Web 2.0 Traffic Measurement – Analysis on Online Map Applications

Song Lin
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science
Tsinghua University, Beijing, China
linsong1984@gmail.com

Zhiguo Gao
IBM China Research Lab
Beijing, China
gaozg@cn.ibm.com

Ke Xu
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science
Tsinghua University, Beijing, China
xuke@tsinghua.edu.cn

ABSTRACT
In recent years, web based online map applications have been getting more and more popular, such as Google Maps, Yahoo Maps. Many new Web 2.0 techniques such as mash-up and AJAX were adopted in these map applications to improve user experiences. But few researches have been done on traffic analysis of the Web 2.0 based online map applications. In this paper, we introduced our research on features of online map applications that previous studies hadn't cover. In our research, we captured map application related HTTP traffic in a campus network while not violating user privacy. We introduced the traffic overview, mash-up and web caching characteristics of four map web sites (Google Maps, Yahoo Maps, Sogou Maps and Baidu Maps). For the first time, the mash-up characteristics of Google map traffic were analyzed using a new method proposed in this paper. The same method could be applied to other mash-up analysis works. These results can help us optimize the future web application designs and CDN based accelerating solution designs.

Categories and Subject Descriptors
C.2 **[Computer-Communication Networks]**: Miscellaneous

General Terms
Measurement, Performance

Keywords
Web 2.0, online map application, mash-up, web caching, traffic characteristics.

1. INTRODUCTION
In recent years, Web 2.0 brings more and more new techniques to many new web applications such as blog, online document processing applications and web based online maps. Web 2.0 enables users to create, share, and distribute contents on the Web easily. It also makes different web sites to share data and services easily. Besides that, a lot of techniques have been used in Web 2.0 sites, such as mash-up, AJAX (Asynchronous Javascript and XML). New usage patterns and new techniques change the network traffic pattern compared to traditional websites. Web based map application is one of the typical Web 2.0 applications that uses these two techniques. In this paper, we studied several web based map applications for our Web 2.0 traffic research.

One of the techniques widely used in Web 2.0 is mash-up. Mash-up technique combines data from more than one source into a single integrated tool (an example is the usage of cartographic data from Google Maps to add locating information to real-estate data), thereby creating a new and distinct web service that was not originally provided by either source[1]. Mash-up brings traffics from multiple HTTP servers at different locations. It encourages users to enjoy more web applications thus could increase requested data exchange. So the impact of mash-up on web traffic patterns is worth analyzing. When talking about mash-up, Google Maps is also frequently mentioned, and it has played important role in the popularization of mash-up. As another key Web 2.0 technique, AJAX allows the client's Web browser UI to respond quickly to inputs, which also encourages more interactivity from users[2]. Google Maps is one of the earliest adopter of AJAX. Developers from other web based map applications followed up quickly to build their own high quality map applications.

Beside these new technologies, we also paid attention to web caching technology. While web caching is considered as an important technology by traditional web researches, the importance of web caching in the new environment of Web 2.0 based map applications needs to be discussed.

So the research in the traffic model of web based map applications can help us understand how new techniques such as mash-up and AJAX have brought impact to the new web, and whether traditional web caching technology is still effective to the Web 2.0 application. From another point of view, web based map applications are widely used and still fast growing. They provide address querying, public transport querying, real-time traffic querying with graphical results. Mashed-up with other web application, they provide convenient services such as travel information sharing, hotel booking etc. Web based map applications become a typical workload in Web 2.0 age, thus its traffic features needs to be studied, which is another motivation of our study.

There are many studies of "traditional" web workloads, such as [3]. But there are only a few research works on Web 2.0 traffic characterization. In 2007, Phillipa Gill and Martin Arlitt et.al [4] analyzed the traffic characterization of the popular Web 2.0 site YouTube. In their paper, usage patterns, file properties, popularity, referencing characteristics, and transfer behaviors of YouTube were examined. Compared with traditional web and media

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
NOSSDAV'09, June 3-5, 2009, Williamsburg, Virginia, USA.
Copyright 2009 ACM 978-1-60558-433-1/09/06...$5.00.

streaming workload characteristics, Web 2.0 provides additional metadata that should be exploited to improve the effectiveness of strategies like caching. Caching could improve user experiences, reduce network bandwidth consumption, as well as the load on YouTube's core server infrastructure.

In [5], Alan Mislove, Massimiliano Marcon et.al studied the characteristics of online social network graphs on large scale, this was the first study on this subject. Their work was valuable for improving current systems and designing new applications of online social networks. The results confirmed the power-law, small-world, and scale-free properties of online social networks.

Although there are widely distributed and fast growing traffic in web based map applications, few research works have been done on them. While our research was being carried on, Fabian Schneider and Sachin Agarwal et al. presented a traffic study of several Web 2.0 applications in [6] including Google Maps, modern web-email, and social networking web sites, and compared their traffic characteristics with the ambient HTTP traffic. They found the key differences between Web 2.0 traffic and all HTTP traffic through statistical analysis. The paper [6] focused on the impact AJAX, while in our paper we focused on the impact of mash-up and the performance of web caching under the new environment of map applications.

We select four web based map applications as our target: Google Maps, Baidu Maps, Sogou Maps and Yahoo Maps. Google Maps and Yahoo Maps are famous web based map applications all over the world, while Baidu Maps and Sogou Maps are two of the most popular map applications in China. We capture map related HTTP information in our campus network while not violating user privacy. Our analysis is based on more than 800 hours of data collection (more than 100,000 sessions and more than 6,000,000 requests were collected) that reflects trends and usage pattern of these map web sites from campus network perspective.

The main challenge in data collection was that we needed to identify map application requests from the large overall traffic effectively. We also needed to propose methods to identify the mash-up sessions. The main contributions of our paper were three folds. First, we collected HTTP basic information of Web 2.0 based map application over an extended period of time while not violating user privacy. Second, we studied the requested pictures concentration attribute of the map traffic with an interesting method and drew the conclusion that CDN servers are quite useful for accelerating web based map applications. Third, we proposed methods to study mash-up characteristics of a website.

The rest of the paper is organized as follows. We give an overview of web based map applications studied in this work and then describe our data collection platform in Section 2. In Section 3 we present the results of our statistical analysis. Finally, we have our conclusions in Section 4.

2. WEB BASED MAP APPLICATIONS OVERVIEW AND DATA COLLETION FRAMEWORK

As has been mentioned in the introduction section, web based map applications are widely used and fast growing. Web based map applications usually offer powerful, friendly map services with some local business information. When logged on to these websites, you can use mouse or keyboard to pan and drag maps to view adjacent sections immediately (no long waiting for new areas to download with the help of AJAX). Images you pan and drag include map or satellite images, or even a hybrid of them. You can find business locations and contact information all in one location, integrated on the map, sometimes with additional information. You can also look up an address and let map applications mark the location and driving directions for you, or perform public transport querying between two location specified by you.

Google Maps and Yahoo Maps offer map service in many countries. They provide different functionality in different countries. Baidu Maps and Sogou Maps focus on providing map service in China.

The typical architecture of map web sites looks like this: Some servers accept user requests, provide general map services, such as drag, zoom, and search. We would like to call these kinds of servers map servers, and call the domains that they use map domain. Web 2.0 based map application uses AJAX to improve user experience. The large map shown on the client browser is a combination of some small pictures (about 8KB for Sogou Maps, 18KB for Google Maps). When map servers need to send user map pictures, they do not directly send these pictures in their response, instead, they redirect user to some other servers, which we would like to call them map image server. Image servers just provide pictures, and can use multi-homing server or CDN to provide better performance and user experiences. For the domains which map image servers are using, we call them map image domains. Table 1 shows some examples for map domains and map image domains of Google Maps, Baidu Maps, Sogou Maps and Yahoo Maps. The results were obtained by using Bro [7] to gather traces on a workstation while we browsed the maps site.

Table 1. Map domains and Map image domains

	Map domain	Map image domain
Google	maps.google.com	khm1.google.com mt1.google.com
Baidu	map.baidu.com	mappng.baidu.com
Sogou	map.sogou.com	pic1.go2map.com
Yahoo	maps.yahoo.com	us.i1.yimg.com

We collected map related basic HTTP information at the Tsinghua University campus network without gathering any information concerning user privacy. The campus network of Tsinghua University has a 10 Gbps backbone and was connected to CERNET (China Education and Research Network) through a 2 Gbps full-duplex network interface. IPv4 and IPv6 are both supported by the network. There are about 40 thousand computers connected to the campus network. It contains about 42 thousand users which we think is large enough that the characteristics we get from the collected data present a general view of web based map applications. The data can help us understand how web based map application may be used by the clients of other large networks.

As the first step of our data collection, we do DNS queries of domains like those listed in Table 1 to get the servers' IP. These servers IP will be used as filter in out capture. Notice that the DNS queries should be redone and updated in the filter frequently. We redo DNS query work every day, in order to identify the IP changes for these servers.

For data collection, we need packet level data capture, TCP level reassembly and basic HTTP level analysis. Among many libpcap-based applications [8], we found that Bro[7], Snort[9] and Argus[10] may be suitable for our data capture requirement. At last we chose Bro[8] to extract summaries of each map web site HTTP transaction in real-time. Bro is a network intrusion detection system that can be used to capture data and construct the HTTP request-response stream for all connections.

Bro uses a specialized policy language that allows a site to tailor Bro's operation[8]. The three policy scripts we need to modify are http.bro, http-reply.bro and http-header.bro. We modified the output scripts of them in order to reduce unnecessary I/O and to avoid logging private information. In http.bro, we did not save the IP address and port number of HTTP clients, but only logged timestamp when the session starts, session id, server IP address and server port. In http-reply.bro, we removed URI parameters and only kept other basic information of the HTTP request-reply, typically including timestamp, session id, handled requesting URI and server host name. In http-header.bro, we did not log COOKIE, and we remove URL parameters in the REFERER header similarly. In the whole process, IPs, COOKIEs, URL parameters, POST data, and HTTP payloads of HTTP clients were never logged for protecting user privacy. Some previous works used mapping to handle private information, but following the principle in network measurement that only data that are needed should be collected, these data were just ignored since we did not need these data in our current research issue.

Among the requests we had captured, we found that some of them come from other web traffic. For example, when we gather Sogou Maps traffic, some traffic from Sogou Blog will be also captured due to filter policy drawbacks. We mainly use URL prefix and HOST from HTTP header to verify the validity and remove invalid packets. This methodology is similar to Schneider's methodology[6] so we do not describe the details here.

3. CHARACTERISTICS OF WEB 2.0 BASED MAPS APPLICATIONS

Among the 843 hours of traffic we have collected, there are over 100,000 valid sessions and more than valid 6,000,000 requests. Figure 1 shows session percentage for the four map applications that we captured. Since users in our case would be charged for international data exchange, some of them used proxies to access the web site located out of China. There are only 103 sessions from Yahoo Maps in our captured data, since Yahoo Maps doesn't have local server in China. Therefore we only focused on the other three web based map applications including Google Maps, Sogou Maps and Baidu Maps. Traffic from map servers contains the characteristics introduced by Web 2.0 techniques such as AJAX and mash-up. Adding traffic from map image servers may interfere features of these technologies and traditional HTTP traffic, so we focused on traffic from map servers in traffic statistics.

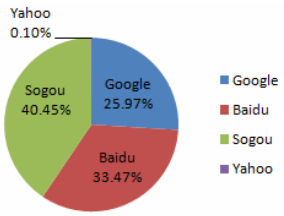

Figure 1 Session percentage for different applications

3.1 Basic Traffic Information Distribution over Time

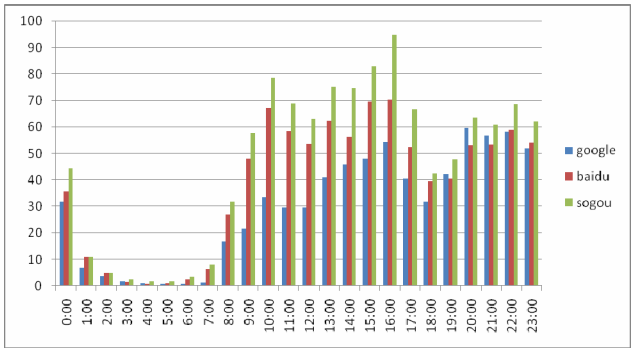

Figure 2 Average session count in every hour interval

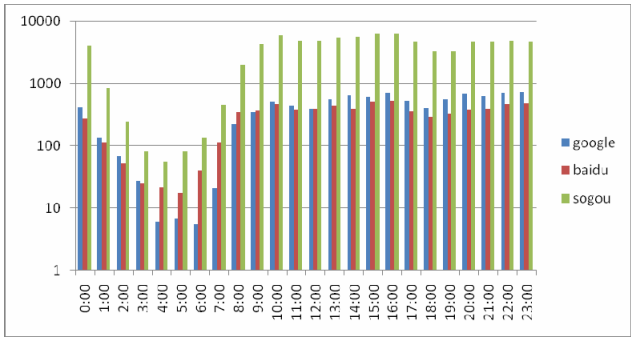

Figure 3 Average request count in every hour interval

Figure 2 and 3 show the fraction of average session count and request count at a particular time of day. The first three bars are traffic information between 0:00 and 1:00 for Google, Baidu and Sogou respectively, and so on. We can see the map traffic vary by time. The session and request count are much smaller at late-night than other time slot. Another trough appears at supper time (18:00 to 19:00). For different web sites, Sogou has a slightly larger session number, which shows its relatively popularity, and has a larger request number, which may also be caused by shorter interval request time or implement style.

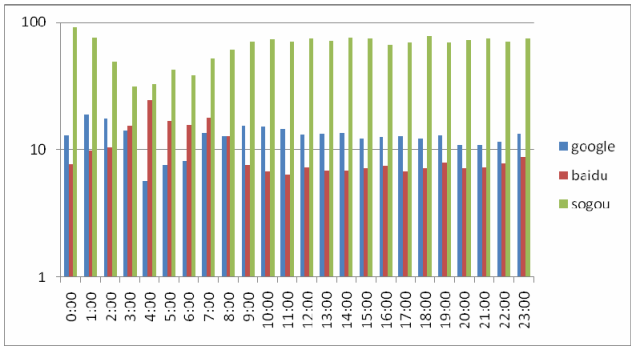

Figure 4 Average request count per session in every hour interval

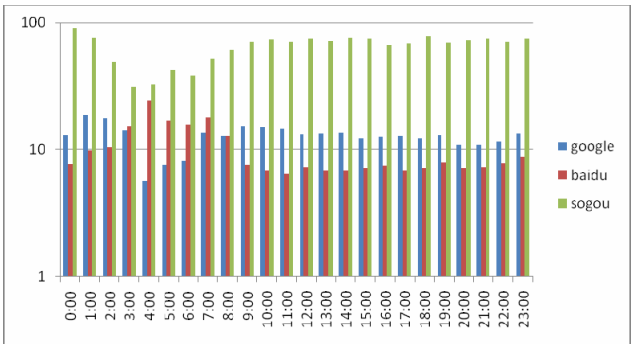

Figure 5 Average HTTP payload size per session in every hour interval

Figure 4 and 5 show the average request count and HTTP payload size per session, by the interval of an hour. For the same website, these two main characteristics remain at a similar level. For a comparison between web sites, Sogou Maps still shows a larger request count and HTTP payload size per session, since users always need more interactions with Sogou to get the detail information of the map.

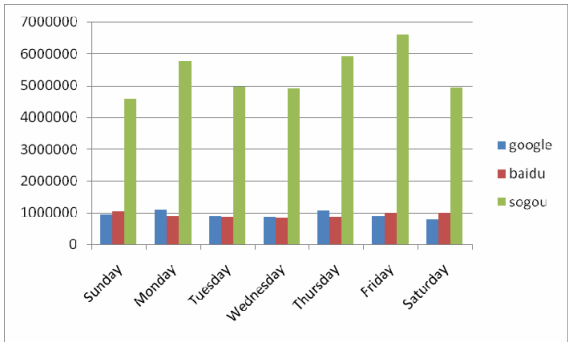

Figure 6 Traffic distributions along a week

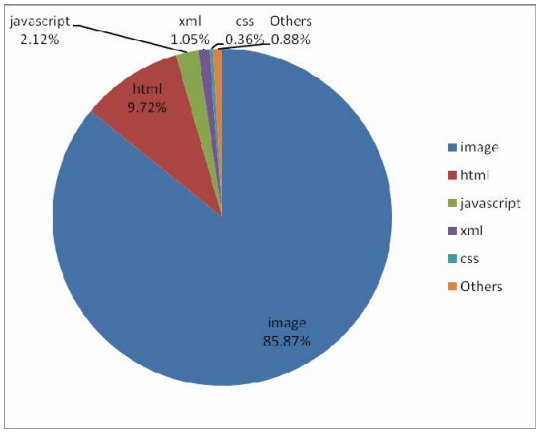

Figure 7 The composition of traffic

Figure 6 shows the total traffic distributions in a week. Generally speaking Friday has a larger traffic value perhaps because people need to make their plans for the weekend. Moreover, when we deep dived the composition of the traffic from Sogou Maps shown in figure 7, we found images count for about 86% of traffic, java script only occupied 2% of traffic since java script only download once and cached in client when using AJAX technique.

3.2 Mash-up

In Web 2.0 application development, mash-up has always been used to combine data from more than one source into a single integrated application. Integrating Google Maps service has always been considered as a typical example for mash-up. Through the Google Maps API, users can embed Google Maps in their own web pages with JavaScript. The API provides a number of utilities for manipulating maps (just like on the http://maps.google.com web page) and adding content to the map through a variety of services, allowing website-constructors to create robust maps applications on websites[11]. Using cartographic data from Google Maps, other web sites can provide more fancy services such as hotel booking service, driving navigation. Users could also add marks or tags into the Google Maps through API. From the traffic view point, when Google was mashed up, two parts of traffic will be collected, one part comes from the service websites, the other part comes from Google Maps.

To tell whether a Google Maps session comes from a mash-up, or from a direct visit to Google Maps (maps.google.com or its local domain at different countries), we proposed a method to solve the issue. The REFERER of request header in each session was checked, if it comes from a website other than the current session domain, we take that it comes from a mash-up. We call it likely mash-up session. The major inaccuracy of this method occurs then you click a hyperlink on a webpage, REFERER may also be the page URL you are browsing, and this session will be marked as a likely mash-up session too. We calculated the percentage of likely mash-up sessions from Google Maps, Sogou Maps and Baidu Maps, the percentages of likely mash-up sessions out of all sessions are shown in Table 2.

Table 2. Percentages of likely mash-up sessions for different web sites

	Total sessions	likely mash-up sessions	percentage
Google	26167	10875	41.6%
Baidu	33720	571	1.7%
Sogou	40750	2136	5.2%

We found that Baidu Maps and Sogou Maps were seldom used in mash-up. Only 1.7% and 5.2% sessions of Baidu Maps and Sogou Maps were mash-up sessions. The reason was that Sogou and Baidu doesn't provide easy-to-use API for web developers. Although there might be a little inaccuracy of our method, it is obvious that Google Maps still gets large part of its sessions from mash-up usage. To identify where these likely mash-up sessions come from, we checked their REFERERs. For the values in REFERERs, we generally used its value before the first "/" for classifying except www.google.com/ig, which can be considered as a mash-up of Google applications, so we would like to separate it from other URLs that begin with www.google.com. Figure 8 shows the distribution of the 26167 sessions coming from Google Maps.

From the figure, we found that quite a few come from other Google web applications such as Picasa, Panoramio. For example, Google Picasa has a Google Maps mash-up on everybody's homepage to show the location of picture; Panoramio is a web site from Google for user to upload pictures all over the world. Others are websites for sharing traval information. Some are tickets or hotels booking websites. All of these websites have clear motivations to mash-up with Google Maps for showing locating information and to improve their service quality. The only website that can be considered as inaccuracy in the top 15 websites is www.google.com. Some of the sessions with a REFERER of "www.google.com" may come from a direct click from the hyperlink on the Google homepage. But generally speaking, the method that proposed in this paper can be used to detect mash-up works fine and can be of great help in the future of mash-up study.

As we can see, mash-ups really take up a large part of the sessions of Google Maps. It is reasonable for other websites to consider whether and how mash-up can help to popularize their own website.

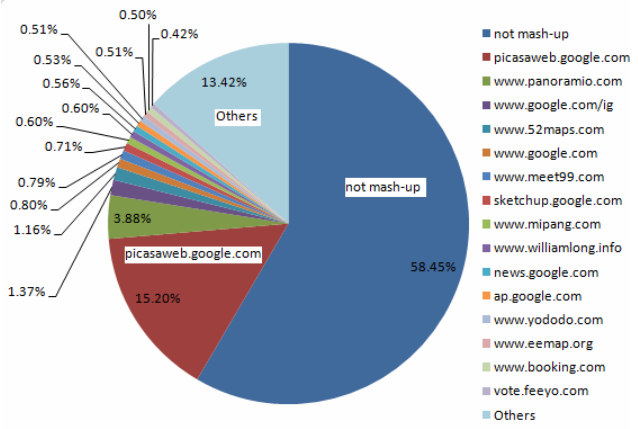

Figure 8 Session percentages of Google Maps that comes from different web site mash-ups

3.3 Requested Pictures Concentration

Since a relatively larger Sogou Maps traffics were collected, we paid more attention to them. As mentioned before, map servers handle user requests and redirect the image requests to some map image servers. These images take up to about 68.7% of total Sogou Maps traffic, it is worthy to deeply analyze the content of these images. We need to know how the locations of these requested images distribute in the whole map, because the distribution is important for designing caching policy and accelerating application through CDN. We will give an intuitive view on their concentration and some statistical results.

At first, we'll introduce the method on how to calculate the requesting location through the collected request URL. To do this work, we need to define zoom level since Web 2.0 based map website can be zoomed at different levels. We define the level that we can not zoom out any more as zoom level 0. At this level we usually see the whole world map. As we keep zooming in, we get zoom level 1, zoom level 2, and so on. If you zoom in a picture p at a level, it will need 4 pictures in the next level for the same location of picture p. Requests to pictures on Sogou image servers like pic1.go2map.com have the predefined format as /seamless/0/*k*/*z*/*a*/*b*/*x*_*y*.GIF, where *k* shows category information such as map, satellite, etc. Zoom level equals to 728 – *z*. For the levels above zoom level 4, Sogou only provides pictures around China, and the coordinate of left bottom corner and right top corner of the picture in these levels can be calculated by

$$\left(x/2^{zoomlevel}, y/2^{zoomlevel}\right) \text{and} \left((x+1)/2^{zoomlevel}, (y+1)/2^{zoomlevel}\right)$$

respectively, where x, y come from the URL. It's worth mentioning that other map web sites may have similar URL format, eg. Baidu Maps.

With this knowledge we can dig out the distribution of the requested pictures' distribution for Sogou Maps. Altogether we have 2,249,373 picture requests of Sogou Maps. A visit to a picture contributes a point uniformly shared by the range it covers. Figure 9 and Figure 10 show how many points every location gets. We can see in the picture of China (Figure 9), requests concentrate at large cities or famous cities for travelling (eg. Dalian), and around Beijing where our data collect point locates. In the picture of Beijing (Figure 10), requests concentrate around the data collection point (Tsinghua University, Point A marked with red circle) and the center of Beijing. Another hot point is the place where Beijing 2008 Olympic Games were held.

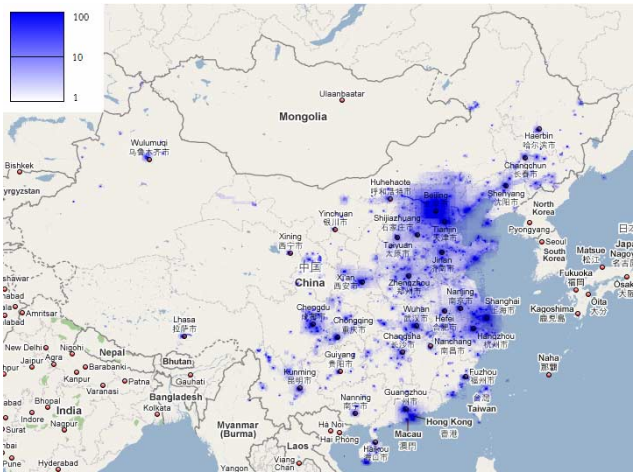

Figure 9 Requested pictures distribution for China

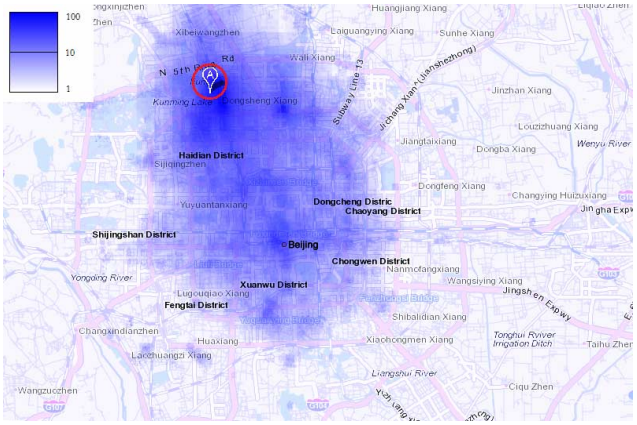

Figure 10 Requested pictures distribution for Beijing

From these figures we found that the requested pictures have significant hotspots as we expected. In addition, we can see that

hotspots are related with user location. We can safely infer that requests from different locations have different hotspots, thus CDN servers are probably suitable for helping improve the performance of web based map applications through utilizing this traffic characteristics.

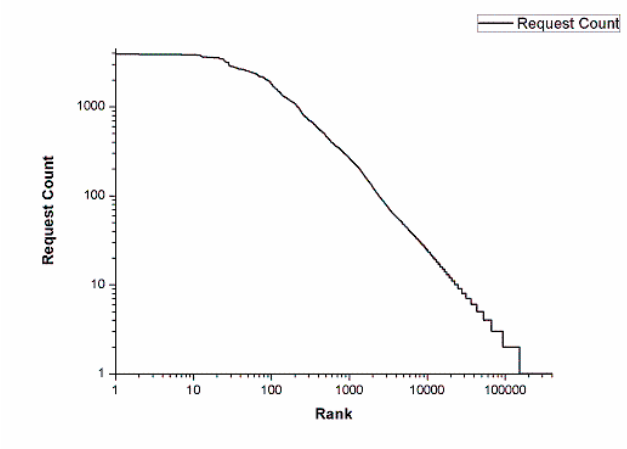

Figure 11 Ranked picture request count

For another important factor in cache policy design, we examined how dense requests of pictures really are. We counted all of the requests for the Sogou Maps pictures and sort them according to the requested count. We draw the rank of picture and request count of picture in a log-log coordinates (Figure 11). Main part of the result (rank > 100) follows the Zipf's Law[12]. The most popular pictures are the map pictures shown in home page of Sogou Maps, followed by the map pictures shown if you select the city of Beijing in home page.

After examining the density of picture requests, we observed that 10% (about 1.8 GB) of the requested pictures in our captured data took up 74.6% of the total requests, 20% of them took up 81.8% of the total requests. Since not all the pictures appear in our capture data, less than 10% of the total available pictures should take up 74.6% of the total requests.

For the analysis above, we can infer that the location sensitive hotspots and high density properties of map pictures requests give good performance of CDN servers. The map pictures took up about 68.7% of total Sogou Maps traffic. We can conclude that CDN servers are quite useful for accelerating web based map applications.

4. CONCLUSION AND FUTURE WORK

In this paper, we captured basic HTTP information of Web 2.0 based map application at our campus network while not violating user privacy. We collected and examined the traffic from four popular map web sites including Google Maps, Yahoo Maps, Baidu Maps and Sogou Maps. More than 800 hours of traffic data (more than 100,000 sessions and more than 6,000,000 requests were collected) were collected that could reflect workload trends of these map web sites from campus network perspective.

It is the first time that the mash-up characteristics of Google Maps traffic was analyzed through a method proposed in this paper. This method can be applied to other mash-up analysis work. We also analyzed other characteristics of web based map applications and drew the conclusion that map applications are more and more popular in mash-up. For example, 40% of Google Maps sessions come from mash-up from other website.

Furthermore, we found that cache is still useful in web based map applications, because the requests to pictures follow the Zipf's Law. In additional, requests from different locations have different concentrations, thus CDN servers are suitable for utilizing this attribute and should have a better performance in web based map applications. CDN server can help Web 2.0 based web applications to enjoy a shorter IRT.

Due to page limit and the similarity with [9] in the study of AJAX characteristics, the study of AJAX characteristic is not discussed here. For future works, we are interested in the comparison of mash-up and not mash-up traffic model, web caching and some further study of AJAX characteristics.

5. ACKNOWLEDGMENTS

We would like to thank network administrators of Tsinghua Univ. Campus network, Jilong Wang and Qianli Zhang to help collect the traffic data. This research is supported by NSFC-RGC Joint Research Project (20731160014), 973 Project of China (2009CB320501), 863 Project of China (2008AA01A326), and Program for New Century Excellent Talents in University.

6. REFERENCES

[1] http://en.wikipedia.org/wiki/AJAX

[2] http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)

[3] M. Arlitt and C. Williamson. 1997. Internet Web Servers: Workload Characterization and Performance Implications. IEEE/ACM Trans. on Networking, 1997(5): 631-645.

[4] Phillipa Gill,Martin Arlitt et al. 2007. YouTube Traffic Characterization: A View From the Edge. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. (San Diego, CA, USA, October 24-26, 2007). IMC'07. ACM Press, New York, NY. 15 – 28. DOI=http://doi.acm.org/10.1145/1298306.1298310

[5] Alan Mislove, Massimiliano Marcon et al. 2007. Measurement and Analysis of Online Social Networks. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. (San Diego, CA, USA, October 24-26, 2007). IMC'07. ACM Press, New York, NY. 29 - 42. DOI=http://doi.acm.org/10.1145/1298306.1298311

[6] Fabian Schneider, Sachin Agarwal, et al. 2008. The New Web: Characterizing AJAX Traffic. Proceedings of the 9th International Conference on Passive and Active Network Measurement. PAM'08. Springer Berlin / Heidelberg. 4979: 31-40.

[7] http://www.bro-ids.org

[8] http://www.stearns.org/doc/pcap-apps.html

[9] http://www.qosient.com/argus/

[10] http://www.snort.org/

[11] http://code.google.com/apis/maps/

[12] G. Zipf. 1949. Human Behavior and the Principle of Least Effort. Addison-Wesley (Reading MA).