대용량 블로그 워크로드 분석

전명재*, 전병찬**, 이영규***

요약

본 논문에서는 대용량 블로그 호스팅 서비스 중 하나인 티스토리의 웹 접근 로그 분석을 통해 웹2.0 워크로드의 특성을 설명한다. 티스토리 워크로드와 기존의 웹 워크로드의 비교를 통해 Web 2.0에서의 컨텐트 생성은 사용자 참여도에 관한 특성, 파일 종류에 관한 특성, 그리고 정적 혹은 동적으로 생성된 파일의 특성에 많은 변화를 야기한다는 것을 밝혔다. 특히, 본 논문에서는 블로그의 파일들을 파일 수준과 아티클 수준을 함께 분석하는 계층적 접근법을 제시하여 사용자에 의한 컨텐트 생성에 의한 특성들을 심도있게 분석하였다. 발견된 특성들은 수학적인 분포들로 모델링이 되었고, 이를 통해 추후 워크로드 생성을 할 수 있는 가능성을 제시하였다.

Empirical Analysis of a Large-Scale Blog Workload

Myeong-jae Jeon*, Byoung-Chan Jeon**, Young-kyu Lee***

ABSTRACT

This paper presents a detailed characterization study of Web 2.0 workload through an analysis of real-world Web access logs collected from Tistory, one of the largest blog hosting services in South Korea. Compared to traditional Web workloads, Tistory's workload shows different characteristics; the content creation from a number of end users drastically changes user participation characteristics, file type properties, and static and dynamic content behavior. Therefore, we delve deeply into the user-created content by following two-tier hierarchical approach that analyzes the Tistory's workload into the file and article level. Through this paper, the observed characteristics are approximated by theoretical distributions to provide data for generating synthetic workloads.

Key Words : Web 2.0, User-created Content, Workload Characterization, Measurement, Modeling

* Department of Computer Science, Rice University, USA (✉myeongjae@gmail.com)
** 청운대학교 방송영상학과
*** 혜천대학 사회복지과
· 제1저자(First Author) : 전명재 · 교신저자(Correspondent Author) : 이영규
· 접수일(2010년 12월 24일), 수정일(1차 : 2011년 1월 24일), 게재확정일(2011년 1월 27일)

Ⅰ. Introduction

Understanding the workload characteristics of the hosting servers is critical to improving the current systems and designing new systems for the Web 2.0 services. However, there has been a lack of research on the workload characterization of the Web 2.0 servers. The blog reflects the unique features of Web 2.0 applications in that, Internet end users create and read content in the form of articles. In this paper, we empirically analyze a real-world workload trace gathered from Tistory [1], one of the most popular and large-scale blog services in South Korea. The blogs hosted by Tistory generate 8 million requests and 400 GB of network traffic per day.

We first revisit traditional Web workloads to provide Tistory's unique results that stem from the active content creation by end users. We then delve into such content by following a two-tier hierarchical approach that analyzes the server-side workload into the file and article level. In particular, the article-level analysis investigates user activities on producing and consuming blog articles, thus providing more accurate, user-oriented analysis results.

A number of interesting characteristics are found. For example: 1) our workload shows that the file and transfer size distributions are heavy-tailed only for non-multimedia content (i.e. files excluding audio and videos); 2) users prefer to read articles posted along with article files such as image and audio files, but generate no hot spot for the article re-referencing pattern. In addition to these findings, we also approximated the characteristics of blog content generation/consumption using several statistical models.

Ⅱ. Background

In a *blog page*, the *template* is used to form the design of blog page using a combination of images, CSS, and JavaScript. The *article* is the unit of blog posting and consists of *article text* and *article file*; article text is simply text written by a blogger, whereas article file refers to files uploaded when a blogger publishes an article. Article files are additionally sorted into *article image*, *article audio*, and *article video*, according to the file type.

Blog servers provide users with blog content that is produced in two manners. The *dynamic content* is generated on-the-fly by server-side scripting when it is sent to clients, and the *static content* is stored in disk and fetched using file system operations. In Tistory, generating blog pages requires server-side scripting to dynamically embed text like article text. Besides such text, all other content is managed by server file system, thus belonging to static content.

We collected the access logs of blog servers over the period of 12 consecutive days. After eliminating unsuccessful requests from the access logs [2], a total of 50,118,065 requests were obtained, wherein we observed 948,290 visitors and 11,629 blogs.

표 1. 파일 타입별 통계량
Table 1. Statistics of unique files for each type

	html	Java Script	CSS	image	audio	video	flash
Files (#)	123	1,438	7,989	401,921	15,414	409	1,208
Avg. Size (KB)	18.42	3.042	7.020	132.8	4,427	4,163	603.2
Stdev	30.17	5.067	7.588	269.6	2,164	3,645	1,301

III. Re-evaluation of Web Characteristics

1) *User Participation.* An analysis of blog user's participation is done by inspecting HTTP request types. Specifically, GET request retrieves files from the Web server, whereas POST request is used when users upload content, make trackbacks, and post comments.

A breakdown of total 50,118,065 HTTP requests shows that POST request occupies 2.7% (1,314,064), a fairly larger portion; in 1998 World Cup Web site, the access logs show a significantly smaller portion of POST with 0.06% [3]. The most popular use of POST is author requests for uploads of content such as article images, article audio, and article videos, which comprise a total of 348,530 attempts. Although the fraction of POST seems to be insignificant, it clearly reveals the tendency of users' participation in publishing content and online interactions.

2) *File Type Properties.* Given the users actively participating in blogging, it is rationalized that file type properties are heavily affected by content that users supply. To this extent, the following illustrates network or storage utilization influenced by the various types of content.

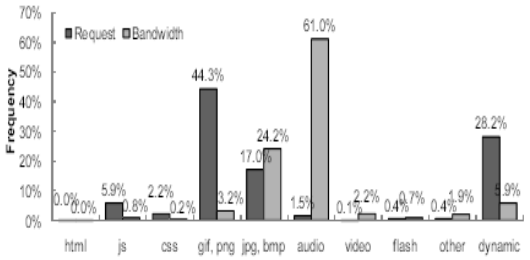


그림 1. 파일 타입별 참조 횟수 및 대역폭 사용
Fig. 1 The number of requests and bandwidth used for each file type

Table 1 presents the statistical summary of unique files for each content type. It can be seen that images significantly dominate in number (401,921), while audio and videos dominate in terms of size (over 4MB). Especially, the large number of images is primarily due to the large photographic images (JPG and BMP) that bloggers supply, with more than half of all stored images. In addition, we observed that many bloggers upload audio content onto their blogs, but not videos.

As a result of the prevalence of large, user-supplied content, we found that the average transfer size is about 100KB, several times larger than that (5.7KB ~ 13KB) illustrated in previous works [2][4]. Furthermore, a breakdown of file transfers given in Fig. 1 shows additional three properties distinct from previous studies [2][4]; 1) Audio files have large traffic volume with a few requests; 2) The bandwidth consumed by images is much smaller than traditional Web workloads, despite the comparable percentage of the number of requests; 3) HTML requests are drastically reduced from 10% ~ 30% to 0.003% due to the prevalence of dynamic Web pages, which noticeably replace static Web pages.

Even with the reduced bandwidth usage by images, photographic images are still responsible for one-fourth of the total usage. Specifically, they account for 17% of total requests and 24.2% of total bandwidth usage, thus ranking second in bandwidth usage. The influence of videos is still minor.

3) *Static Content vs. Dynamic Content.* Processing requests for dynamic content often involves a substantial amount of time for invoking programs via server-side scripts.

표 2. 동적/정적 파일 통계량
Table 2. Summary of dynamic/static content

	Static	Dynamic
Total requests	36,138,738	13,979,327
Total processing time (s)	22,573,520	4,424,946
Mean processing time (us)	624,635	316,535
Stdev of processing time (us)	10,738	557
Total bytes transferred (GB)	4452.53	276
Mean transfer size (KB)	128.74	20.97
Stdev of transfer size (KB)	739.87	336.12

Therefore, Web servers that extensively service dynamic content sometimes show high CPU usage rather than the pressure on network or storage bandwidth [5]. Also in Tistory, the use of dynamic Web page has completely replaced the static Web pages, constituting 28% of the total number of requests. Thus, we measure the bandwidth usage and response time for each content type.

Table 2 compares the processing time and bandwidth usage from the requests classified by content type. The results show that most time and bandwidth is consumed by static content even though the number of requests for it is somewhat comparable to dynamic content. In detail, approximately 94% of the total bandwidth (over 4.4TB) is used for delivering the static content. The average processing time of dynamic content and static content is 0.31 seconds and 0.62 seconds respectively, giving a twofold difference. Moreover, the sum of time required to treat all static content requests constitutes 82% of the total processing time for all requests. This finding contradicts the observation in [5] where dynamic content slows down Web servers considerably.

In order to understand the reason for the biased processing time, we measured its average of static content with respect to different range of size. For the content smaller than 16KB, the time is bounded below 10 ms, but increases proportionally to the other larger content size. Processing a request for 256KB static content was delayed up to 1.3 seconds.

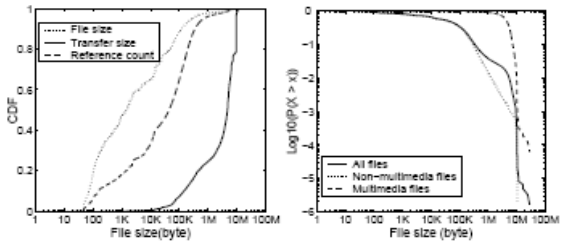


그림 2. 정적 파일 분포. (a) 파일 크기, 참조 횟수, 전송량 누적분포, (b) 멀티미디어, 비 멀티미디어 파일 크기의 LLCD
Fig. 2 Distributions of static content. (a) CDF of file size, reference count, and transfer size, (b) LLCD of file size for all, multimedia, and non-multimedia types

Fundamentally, these unique characteristics of Tistory are obtained by the active content creation that simultaneously and unpredictably involves the transfer of the content. Thus, the known characteristics of Web server workloads might not be reliable enough to be used in blog servers. This motivates our further study of static content, since it may present new challenges to underlying systems.

For the study of static content, we follow a two-tier hierarchical approach that describes the traffic to blog services at file level and article level.

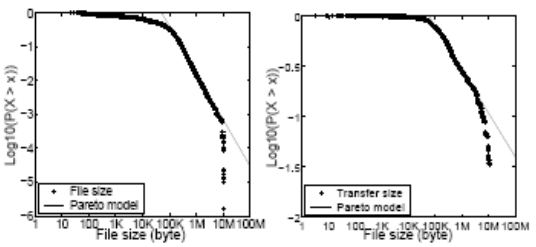
IV. File-Level Analysis of Static Content

In this section, we show our analysis of static

content (static files) in workloads. The static content can be categorized into two major file types, multimedia and non-multimedia files. The multimedia content includes audio and video files, and the non-multimedia content includes static content such as images, JavaScript, CSS, and Flash.

4.1 File Size and File Transfer Size

Fig. 2(a) presents a CDF of static content with respect to file size, reference count, and transfer size. In the figure, the file transfer size denotes aggregate bandwidth usage for a specific file size. It has been simply calculated by '*file size* x *reference count*'.



(a) File Size (b) File Transfer Size

그림 3. 비 멀티미디어 파일의 크기 및 전송량 LLCD
Fig. 3 LLCD transformation for non-multimedia files

From Fig. 2(a), we observe that a small number of large files requires high server bandwidth, whereas a large number of small files requires low server bandwidth. For instance, 92% of total static files are smaller than 400KB, whereas their bandwidth demands are less than 20% of total server bandwidth. A major part of the bandwidth consumption is generated by the large-size article files posted by users; in particular, among the unique files larger than 400KB, photographic images (59%) and audio files (33%) are the two most prevalent types.

Moreover, in Fig. 2(a) we see that the file size distribution visually follows a heavy-tailed distribution, whereas the file transfer size distribution does not show the visual trend of a heavy-tailed distribution. Thus, we confirm our visual finding from the figure by modeling the distributions with a Pareto distribution, which is a well-known model for representing the heavy-tailed distribution. Thus, we transformed the CDF plots of file size and transfer size to the Log-Log Complementary Distribution (LLCD) plots and investigated their R^2 *goodness-of-fit* test values.

From our modeling results, we found that the LLCDs of file size and transfer size could not be modeled in a Pareto distribution. Surprisingly this observation is completely against those observed in conventional Web workloads where the file size and the transfer size show heavy tailed distributions [2][6].

표 4. 비 멀티미디어 파일의 크기 및 전송량 모델 파라미터
Table 4. Approximation results for file size and transfer size distributions of non-multimedia files

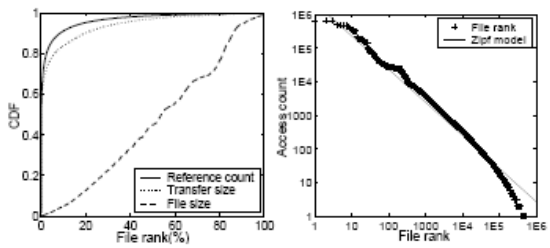
	Transfer size	File size
tail index (α)	0.43	1.37
goodness-of-fit (R2)	0.98	0.99

For further analysis, we attempted to model the LLCD distributions of file size and transfer size for multimedia and non-multimedia files. We found that only non-multimedia files showed heavy-tailed distributions for their file size and transfer size.

Fig. 3 provides the LLCD of those files for both file size and transfer size. From the figure, we see that both can be neatly modeled in a truncated Pareto distribution where there exists a natural upper

bound that curtails the tail. The tail index (α) is shown in Table II with the confidence level (R^2). The tail length of the LLCD ranges from an order of magnitude for file transfer size to three orders of magnitude for file size. Although the distributions strongly fit in the Pareto distribution model with high R^2 values, each presents a fairly different value. Table 4 shows that is 0.43 for the file transfer size and 1.37 for the file size, indicating that the distribution of file transfer size is much more heavy-tailed than that of the file size. Obviously, this result is mainly due to the dominant bandwidth usage by photographic images, which is 59% of the total.

The implication of our data is that conventional Web servers may not be perfectly suited for blog service workloads. The file size and transfer size of Web workloads have been known to be heavy-tailed in several studies.



(a) CDF (b) Zipf Analysis

그림 4. 파일 참조횟수 순위화 후 누적분포. (a) 파일 크기, 참조 횟수, 전송량 누적분포, (b) Zipf 모델링

Fig. 4. Concentration of references for ranked files: (a) CDF of file size, reference count, and transfer size, (b) Zipf analysis (access count versus rank)

However, the large-scale blog workload shows that both the file size and transfer size distributions are heavy-tailed only for non-multimedia content. Assuming all the blog content files are managed by a central server, the blogs may generate significantly different request patterns to the underlying server systems.

4.2 File Referencing Behavior

To analyze file referencing behaviors, we measured the relationship between popular files and their actual storage and network bandwidth usages. For this purpose, files are sorted by their reference counts and the most popular file is assumed to be the highest in rank. Disk space usage, network usage, and reference count are then accumulated into a CDF graph, as shown in Fig. 4(a). We see from the figure that 10% of the most frequently referenced files are responsible for 84% of bytes transferred over the network and 91% of total requests. The files, however, only account for 4.8% of the overall consumption in the storage.

Previously we observed that both large files (>400KB) and popular files (top <10%) influenced a huge volume of traffic. This is because 70% of total bandwidth is used by the files that lie on the intersection of the two groups.

A typical method of approximating the popularity distribution is to exploit Zipf's law. According to Zipf's law, the number of file references shows a consistent decline as the file rank decreases. If (R) represents the rank of a file, then the number of references (P) is: $P \sim R^{-\beta}$ where β is the slope value. The formula shows that the number of references is inversely proportional to the rank of the file. In other words, if the popularity follows Zipf's law, a linear shape is shown in the plot by placing log-scaled R on the horizontal-axis and log-scaled P on the vertical

axis. Fig. 4(b) shows that Zipf's law approximates the popularity of our traces well, with β and R^2 goodness-of-fit value close to 0.99 and 0.98 respectively.

V. Article-level Analysis of Static Content

The methodology for collecting article access logs is explained in [7]. After eliminating a few articles (less than 3%) with unknown size, we finally found a total of 87,332 unique articles with 235,586 article files stored in the blog servers and 1,088,210 accesses, which accompany 2,909,894 article file transfers, for such articles during the trace period.

In addition to characterizing blog articles in terms of their sizes, we further analyze the inter-reference time of article accesses to investigate whether subsequent accesses exhibit temporal locality.

1) Unique Article. Fig. 5(a) shows a histogram of unique articles with respect to the number of article files. The article contains 2.74 article files on average and the median and the standard deviation are 1 and 8.89 respectively.

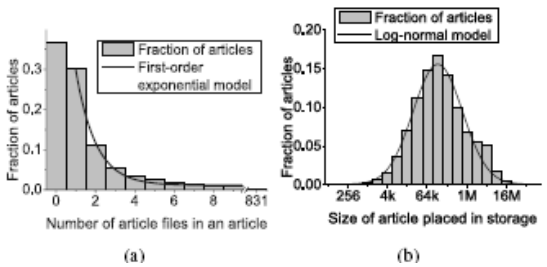


그림 5. 아티클 분포. (a) 아티클 수, (b) 아티클 크기
Fig. 5 Histograms of unique articles: (a) the number of article files, (b) article size

The number of articles having less than 4 article files occupies 87% of the total, indicating that bloggers prefer to publish their own articles with a small number of article files.We can approximate the histogram in Fig. 5(a) with a first order exponential distribution ($0.01+0.77e^{-x/1.0}$) except for zero-file articles (i.e. articles with only text). The goodness-of-fit (R^2) value of 0.99 shows a strong fit to the distribution. This approximation explains that articles exist in blogs with an exponentially decreasing rate as the number of article files increases.

Fig. 5(b) shows a histogram of unique articles with respect to their sizes. Here, the average size is 1.1MB, the standard deviation is 4.6, and the median size is 203KB. Articles are mostly located within a range of 32KB ~ 2MB, accounting for approximately 74% of the total. We can approximate the histogram with a log-normal distribution ($\mu = 12.3$, $\sigma = 1.80$, natural log), as shown in Fig. 5(b). It shows a strong R^2 value of 0.98.

2) Access Frequency of Unique Article. Fig. 6(a) shows a histogram of articles with regard to their access frequency. The average access frequency is 2.67, the standard deviation is 6.12, and the median is 1. The number of articles composed of less than 4 article files accounts for 86% of the total. When the percentage of access for zero-file articles is compared with that of unique articles, we observe that blog visitors prefer to read articles posted along with article files such as image and audio files. The histogram of Fig. 6(a) can also be approximated with a first order exponential distribution ($0.01+1.10e^{-x/0.98}$) except for zero-file article accesses. The R^2 value of 0.97 shows a strong fit to the distribution.

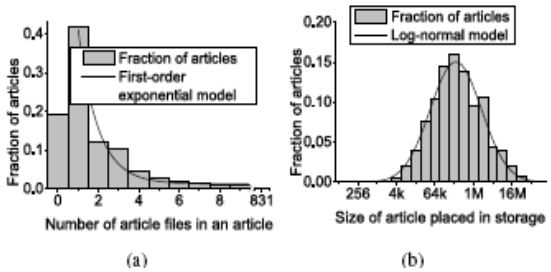


그림 6. 아티클 참조량 분포. (a) 아티클 수, (b) 아티클 크기

Fig. 6 Histograms of accesses to unique articles: (a) the number of article files, (b) article size

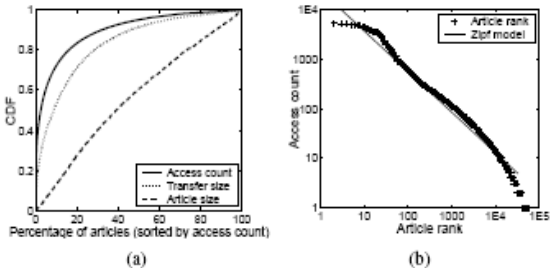


그림 7. 아티클 참조횟수 순위화 후 누적분포

Fig. 7 Concentration of references for ranked articles

Fig. 6(b) shows a histogram of the size of articles. The average size of articles is 1.07MB, the standard deviation is 3.46, and the median is 202KB. Articles are aggregated within a range of 32KB ~ 2MB, occupying approximately 74% of the total. We can approximate the histogram with a log-normal distribution (μ = 12.3, σ = 1.80, natural log), as shown in Fig. 6(b). It shows a strong R^2 value of 0.98.

Similar to exponential distribution, the log-normal distribution is common when the average is low and the variance is large. Our analysis shows that articles' access and creation patterns of most users are highly concentrated in small articles or articles with a small number of article files.

2) Article Referencing Behavior. We conducted the rank-based analysis, as discussed in Section 4.2. We first sorted the articles by their access frequency. In Fig. 7(a), we show the CDFs of the storage and network bandwidth usages, and access frequency of the article files. The graphs show that 10% of most frequently accessed articles are responsible for 73% of total reference counts and 53% of bytes used to transfer total articles.

This rate is much smaller than the value observed in our analysis of file popularity that 90% of requests and 84% of bytes are concentrated on the top 10% of most frequently requested files. Due to the diluted concentration on the articles, modeling the popularity with Zipf's law yields a lower slope value than the case of file popularity. The β and the R^2 are 0.82 and 0.98 respectively, as plotted in Fig. 7(b).

3) Temporal Locality. We found that the access patterns of articles exhibited weak temporal locality. Fig. 8 shows the CDF of inter-reference time of articles. We observed that the articles that were revisited within 30 minutes and 1 hour account for 31% and 62% of the total, respectively.

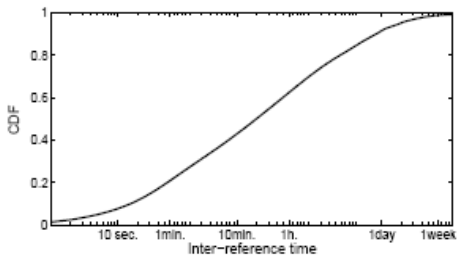


그림 8. 아티클 재 참조 시간의 누적분포

Fig. 8 CDF of inter-reference time for articles

A noticeable trend in the graph is a monotonous increase as the interval is shifted toward higher values, and thus no hot spot is detected for the re-referencing pattern. One might expect that the temporal locality would be shown in article accesses since the top 10% of popular articles take up 73% of total reference counts. We speculate that this decayed locality may be strongly affected by active content creation, where article references are distributed among various articles.

VI. Related Work

Several studies have focused on the patterns or evolution of user generated content in the Web 2.0. Guo *et al.* [8] studied services similar to blogs and found that the user posting behavior exhibits strong daily and weekly patterns. Cha *et al.* [9] studied information propagation in Flickr and showed that popular photos do not spread widely and quickly throughout the network. Burke *et al.* [10] analyzed server log data from Facebook and found that newcomers who see their friends contributing go on to share more content themselves. Leskovec *et al.* [11] studied the temporal and topological patterns of information propagation in blogs, and found that the popularity of blog posts drops with a power law.

The prior work most similar to ours is that of Duarte *et al.* [12]. They analyzed HTTP requests from a Brazilian blog service. One of their two findings in the server context is relevant to ours; file transfer size can be modeled by a Pareto distribution. As a complementary study to [12], this work underscores the importance of the comprehensive analysis of aggregated server accesses in the server context.

VII. Conclusion

We have investigated user activities, the behavior of user-created content and article distributions in blog servers. We also modeled detailed characteristics using statistical distributions. This paper will provide useful guidelines when modeling Web 2.0 workloads, performing workload synthesis, and designing new systems for Web 2.0 applications.

We believe that the observations presented in this paper will provide useful guidelines when modeling OSN workloads, performing workload synthesis, and designing new systems for hosting OSNs.

References

[1] http://www.tistory.com

[2] M. F. Arlitt and C. L. Williamson, *"Internet web servers: workload characterization and performance implications,"* IEEE/ACM Trans. Netw., 1997.

[3] http://ita.ee.lbl.gov/html/contrib/WorldCup.html

[4] L. Bent, M. Rabinovich, G. M. Voelker, and Z. Xiao, *"Characterization of a large web site population with implications for content delivery,"* WWW, 2004.

[5] V. Holmedahl, B. Smith, and T. Yang, *"Cooperative caching of dynamic content on a distributed web server,"* HPDC, 1998.

[6] C. W. A. Williams, M. Arlitt and K. Barker, *"Web workload characterization: Ten years later,"* Web Content Delivery. Springer, 2005.

[7] M. Jeon, J. Hwang, J. Jang, E. Seo, and J. Lee, *"Characterization of a large-scale blog traffic,"* TR/CS-2009-300 KAIST, 2009.

[8] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao, *"Analyzing patterns of user content generation in online*

social networks," ACM SIGKDD, 2009.

[9] M. Cha, A. Mislove, and K. P. Gummadi, *"A measurement driven analysis of information propagation in the flickr social network,"* WWW, 2009.

[10] M. Burke, C. Marlow, and T. Lento, *"Feed me: motivating newcomer contribution in social network sites,"* ACM CHI, 2009.

[11] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst, *"Cascading behavior in large blog graphs,"* SDM, 2007.

[12] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida, *"Traffic characteristics and communication patterns in blogosphere,"* ICWSM, 2007.

저자소개

전명재(Myeong-jae Jeon)

2005년 광운대학교 컴퓨터공학과 학사

2009년 KAIST 전산학과 석사

현재: Ph.D. Student in Computer Science, Rice University

※ 관심분야 : Operating Systems, Machine Virtualization, Computer Architecture

전병찬(Byoung-Chan Jeon)

1994년 수원대학교 전자계산학과 (공학석사)

2002년 순천향대학교 전산학과 (공학박사)

2005년~현재 청운대학교 방송영상학과 교수

※ 관심분야 : 컴퓨터구조, 홈 네트워크, 모바일

이영규 (Young-Kyu Lee)

1999년 순천향대학교 전산학과(공학박사)

1997년~2001년 극동정보대학 정보처리과 교수

2001년~현재 혜천대학 사회복지과 교수

※ 관심분야: 병렬처리, 데이터베이스