24 Hours in the Blogosphere

Matthew Hurst
Intelliseek, Inc/BlogPulse.com
mhurst@intelliseek.com

Abstract

The majority of weblog posting events are recorded and distributed by ping services. Weblogging is an international phenomenon, a fact reflected in the geographic information presented in many online profiles. By examining 24 hours of ping data from Weblogs.com, a popular ping server, and combining temporal analysis with URL decomposition and the profile information associated with webloggers, we can begin to understand a number of aspects of weblog posting behaviour.

Introduction

On the 28th of July 2005, Weblogs.com broadcast 816, 704 pings - each ping indicates some activity around a resource on the web. In this paper we analyse this single day's data and use it, together with additional resources, to get an understanding of the blogosphere and its denizens.

This paper investigates the day's ping data with the following goals:

- to understand the type of analytics that can be acheived over large volumes of blog data - one clear application is a system that monitors the data stream and reports concrete observations (i.e. monitoring observable events, not predicted or implied events).
- to understand the assumptions that are made when making observations - for example, the investigation uses profile data drawn from personal information volunteered by blog authors, but is that data reliable?
- to generate hypotheses about the blogosphere - a set of statements that may be further investigated with other data sets.

This paper is a field study of the blogosphere.

Ping Servers

Weblogs.com is one of several ping servers. A ping server acts as a centralized dispatch service for the notification of updates to web resources, specifically RSS feeds and web logs (blogs). The intended use is: whenever a web hosted resource (in the case of weblogs.com - specifically weblogs) is updated, it informs the ping server. The ping server records this 'ping' with some minimal amount of information (the weblog URL and the time of the ping). Any system interested in updated weblogs may then *poll* the ping server for recent updates, digest the records and then process the results accordingly (a step which often involves hitting each of the updated resources and downloading the latest version).

The mechanisms and structured content involved in this process tend to be light-weight and require only the loosest of validations - there is no mechanism or declaration to determine if the URL actually is a blog, if it has actually updated and so on. The entire architecture revolves around a system of trust.

As RSS aggregation and feed reader clients of one sort of another have become more and more popular, so too have ping servers become more and more visible, attracting new and interesting varieties of use. For those interested in weblogs, ping servers represent a vital central point for integrating publication events into a collection mechanism. However, they have also introduced problems into the blog specific arena including spam and non-blog RSS updates.[1]

The Data

Weblogs.com archives ping information in hourly chunks per day. Each hour of pings is available as an html file (at the time of writing, the xml archives were not available). Each item in the html file records an ordinal, a url/title and a time (at the minute granularity). For example, the first hour on the 28th of July 2005 contains 29, 634 pings, the first of which is

```
1.  Reality TV Magazine 12:59 AM
```

and the last of which is

```
29634.  Deborah Elizabeth Finn 12:00 AM
```

This raw level of data suggests some immediate forms of analysis: URL analysis and temporal analysis.

[1] As of the time of writing, most weblog search engines attempt to separate true blog data from other types of updates that use ping servers. One engine (IceRocket) has sidestepped this complexity by claiming to be something broader than a true weblog search engine.

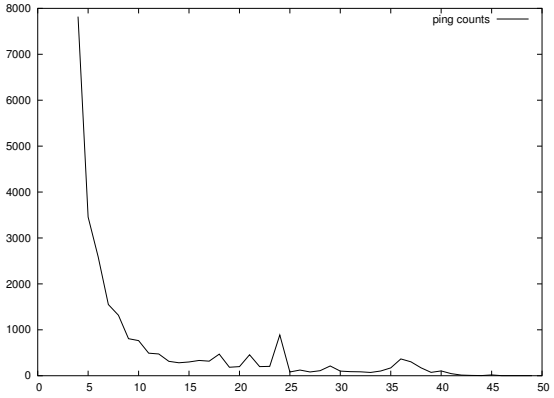

Figure 1: The number of URLs per total pings received by Weblogs.com

URL Analysis

We first consider the entire URL, theoretically, indicating a unique weblog; secondly, we have domains: important as many blogs are hosted by a service which is unique at the domain level (e.g. typepad, livejournal, xanga, blogspot); finally, we have hosts: these are interesting as many unique blogs are identified by hosts information dynamically mapped to a users blog.

Table 1 shows a breakdown of the number of pings received from individual URLs. 51.74 % of all pings are accounted for by URLs that pinged only once in the 24 hour period. The highest number of pings was 49, received from `http://neoamin-net.staging.daemon.co.za`

Ping Count	Number of URLs	% of all pings
1	422, 526	51.74
2	40, 166	9.84
3	14, 737	5.41
4	7, 821	3.83
5	3, 456	2.12

Table 1: Counts of URLs by number of pings.

If we plot the count of URLs against the number of pings received, it is not surprising to see rapid decay as the data in Table 1 suggests. What is notable is that when the details of the tail (shown in Figure 1) are plotted, the decay is not smooth. In fact, there is a clear spike at the 24 count. This is the first trace that we see of non-human posting mechanisms at work. A random sample of 100 URLs taken from the total of 503, 187 URLs and hand labeled resulted in 54 % spam blogs. Clearly, Weblogs.com is a highly compromised service.

Table 3 shows the pings per domain, the URLs for each domain and the ratio of pings per URL - the average number of pings from URLs in the domain. The majority of domains average between 1 and 2 pings per URL, something that we might believe is normal posting behaviour. Only five of those domains listed have ping ratios above 2.0:

Ping count	Host
178, 128	http://spaces.msn.com
53, 049	http://search-now130.com
40, 231	http://findallarticles.com
38, 516	http://search-now140.com
28, 491	http://www.look4articles.com

Table 2: Ping counts per host.

Domain	pings	URLs	ratio
blogspot.com	262, 702	62, 012	4.24
msn.com	178, 128	146, 917	1.21
search-now130.com	53, 049	53, 041	1.00
findallarticles.com	40, 231	40, 225	1.00
search-now140.com	38, 516	38, 509	1.00
look4articles.com	28, 491	28, 478	1.00
findactions.com	23, 054	23, 049	1.00
myblog.de	13, 537	5, 281	2.56
find4news.com	13, 378	13, 375	1.00
livejournal.com	6, 509	3, 919	1.66
wefindforyou.com	5, 972	4, 472	1.34
blogdrive.com	3, 734	2, 604	1.43
persianblog.com	3, 400	2, 547	1.33
blogspirit.com	3, 162	568	5.57
blogfa.com	2, 979	2, 231	1.34
hautetfort.com	2, 851	312	9.14
typepad.com	2, 720	1, 627	1.67
cocolog-nifty.com	2, 540	2, 001	1.27
livedoor.jp	2, 228	1, 350	1.65
obtainamazing.info	1, 920	719	2.67

Table 3: Ping counts and number of unique URLs per domain.

blogspot.com, myblog.de, blogspirit.com, hautetfort.com and obtainamazing.info.

Temporal Analysis

If we plot the pings per minute for some of the top ranked domains, we can observe some remarkable differences. Figure 2 contains two time plots. The first compares Blogspot and Spaces and the second compares search-now130 and findallarticles. In the first, we can observe a strong trend which peaks around minute 650 (approximately 11 am). Intuitively, we might describe this pattern as a natural pattern of blog posting and make the following hypothesis:

HYP 1: *Bloggers post at a steady rate throughout the day with a surge around 11 am as well as a lesser surge approaching midnight.*

Further analysis, including the inclusion of location information pulled from profiles will modify this hypothesis considerably.

If we compare this strong 'natural' trend with the flat trend shown in the second graph in Figure 2 we arrive at a further hypothesis:

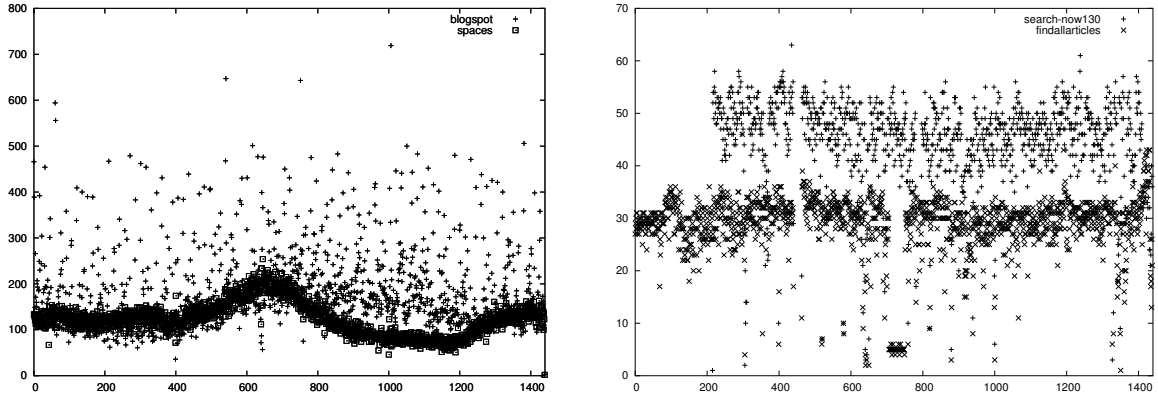

Figure 2: Pings per minute for Blogspot and Spaces (left) and pings per minute for search-now130 and findallarticles (right).

HYP 2: *In aggregate, bloggers post in predictable but non-uniform patterns throughout the day.*

In other words, humans look like humans and robots look like robots.

The variance in post counts per minute for Blogspot and Spaces is further analysed in Figure 3 which clearly illustrates the fact that Blogspot has a wide variance compared with that of Spaces. This leads to another hypothesis regarding the temporal profile of hosted blogs:

HYP 3: *Variance in posts counts per minute for a hosted blog system is correlated with the percentage of spam blogs hosted by that system.*

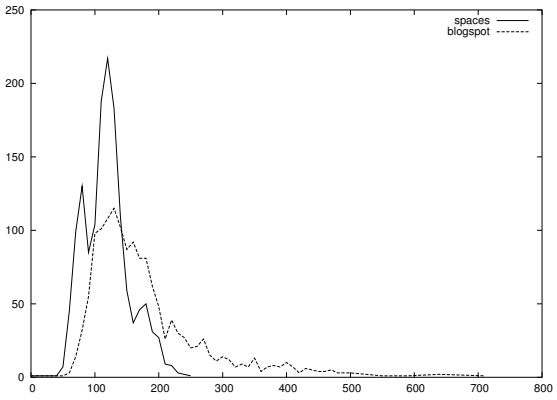

Figure 3: Distribution of pings per minute for Spaces and Blogspot.

Profile Analysis

The previous two sections have looked at analytical possibilities and results for the object data provided by Weblogs.com. In what follows, we use information drawn from blog profiles to segment the data by a number of features. Before presenting this analysis, we first give a brief introduction to blog profiles and how the information is captured.

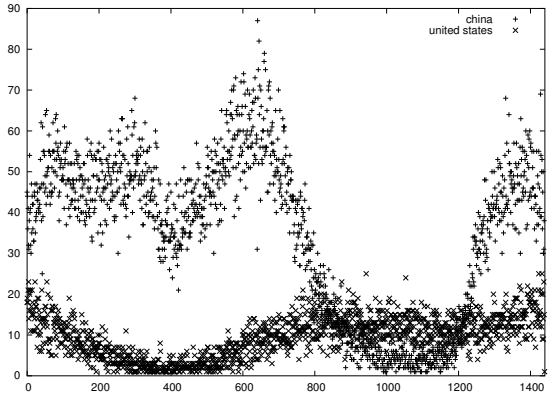

Figure 4: Ping times for China and United States Space bloggers

Most blog hosting systems require or allow the user to register certain types of profile information. This often includes name, age, gender and location. The author may elect to have this information displayed, in which case it may appear in one or more places including the blog itself and a special profile or about-me page.

As the information is generally of a simple key-value pair type, the presentation is often a simple highly structured part of the page. To extract this information, we use a method called automated wrapper induction (similar in motivation to supervised methods such as (CHJ02), (Bre03)). The specific technique is not described in detail here, but briefly it:

- collects all the text nodes in an html document and represents them using a combination of tag path (representing structural and presentational tags in different spaces) and text features,
- groups text nodes that are 'similar' according to a measure based on the tag path representation,
- clusters sequences of text nodes drawn from groups of similar nodes.

The results of this system are filtered by text strings

Feature	Spaces	Blogspot
Location	119, 954 (81.65 %)	21, 304 (34.26 %)
Age	83, 327 (56.72 %)	12, 462 (20.14 %)
Occupation	20, 403 (13.89 %)	11, 830 (19.03 %)

Table 4: Profile feature yields for Spaces and Blogspot.

that are strong indicators of profile information, such as `gender`, `age` and so on.

By running this system within a simple crawl that either passes the blog front page or looks for a known profile page (e.g. in the case of blogspot it searches for a link in a particular location on the page) we can derive profile data for many of the bloggers in the data set.

Extracting Profile Information from Weblogs

When extracting features from profiles, some criteria are needed for determining whether or not the value found is good or not. For these experiments we used some very simple tests. For location and occupation, a valid field simply had to have some text in it. For age, the field had to contain a sequence of numbers. Of course, there is some risk in this apprach, particularly for location and occupation values. Some percentage of values will be good locations, for example, including country, state and metro names. Others will include some amount of unconventional material (`Location: la la land`). This study does not aim to resolve this issue, but we note the problem.

Table 4 presents the counts of values retrieved from the profile harvest. It shows that for location and age, Spaces provides more data than Blogspot. Occupation shows the reverse with Blogspot having a greater share.

HYP 4: *Different hosting systems either attract users more likely to disclose personal information or are better at eliciting this type of information.*

HYP 5: *Blogspot users are more disposed to provide occupation information than Spaces users.*

Table 5 shows the break down of Spaces and Blogspot bloggers by country. The values used to segment that data are not strictly interpretations of the text found in the relevant profile fields. Rather, they are the last element of that field and therefore an object level string. This breakdown relies on the most common structure of these entries which presents metro, state and country level information. It should be noted that there are some cases where a country is not given, but where other information is provided from which a country could be inferred (e.g. `New York`).

The two tables show the top 20 countries by count for Spaces and Blogspot. The third column shows how many of these also provided age information. The next column shows the average age and the final column shows the per capita of bloggers as bloggers per million of population.

A broad comparison lets us see that

- the average age of Spaces bloggers is less than that of Blogspot bloggers.

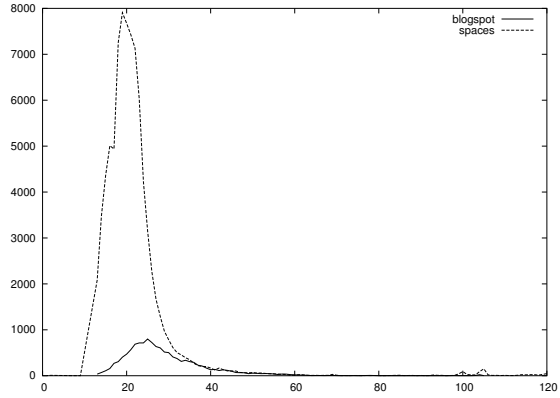

Figure 5: Age distribution in Blogspot and Spaces.

- the ranking of countries is considerably different, for example for Spaces `china` is ranked first whereas for Blogspot it isn't even in the top 20.

In addition, Table 5 shows the percentages of bloggers in each system who registered location information who also registered age. For example, 90.33 % of bloggers in Spaces that registered the location mexico also registered an age.

HYP 6: *There is a relationship between the location of a blogger and the service that they use.*

Finally, Table 5 shows the per capita of bloggers for each system by country. This value is per million of population. For example, 30.30 per million Chinese registered at Spaces and declared their location to be China. There are some very interesting outliers in this data. Taiwan, for example, has 472.95 per million capita in Spaces, compared with 40.53 for the United States.

HYP 7: *Asian countries have a higher blogger per capita ratio.*

Combining Temporal and Geographic Features

Figure 4 plots the pings per minute for bloggers using Spaces from China and those from the United States. As the time is normalized to some US time what we see as a large peak in the centre for China is actually bloggers blogging around midnight. In contrast, the US bloggers hit their stride in the run up to lunch time and don't let up until the wee hours. This analysis contests HYP 1 describing a general distribution of blogger ping times. In fact, projecting this sample onto the overall trend, we might hypothesise the total percentage of users from China based on the dominance of the peak around US noon. In fact, Figure 7 suggests that the main peak of activity in the US is around midnight, further supporting the explanation of the Chinese influence on the dominant peak found in Figure 2.

Figure 5 shows the distribution of ages in Blogspot and Spaces. The average age in Blogspot is 29.19, and the average for Spaces is 22.04. The average age for United States bloggers in Blogspot is 30.57. The average age for Chinese bloggers in Spaces is 21.76.

Spaces

Country	Count	w/Age	$a\bar{g}e$	per cap
china	39, 584	62.60	21.76	30.30
united states	11, 985	69.64	21.02	40.53
taiwan	10, 828	65.40	22.29	472.95
japan	7, 436	54.57	23.58	58.36
brazil	5, 817	71.67	22.38	31.26
united kingdom	4, 928	76.95	20.01	81.53
canada	4, 483	77.52	22.37	136.66
australia	4, 454	74.00	20.22	221.70
spain	3, 785	79.34	20.84	93.82
mexico	3, 259	90.33	20.56	30.69
france	2, 296	80.31	25.22	37.85
italy	2, 212	86.93	21.34	38.07
hong kong sar	2, 029	61.46	21.98	294.11
thailand	1, 672	76.61	20.78	25.55
netherlands	1, 368	81.36	28.81	83.38
argentina	1, 265	85.53	21.55	31.99
peru	675	80.00	20.98	24.17
belgium	577	81.98	27.51	55.67
chile	569	78.73	22.10	35.60
portugal	565	79.82	23.45	53.47
india	154	69.48	26.63	0.14
singapore	621	64.73	20.72	140.32
philippines	0	0	0	0
malaysia	450	77.56	21.10	18.79
germany	561	79.32	21.96	6.8
sweden	262	77.10	29.62	29.11

Blogspot

Country	Count	w/Age	$a\bar{g}e$	per cap
united states	9, 427	56.17	30.57	31.88
canada	1, 170	58.21	29.87	35.67
united kingdom	913	57.61	30.02	15.11
india	561	67.20	25.02	0.52
portugal	522	62.26	30.92	49.40
australia	478	58.16	27.50	23.79
brazil	470	67.66	27.10	2.52
singapore	460	72.82	24.28	103.94
malaysia	445	67.42	23.68	18.58
mexico	434	57.37	25.07	4.09
spain	420	67.86	29.78	10.41
chile	274	76.64	27.99	17.15
philippines	267	73.78	24.86	3.04
argentina	232	68.54	28.14	5.87
taiwan	187	66.84	26.67	8.17
france	184	61.96	30.78	3.03
germany	172	58.72	29.61	2.09
japan	169	52.07	30.00	1.33
sweden	164	63.41	28.39	18.22
netherlands	162	70.37	33.25	9.87
china	47	63.83	24.43	0.04
italy	98	66.33	29.48	1.69
hong kong	158	46.20	22.17	22.90
thailand	43	60.47	26.88	0.66
peru	51	70.59	28.39	1.83
belgium	64	75.00	31.83	6.17

Table 5: Geographic distribution for Spaces and Blogspot bloggers. For each host, the top 20 locations are listed. Each table is augmented with an addition 6 locations to make complete comparison posisble.

HYP 8: *There is a relationship between the age of a blogger and the service that they use.*

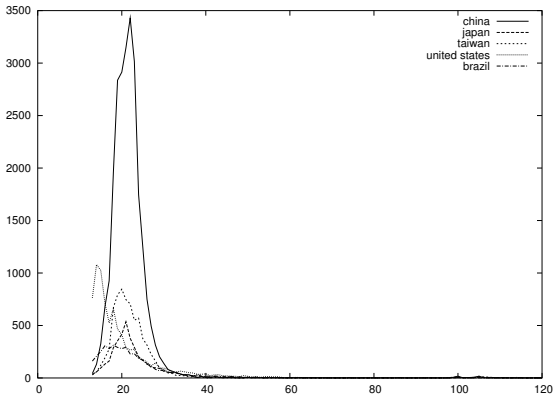

Figure 6: Age distribution in Spaces for different countries.

Figure 6 shows the distribution of ages for the top 5 countries in the Spaces segment. It illustrates the clear differences between countries. This kind of data has relevance for systems that aim to infer demographic information about bloggers.

HYP 9: *There is a relationsconthip between the location of a blogger and the age of the blogger.*

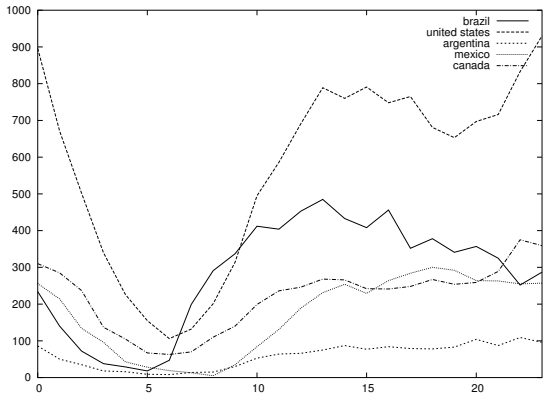

Figure 7: Pings per hour for a selection of American countries.

Figure 7 shows pings per hour for Space blogs for a selection of countries in the American continent. There appears to be three classes of patterns here. Firstly, as for Argentina and Mexico, a relatively flat trend with a dip in the nught; secondly, as for Brazil, a more curved trend with a dip at night; thirdly, as for the US and Canada, a peak or plateau followed by a second burst late at night followed by the later dip common to all.

HYP 10: *There are different patterns to blog posting that are culturally dependant.*

Conclusions

This paper has investigated a number of facets of bloggers, their publishing behaviour and the mechanisms used to broadcast publication. It has highlighted a number of potentially significant variations across factors such as the hosting system, the registered location of the blogger and the registered age of the blogger.

References

Thomas Breuel. Information extraction from html documents by structural matching. In *Second International Workshop on Web Document Analysis*, 2003.

William Cohen, Matthew Hurst, and Lee Jensen. A flexible learning system for wrapping tables and lists in html documents. In *Proceedings of International WWW Conference (11)*, 2002.