AAAI Technical Report WS-12-03
Social Media Visualization

A Temporal Analysis of Posting Behavior in Social Media Streams

Bumsuk Lee
Database Systems Laboratory,
The Catholic University of Korea
bumsuk@catholic.ac.kr

Abstract

In this work, we investigated the social media streams to understand their characteristics and their temporal aspects. We assumed that each blogger has different temporal preference for posting. To investigate this hypothesis, we analyzed a massive dataset, nearly 700,000 blog articles, with the consideration of two factors which are day of the week and time of the day. The comparison was done in manifold ways: Blogosphere vs. Twitter, commercial blogs vs. non commercial blogs, and their individuals. We hope that this work provides a hint to develop a personalized system which can be used for the reduction of the system resources for pull/fetch technology.

1. Introduction

In recent years, we have observed huge success in micro-blog services like Twitter, and many researchers have done their analysis on Twitter (Kwak et al. 2010). However, the traditional blog services are still popular. According to the latest report from eMarketer.com, there is an increase in the number of bloggers. In 2010, 112.7 million people (51% of the Internet users in the United States) read blogs and 26.2 million people had their own active blog. The "State of the Blogosphere 2011" report from Technorati.com showed that 82% of the bloggers who were surveyed were using Twitter but 77% of them used Twitter to promote their blogs. These facts show that the Blogosphere is an attractive resource with plenty of beneficial information.

The blog search engines work similar to the Web search engines, but they have to check for updates from the collected blogs more frequently to provide up-to-date information (Lee et al. 2008). However, the report from Technorati shows only 11% of bloggers post daily. Thus, an aggressive strategy for checking updates not only wastes the system and network resources, but also threatens the stability of the blog server as repeatedly crawling the server from multiple locations is quiet similar to the distributed denial of service attack (Sia et al. 2007).

In this work, we present the temporal aspects of posting behavior on social media streams to understand their characteristics. Analysis results show that each blogger has their own preference for the time of posting, in terms of the day of the week and also, the time of the day. This gives a hint for developing an alternative pull/fetch technology with optimal pulling period for each feed instead of a single common polling rate for all feeds.

This paper is organized as follows: In the next section, we explain the previous studies on social streams which are helpful to remind you about our research. In section 3, we present our analysis results, and finally, we discuss the results and conclude the paper in Section 4.

2. Related Work

Analysis of Social Streams. Analysis results for the social media streams have been published before. BlogScope is an information discovery and text analysis system that offers a set of unique features, including spatio-temporal analysis, flexible navigation, keyword correlations and ranking functions for query (Bansal and Koudas 2007). Chi et al. proposed a novel technique that captures the structure and temporal dynamics of blog communities (Chi et al. 2007). Leskovec et al. analyzed the largest blog datasets, with 45,000 blogs and 2.2 million blog-postings, to discover the patterns of information propagation in blogosphere (Leskovec et al. 2007). In another paper, they observed how blogs behave over time. The time behavior of blogs is bursty (McGlohon et al. 2007).

While these works focused on blogs, recent works (Ahn et al. 2010; Yang and Leskovec 2011) are focusing on Twitter and its visualization. Although there are many studies on the social streams, our study is more concerned with bloggers' behavior on temporal bias.

Feed and Pull Technology. There are more related works to ours. A research group analyzed client behavior and feed characteristics (Liu, Ramasubramanian, and Sirer 2005). They collected snapshots of RSS content by actively polling every hour from 99,714 feeds listed in the feed directory on syndic8.com. Update rate and amount of change were investigated, and they found out a notable fact that the feed update rates exhibited two extremes: either very frequent or very rare. More than 55% of recently updated feeds were updated again within the first hour while 25% feeds were not updated during the entire polling period. They argued that significant bandwidth savings can be obtained by using the optimal pulling period for each feed instead of a single common polling rate for all feeds. This argument is still effective because many applications on the smart devices use pull/fetch technology.

Later, Hmedeh et al. characterized Web syndication behavior and content (Hmedeh et al. 2011). They analyzed three factors: publication activity, items structure and length; vocabulary of RSS content. Several solutions have already been proposed to the pull technology (Adam, Bouras, and Poulopoulos 2010; Lee and Hwang 2009; Han et al. 2008; Bright, Gal, and Raschid 2006). These papers proposed adaptive methods based on the pull technology. In addition to that, Urbansky et al. presented adaptive feed polling algorithms which learn from the previous behaviors of feeds and predict their future behavior (Urbansky et al. 2011). Our analysis result could be applied all these research to improve their performance.

3. Temporal Activity Analysis

Dataset and General Characteristics. For this study, a massive dataset, nearly 700,000 blog articles, from 15,000 blogs were gathered between November 28, 2008 and December 18, 2008. As the feed service is based on the pull technology, we had to check for updates of the blog feeds as frequently as we can in order to avoid missing any content. For our first analysis, we observed the temporal aspect of all datasets. Figure 1 describes the average number of posts by the day of the week and by the time of the day. Tuesday is the most popular day for posting and on Tuesday a blogger posted about 3 articles on average. In 24 hours, most of the posting activities took place from 10 p.m. to 2 a.m., but we can see a convex at 2 p.m.

Although we used the raw data for Figure 1, in order to see the average numbers of posts, the rest of this paper use the normalized percentage of the posting activity.

Blogosphere vs. Twitter. We compared our data with Twitter (Cheng et al. 2009) as Twitter is one of the most popular micro-blog services. The posting activities on the Blogosphere and on Twitter are described in Figure 2.

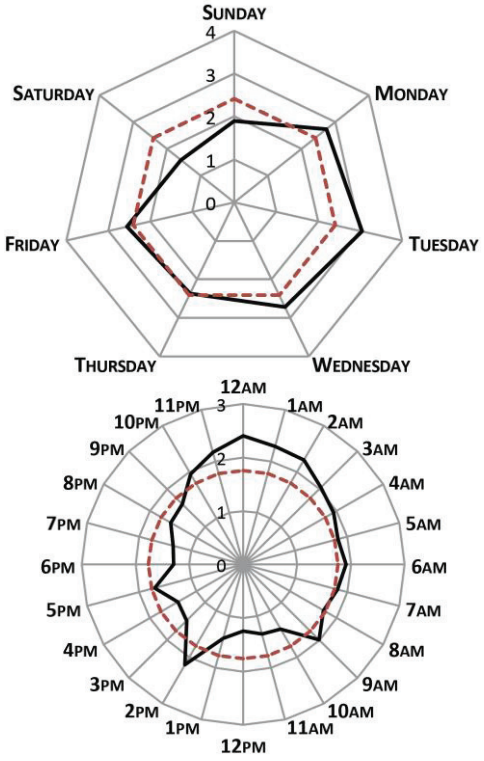

Figure 1: Posting activities; solid line denotes the average numbers of posts by the day of week and by the time of day; dotted line denotes the average of the total data.

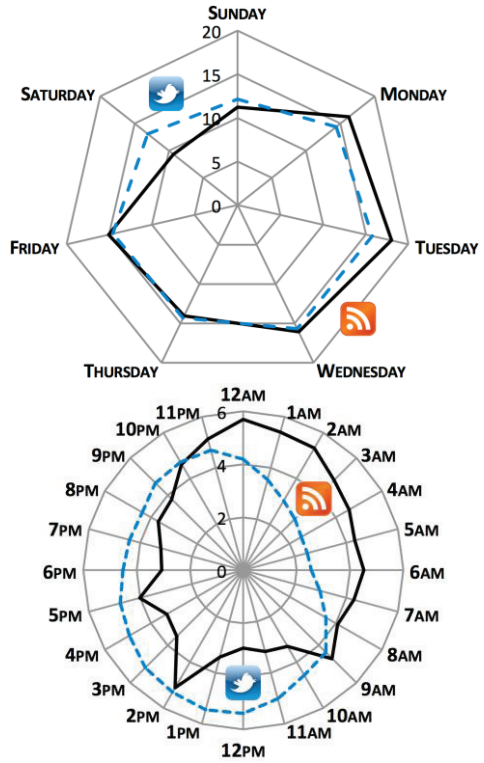

Figure 2: Comparison between Blogosphere (solid line) and Twitter (dotted line).

We observed an interesting fact from this comparison. For the day of the week, the overall aspects of two datasets were similar, but the graph for the Blogosphere was more dynamic. The standard deviation of each dataset was calculated to see their dynamics. They were 2.81 for the Blogosphere and 1.21 for Twitter. The difference between the two data was observed more explicitly on the time of the day. As we aforementioned, many bloggers posted their articles from 10 p.m. to 2 a.m. while tweets were mostly generated from 11 a.m. to 3 p.m. This aspect seems to be reasonable as the length of a tweet is relatively short and users can post their tweet easily with their smart devices from anywhere and at any time. To the contrary, writing a blog article usually takes more time as it is longer than a tweet in many cases and bloggers use their computers for writing articles instead of using their smart devices.

Commercial Blogs vs. Non-commercial Blogs. In this subsection, we compared the feeds from commercial blogs and from non-commercial blogs. We categorized the blogs into two groups based on the keywords present in the title and in the description. We investigated 382 personal blogs and 1,983 blogs for this analysis. News channels, radio stations, film distributors, and libraries were included in the group of commercial blogs. Table 1 shows standard deviation of the normalized percentage of the posting activity. Commercial blogs have stronger preference for the day of posting compared to the non-commercial blogs, but an opposite aspect was observed for preferred posting time of the day.

Table 1: Standard deviation of the normalized percentage.

	Commercial blogs	Non commercial blogs
Day of the week	5.25	4.41
Time of the day	1.09	1.34

The rest of this subsection is dedicated to the analysis of the individual feeds and to understand their characteristics in a more detailed manner. Each blog has a strong preference for day and time. Figure 3 illustrates the posting activity of two individual blogs which are commercial. One is a radio station blog and the other is a library blog. The radio station posted their articles mainly on Monday and Friday in the evening while the library preferred Wednesday afternoon for their posting activity. Both blogs did not post any articles during the weekends. The posting time of the library shows a more explicit temporal preference as it does not have any articles from 6 p.m. to 9 a.m. at all.

Figure 4 depicts the posting activities of two non-commercial blogs. Although each blogger has a different preference for posting, the temporal preferences of non-commercial blogs are relatively weak. However, the posting

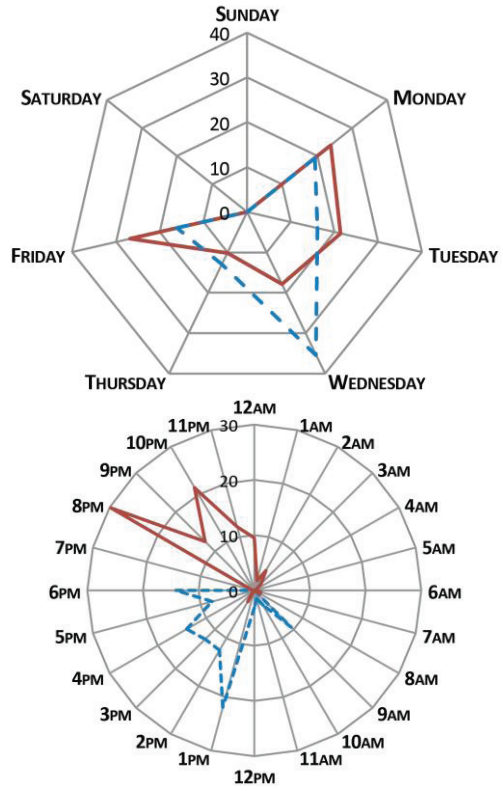

Figure 3: Posting activity of two commercial blogs; solid line is for a radio station and dotted line is for a library.

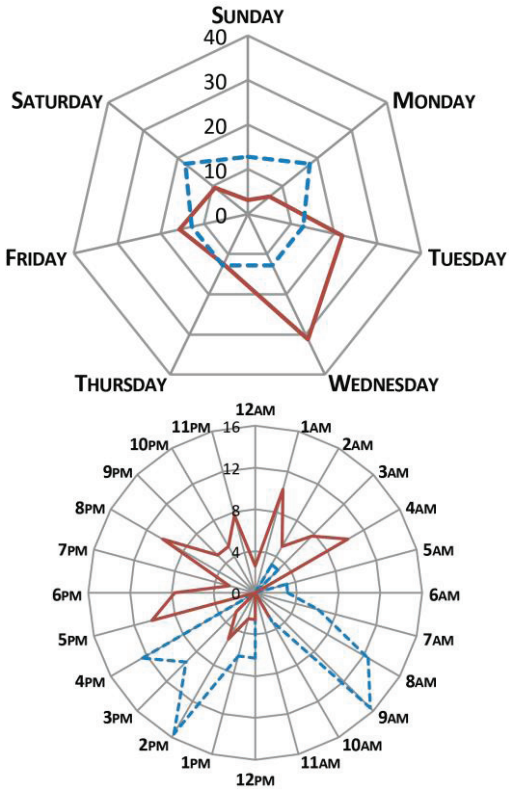

Figure 4: Posting activity of two non-commercial blogs.

time of the non-commercial blogs were widely distributed over time. We can infer from Figure 4 that non-commercial bloggers post their articles whenever they want but commercial bloggers publish their blog articles mainly during their working hours.

4. Conclusion

We investigated and analyzed the social media streams to understand their characteristics and temporal aspects. The displayed results throughout the paper show several insights into the posting behaviors of bloggers. Briefly, each blogger has a different temporal preference for posting. Before the start of this analysis, we expected to find out a hint to implement a personalized scheduler for checking the updates of feeds. As the pull technology or the fetch technology is widely used, the personalized scheduler can be applied to the various systems, including RSS reader and smart device applications. We hope that this work provides the foundation for developing a personalized system that can be used for the reduction of the system resources for pull/fetch technology. Based on the lessons learned, we have been implementing a prototype of the RSS reader that includes checking interval scheduler for each feed. The scheduler estimates the probability of the update existence and moreover based on the day of week and the recent posting activity. The algorithm for prediction was evaluated and the precision was 66 per cent. Its performance was fair as it could reduce the overhead of servers. The algorithm still has to be improved particularly in terms of the posting time of day and should also be evaluated on a larger data set.

References

Adam, G., Bouras, C., and Poulopoulos, V. 2010. Efficient Extraction of News Articles based on RSS Crawling, In *Proceedings of the International Conference on Machine and Web Intelligence*.

Ahn, J., Taieb Maimon, M., Sopan, A., Plaisant, C., Shneiderman, B. 2010. Temporal Visualization of Social Network Dynamics: Prototypes for Nation of Neighbors. In *Proceedings of the 4th International Conference on Social Computing, Behavioral Cultural Modeling and Prediction*.

Bansal, N. and Koudas, N. 2007. BlogScope: Spatio temporal Analysis of the Blogosphere. In *Proceedings of the 16th International Conference on World Wide Web*, 1269 1270.

Bright, L., Gal, A, and Raschid, L. 2006. Adaptive Pull based Policies for Wide Area Data Delivery, *ACM Transactions on Database Systems*, Vol. 31, No. 2, 631 671.

Cheng, A. and Evans, M. 2009. Inside Twitter: An In Depth Look Inside the Twitter World. http://www.sysomos.com/insidetwitter/, Sysomos Inc.

Chi, Y., Zhu, S., Song, X., Tatemura, J., and Tseng, B. L. 2007. Structural and Temporal Analysis of the Blogosphere through Community Factorization. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 163 172.

Han, Y. G., Lee, S. H., Kim, J. H., and Kim, Y. 2008. A New Aggregation Policy for RSS Services, In *Proceedings of the International Workshop on Context Enabled Source and Service Selection, Integration and Adaptation*, 1 7.

Hmedeh, Z., Vouzoukidou, N., Travers, N., Christophides, V., Mouza, C. D., and Scholl, M. 2011. Characterizing Web Syndication Behavior and Content, In *Proceedings of the 12th International Conference on Web Information Systems Engineering*, 29 42.

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*. 591 600.

Lee, B., Im, J., Hwang, B. Y., and Zhang, D. 2008. Design of An RSS Crawler with Adaptive Revisit Manager. In *Proceedings of the 20th International Conference on Software Engineering and Knowledge Engineering*. 219 222.

Lee, B. and Hwang, B. Y. 2009. An Efficient Method Predicting Update Probability on Blogs, In *Proceedings of the 2nd WSEAS International Conference on Engineering Mechanics, Structures and Engineering Geology*, 208 213.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., and Hurst, M. 2007. Cascading Behavior in Large Blog Graphs. In *Proceedings of the 7th SIAM International Conference on Data Mining*.

Liu, H, Ramasubramanian, V., and Sirer, E. G. 2005. Client Behavior and Feed Characteristics of RSS, A Publish Subscribe System for Web Micronews. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*, 29 34.

McGlohon, M., Leskovec, J., Faloutsos, C., Hurst, M., and Glance, N. 2007. Finding Patterns in Blog Shapes and Blog Evolution. In *Proceedings of the 1st International Conference on Weblogs and Social Media*.

Sia, K. C., Cho, J., and Cho, H. K. 2007. Efficient Monitoring Algorithm for Fast News Alerts. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 19, Issue 7, 950 961.

Urbansky, D., Reichert, S., Muthmann, K., Schuster, D., and Schill, A. 2011. An Optimized Web Feed Aggregation Approach for Generic Feed Types, In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 638 641.

Yang, J. and Leskovec, J. 2011. Patterns of Temporal Variation in Online Media. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*.