Object Popularity Distributions in Online Social Networks

Theo Lins
Computer Science Dept.
Federal University of Ouro Preto (UFOP)
Ouro Preto, Brazil
theosl@gmail.com

Fabrício Benevenuto
Computer Science Dept.
Federal University of Ouro Preto (UFOP)
Ouro Preto, Brazil
benevenuto@gmail.com

Wellington Dores
Computer Science Dept.
Federal University of Ouro Preto (UFOP)
Ouro Preto, Brazil
wellingtonjdores@gmail.com

Fabrício Barth
Apontador Corporation
São Paulo, Brazil
fabricio.barth@gmail.com

WebSci'12, June 22–24, 2012, Evanston, Illinois, USA.

Abstract

The extreme popularity of online social networks (OSNs) might have the potential to reshape the Internet traffic in the Future. Unlike traditional Web servers where popular objects are required anywhere, requests to online social networks are dominated by content of local interest, where users typically consume content created by friends. This paper investigates changes in the access patterns of social networks in comparison with traditional Web workloads. Using real data from popular systems like Orkut and YouTube we show that popularity distributions of objects in OSNs is less skewed in comparison with traditional Web servers. This observation suggests that we might need to rethink the strategies currently employed in Web caching systems in view of the growing wave of popularity of social systems.

Author Keywords

Social networks, Internet of the Future, workloads

ACM Classification Keywords

C.4 [Computer System Organization]: Performance of Systems—Measurement techniques

General Terms

Measurement, Design

Introduction

Since its beginning the Internet has received a large wave of applications including the Web and Peer-to-Peer networks, in which the different traffic patterns helped reshaping its infrastructure. Recently, online social networking applications have emerged as extremely popular applications. According to Alexa.com, social networks like Facebook and Twitter are among the top 10 most visited websites in the world, both in terms of unique users and in terms of time spent on these websites. With 750 million users, if Facebook was a country, it would be the third most populous country in the world [8].

Several online social networks allow some features in common. Generally, they allow users to share information with friends and have a page with the user profile that can post or update any content. Content varies from simple text messages to multimedia files like photos or videos. To encourage users to share content, social networks make updates available to users immediately after their friends share the content. Thus, not only users spend much time in these systems, but they also create huge amounts of content. As an example, the photo-sharing service on Facebook is the largest repository of photos of the Internet, containing more than 60 billion images [7]. YouTube receives 24 hours of video per minute [9].

Such quantity of users and content associated with an exponential growth of these systems suggest that social systems have the power to reshape the Internet traffic in the Future. In fact, social networking has been a major topic of discussion in the activity known as the **Future Internet**, a movement that aims at formulating and evaluating alternative architectures for the changes that the Internet might need in the Future [10]. Despite considerable interest, little is known about patterns of access to these systems and how they differ from the access patterns of traditional systems.

Intuitively, there is one crucial difference between traditional publishing of content on the Web and share content through social networks online. When people share content on the Web, they typically make the content accessible to any Internet user. On the other hand, when users share content in online social networks they often intend to reach a certain audience, like friends or followers. Often this audience is explicitly defined by the user or by the website's policy for establishing friendships. For example, pictures and videos on Facebook are generally accessible to the immediate friends of the content creator. Other times, the audience is implicitly limited by the inherent nature of the content. For example, a picture of a user and her friends at a birthday party is likely to be of interest only to close friends of that user's social network.

This crucial difference might affect one of the most important properties of the Web traffic, which is the basis for the design of any Web caching system: the high concentration of popularity in a few objects of the system. This property has shown to be valid for many Web-based systems [1, 11, 12, 2].

This work aims at investigating the following **research hypothesis**: *popularity distributions of objects in OSNs are less skewed in comparison with popularity distributions of traditional Web servers*. To do this we analyze the access patterns of different social systems of the current Web 2.0 and compare them with the workload of traditional Web servers from the namely Web 1.0. Next, we briefly detail the datasets used to investigate our hypothesis.

Datasets

This section presents the datasets we used to investigate the popularity distribution of objects in online social networks. Most of the databases described below have been used in previous works. Therefore, only the important characteristics to the present work are discussed.

Web server of the World Cup'98

Ideally, we would like to compare data obtained from existing social networks to data from Web 1.0, consisting mostly of servers containing statistics pages where Web users were spectators. A dataset that meets these requirements is publicly available. We use public data of the Web server of the World Cup 1998 [1]. In particular, we use a 30-days log (from May 24 to June 24, 1998), containing 69.747 unique objects and 681.469.425 registered requests for these objects.

Orkut

This dataset corresponds to data from Orkut collected and characterized in previous work [4, 5]. This dataset was collected from an aggregator of social networks and has a record of all objects of different social networks accessed by 36.309 users who used the system during the monitored period. To perform our analysis, we use only the requests to photos of Orkut in order to measure the popularity of photos shared in this system. In total the database contains 23.764 images, accessed 121.939 times.

YouTube

Among the current social systems, YouTube is the largest one associated with video distribution. We used a dataset containing 1.666.226 YouTube videos collected in December 2006 [6]. For each video, this dataset contains the number of views of each video. In total these videos received 369.762.000.000.000 accesses.

Uol Mais

UOL Mais[1] is a popular social video sharing system in Brazil. In addition to requests to videos, this kind of systems receive requests to other types of content like images, search, and interactions. To study the popularity of all types of objects accessed in a video sharing systems we also use the dataset of UOL Mais. A detailed description of this dataset can be found in reference [3]. The log used in this work was obtained in the period from 12 December 2007 to 07 January 2008, has 109.239 objects and 3.613.935 requests for access to these objects.

Apontador

Apontador[2] is a location-based social system, very popular in Brazil. In short, Apontador has a database containing about seven million geo-referenced places. Each place is represented by a webpage with information that describes the place, such as name, address, latitude, longitude, category, phone number, etc. Apontador has also features that are similar to FourSquare features, allowing users to post tips about places.

We obtained from Apontador a one-month log (from October 01 to October 31, 2011). As the main objects accessed in this dataset are places/locations, we consider only requests to this kind of object for this dataset. In total, the dataset contains 2.679.540 locations, accessed 27.499.271 times.

Object Popularity Distribution

The idea of having a large concentration of popularity in a few objects is the basis for the construction of hierarchical caching systems and has been widely applied in the design of systems of caches in a very recent past [15, 12]. Next,

[1] http://mais.uol.com.br/

[2] http://www.apontador.com.br

we analyze the characteristics of the object popularity of different Web systems. Figure 1 shows the normalized popularity distribution of accesses to objects of the systems presented in the previous section. The x-axis represents the ranking of the content in percentage, where the 10% ranking represents the first 10% of the objects of each dataset analyzed. The y-axis represents the cumulative percentage of popularity, i.e., for the first 10% objects in the ranking, the y-axis shows the percentage of the requests that targeted these 10% of the objects. It can be noticed the great difference in terms of concentration of popularity that each curve shows. Social systems curves are much less skewed if compared with the concentration of popularity of the World Cup 98's Web objects. As an example, whereas 10% of the objects of the World Cup Web server concentrate 97.18% of accesses, 10% of the Orkut's photos received only 50.33% of accesses.

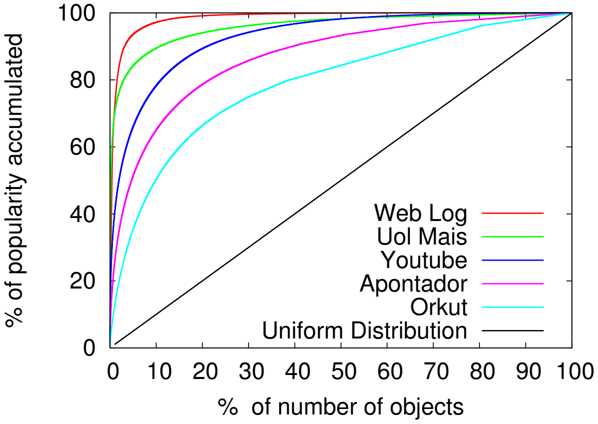

Figure 1: Normalized accesses to objects in different Web systems

In the other systems we can see that the concentration of popularity is also lower in comparison with the World Cup curve. For Uol-Mais, as the dataset contains requests for images (thumbnails) that represent the videos as well as requests for search and navigation in the system, the curve is the closest to the World Cup 98's Web server. The data from YouTube, which accounts only for accesses to videos, we can note that the concentration of popularity is less skewed in comparison with UOL-Mais. For Apontador locations, we can note that the concentration of popularity is even lower, which might reflects the local interest of different objects in this type of system. The popularity distribution of objects showed to be the less skewed for Orkut photos, as Orkut users typically access photos of friends [5, 4], avoiding the formation of popular objects in the system.

To check further the differences in popularity, we measure the disparity between the concentrations of popularity in all the logs. Disparity is a well known metric in economics to measure differences between rich and poor in a country. Typically, the 95th and 5th percentiles are compared. Table 1 shows the measures of disparity for different distributions. The disparity between the 95th and 5th percentile is 20 to Orkut and 45,831 for the world cup 98's server. Even when comparing the disparity from other distributions with the Web's distribution, we note that the disparity for the Web is orders of magnitude higher than that of social systems.

Our observations that distributions of accesses to objects in social systems are much less concentrated than in a typical Web 1.0 server raise important questions about the effectiveness of traditional infrastructure for content distribution today and, especially, in the future if expectations of growth and even greater popularity of

Ratio	Web	UOL Mais	YouTube	Apontador	Orkut
1º / 99º	703,959	334	15,410.5	128	46
5º / 95º	45,831	52	979.62	39	20
10º / 90º	15,119	24	214.61	21	12

Table 1: Popularity Disparity

social systems is confirmed. This is because the current infrastructure is based on caching of a small fraction of objects that dominate the content. The lack of extremely popular objects in sequences of Web requests suggests that it may be necessary to review the infrastructure for distribution of social content in the future. In fact, it is not surprising that recent studies have shown that the content of Facebook could be processed 79% faster and consume 91% less bandwidth [16].

Discussion

Since the first social networks, the popularity of these systems has continued to grow. Using data from large social systems, in this paper, we investigated how a basic principle of Web systems can potentially change with the growth and increasing popularity of online social network systems. We next discuss implications of our findings on the design of distribution infrastructures for social-based systems and discusses research directions we aim to approach in the future.

Our findings suggest that popularity distributions of objects in OSNs are less skewed in comparison with popularity distributions of traditional Web servers. One immediate implication from this observation is that we might need to rethink the current caching designs and strategies meant for web workloads to make them suitable for social-based websites. Our study pointed out that popularity distributions in social systems are flatter than in web workloads. This means that the benefit of traditional caching mechanisms may not hold for social-based systems.

An important aspect we aim at investigating in the future is the physical proximity between content producers and consumers of content in social systems. With the recent growth in use of online social networks via mobile devices, the spatial dimension must be clearly and carefully incorporated into infrastructure projects for the Future Internet [14]. A recent study [13] showed that URLs posted on Twitter tend to spread over short distances, usually close to the content creator. This proximity between producers and consumers could be exploited to allow users to perform uploads content on local servers geographic area, as a city or a state. Such a mechanism could reduce the amount of bandwidth consumed, compared to the strategy to perform an upload to a central server and remote.

Acknowledgments

We are grateful to Fapemig and CNPq for supporting this research. We also would like to thank Apontador (www.apontador.com.br) for the data they provided.

References

[1] M. Arlitt and T. Jin. Workload Characterization of the 1998 World Cup Web Site. 14:30–37, 2000.

[2] F. Benevenuto, F. Duarte, V. Almeida, and J. Almeida. Web Cache Replacement Policies: Properties, Limitations and Implications. In *Latin American Web Congress (LaWeb)*, November 2005.

[3] F. Benevenuto, A. Pereira, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Characterization and analysis of user profiles in online video sharing systems. *Journal of Information and Data Management*, 1(2):115–129, 2010.

[4] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 49–62, 2009.

[5] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user navigation and interactions in online social networks. *Information Sciences*, 195(15):1–24, 2012.

[6] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM Internet Measurement Conference*, 2007.

[7] Needle in a Haystack: Efficient Storage of Billions of Photos, 2009. Facebook Engineering Notes, http://tinyurl.com/cju2og.

[8] Facebook Press Room, Statistics. http://www.facebook.com/press/info.php?statistics.

[9] YouTube Fact Sheet. http://www.youtube.com/t/fact_sheet. Acessado em Março/2010.

[10] A. Gavras, A. Karila, S. Fdida, M. May, and M. Potts. Future internet research and experimentation: the fire initiative. *SIGCOMM Comput. Commun. Rev.*, 37:89–92, July 2007.

[11] D. Menascé, V. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. Meira, Jr. In search of invariants for e-business workloads. In *Proceedings of the 2nd ACM conference on Electronic commerce*, EC '00, pages 56–65, 2000.

[12] S. V. Nagaraj. *Web Caching And Its Applications (Kluwer International Series in Engineering and Computer Science)*. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[13] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 381–393, 2011.

[14] S. Scellato. Beyond the social web: the geo-social revolution. *SIGWEB Newsletter*, pages 5:1–5:5, Sept. 2011.

[15] J. Wang. A Survey of Web Caching Schemes for the Internet. *ACM Computer Communication Review*, 25(9):36–46, 1999.

[16] M. Wittie, V. Pejovic, L. Deek, K. Almeroth, and B. Zhao. Exploiting locality of interest in online social networks. In *ACM Int'l Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, pages 1–12, 2010.