possibility P .

Definition 1. Update possibility P is calculated by using statistics about the days of the week and the posting history. The weight of two factors is regulated by λ .

Definition 1 can be represented by the following equation: $P = \lambda(\text{weekdaysupdated} / \text{weekdayswhole}) + (1 - \lambda)(\text{daysupdated} / \text{dayswhole})$. The update checking module operates when P exceeds the threshold θ , which implies that there is a strong probability of an update on that day. Finding suitable λ and θ values is necessary to improve the precision of the prediction, but we set $\lambda = 0.9$ to assign weight to the weekday statistics, and set $\theta = 0.5$.

4.2 Data selection

The evaluation was designed to measure the precision of the prediction. The prediction requires more than one week of data, because the probability can only be calculated based on the posting history. The selected blogs had an interval of more than seven days between the oldest and newest data. We excluded blogs that had more than four articles per day on average. Predicting blogs with such a high update rate might not be needed, because it is likely that they will be updated each day regardless. In total, 12,115 blogs were selected for evaluation based on the aforementioned conditions.

After this work, data are represented by either a 1 or a 0. It is possible to achieve a prediction precision similar to the example shown in Fig. 5. The table describes the predicted update probabilities versus the real updates. The red cells indicate an incorrect prediction, while the value in parentheses is the real update. The measurement of precision started from November 29th in this example.

Sat.	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.
1	1	1	1	1	1	0
1	1(0)	1(0)	1	1(0)	1	0
1	1(0)	1	1(0)	1	1	0(1)
1	0	1	1			

$\lambda = 0.9$ $\theta = 0.5$

Fig. 5 Result table of precision measurement

For example, the update possibility P on December 7th was calculated at 0.517, as shown below, which exceeds the threshold. The update-checking module operates, but there was no update on that day (i.e., this was an incorrect prediction, $P = 0.9(1/2) + (1 - 0.9)(10/15) = 0.517$). The update probability P for the next day was 0.5125, which also exceeded the threshold. There was an updated article on that day (i.e., this was a correct prediction, $P = 0.9(1/2) + (1 - 0.9)(10/16) = 0.5125$). The precision is defined as follows:

Definition 2. Precision $Prec.$ is the number of days there was a correct prediction over the total predicted days. $Prec. = \frac{\text{the number of correct predictions}}{\text{the number of predicted days}}$.

If the predictions were correct over the entire period, then the precision would be one, whereas the precision would be zero if there were no correct predictions. The example in Fig. 5 has a precision of 0.66; there were twelve days of correct predictions out of a total of eighteen days. The precision was calculated for each of the 12,115 blogs, and the average precision was 0.76. Fig. 6 shows the blog distribution chart for the precision calculations. Most of the studied blogs had a precision between 0.8 and 0.9, as shown in Fig. 6. About 93.06% of the blogs had a precision above 0.5.

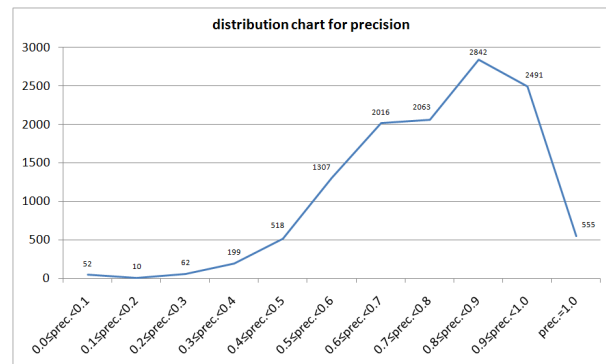


Fig. 6 A distribution chart for precision

5. Conclusion

This paper analyzed postings from 15,119 blogs to determine the preferred days of the week and hours of the day for content updates. Additionally, we proposed a method to predict updates, and we evaluated the precision of this method by using real-world blogs. The average prediction precision for the 12,115 blogs was 0.76, and 93.06% of the blogs had a precision above 0.5. The prediction method

should be applied over a longer period of time than in this paper. There should be studies on suitable λ and θ values, to improve the precision of the predictions. This paper proposed a method to perform update prediction over a period of days, and we expect that refinements will result in improved efficiency.

References

- [1] T. O'reilly, "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *Communications & Strategies*, No. 65, pp. 17-37, 1st Quarter, 2007.
- [2] Technorati, "State of the Blogosphere," <http://www.technorati.com/blogging/state-of-the-blogosphere/>, 2008.
- [3] <http://en.wikipedia.org/wiki/RSS>, 2008.
- [4] Berkman Center, "RSS 2.0 at Harvard Law," <http://cyber.law.harvard.edu/rss/index.html>, 2008.
- [5] M. Olson and U. Oqbuji, "The Python Web services developer: RSS for Python," <http://www.ibm.com/developerworks/webservices/library/ws-pyth11.html>, November 2002.
- [6] X. Li, J. Yan, Z. Deng, L. Ji, W. Fan, B. Zhang, and Z. Chen, "A novel clustering-based RSS aggregator," *In Proc. of 16th Int'l Conf. on WWW*, pp. 1309-1310, 2007.
- [7] K. C. Sia, J. Cho, and H. K. Cho, "Efficient Monitoring Algorithm for Fast News Alerts," *Knowledge and Data Engineering*, IEEE Transactions on Vol. 19, Issue 7, pp. 950-961, July 2007.
- [8] B. Lee, J. W. Im, B. Hwang, D. Zhang "Design of An RSS Crawler with Adaptive Revisit Manager," *In Proc. of the 20th Int'l Conf. on Software Engineering and Knowledge Engineering*, pp. 219-222, July 2008.
- [9] H. Liu, V. Ramasubramanian, and E. G. Sirer, "Client behavior and feed characteristics of rss, a publish-subscribe system for web micronews," *In Proc. of the ACM Internet Measurement Conference*, 2005.
- [10] <http://www.thecounter.com>, 2008.