

A Large-Scale Study on Persian Weblogs

Vahed Qazvinian¹, Abtin Rassolian¹, Mohammad Shafiei¹, and Jafar Adibi²

¹ Computer Engineering Department, Sharif University of Technology, Tehran, Iran
{qazvinian, rassolian, m_shafiei}@ce.sharif.edu

² Information Science Institute, University of Southern California, USA
adibi@isi.edu

Abstract. Weblogs are becoming an important part of today’s web. Interactions between bloggers cause in the formation of a large social network in every blogosphere. Analysis of this network gives a lot of information in behavioral aspects of bloggers and blog readers. In this paper we introduce the largest dataset of Persian Weblogs that contains comments. Our contribution is twofold: first, we provide basic analysis on the blogosphere, and second we introduce a simple model for distribution of comments in Persian Blogs.

1 Introduction

Nowadays *weblogs*, (aka. *blog*) play an important role in social interactions. People who maintain blogs and update them, so called *bloggers*, involve in a series of interactions and interconnections with other people [8]. Each blog usually has a constant regular number of readers. These readers might make links to that blog or comment its posts which is said to be a motive for future postings[21]. Although, not until 1997, the term *blog* had been coined [5], today many people maintain blogs of their own and update it regularly, writing their feelings, thoughts, and any other thing they desire. Persian Weblogs have not been exceptions. There is a growing amount of interest in Persians to do blogging. As stated in [1] by October 2005, Persian weblogs were estimated to be about 700,000 (out of an estimated total value of 100 million blogs worldwide), of which about 40,000-110,000 are active , mostly written in Persian language. Of course this is not limited to Persian bloggers. Blogs are popular across the world [3,14], and Iran actually ranks 9th, in the number of blog users in the world.³

With the above discussion, the role of blogs as a social network is clear, and so, many researches are devoted to analyzing relations in blogs. From political conclusions [5] to finding mutual awareness and communities [16]. Unfortunately works on Persian weblogs are not so many and there is a vast area for new work in the field. This could be the main reason for us to devote our analysis on Persian weblogs.

In this paper we will present a large-scale study on Persian weblogs, and introduce the first and largest Persian weblog dataset.⁴ We will also have investigations on this dataset, and analyze the commenting behaviors of Persian bloggers.

³ <http://Persianweblog.com/articles/show.aspx?id=27>

⁴ To our knowledge it is the largest available blog dataset that contains comments

Our paper will also address a new model on the distribution of comments following a posting.

In the rest of the paper we will first overview some related works, in section 2. After that the data gathering process and characteristics is discussed in section 3. Section 4 and section 5 will describe link types in Persian blogs and basic analysis respectively. A model for distribution of comments over time is given in section 6 and at last the paper is concluded in section 7.

2 Related Work

Various studies have been done on the structure and size of blogspheres. [6] et al. have studied the macro and micro behaviors of blogspace dynamics. Political aspects of blogs are studied in [5]. Oka et al. [19] use frequency segments, and sequential occurrences of terms over time, to extract topics in weblogs. Kumar et al. [15] discuss extraction of bursty communities from blogspace through bursts of hyperlinkings using posts as the unit of analysis. There are also some other works.

Works on cross language blogs and studying structure of non internationally-used blogs are rare. Nanno et al. [18] have studied and monitored Japanese blogs. Works on Persian blogs are even less. To the knowledge of authors only SheykhEsmaili et al. [20] have studied Persian blogspheres in their paper, investigating series of hubs and authorities [7] with simple HITS [13] and PageRank [9] algorithms. The Persian data corpus used in [20] is rather small, not including comments and texts. This shows that there is a vast area for more research on Persian weblogs.

Above researches, and many others on weblogs, focus on post data. Yet few studies have been done on comments. Herrig et al. [12] studies a very small comment dataset of 203 weblogs (average of 0.3 comment per post). A larger scale study on comments investigates the relations of comments and posts, and extracts commenting pattern based on blog popularity [17]. Trevino et al. [21] and Gumbrecht et al. [10] study the importance of comments in blog analysis, and conclude that comments play a essential role in interactive nature of blogs. As we know, no work has been done on analyzing comments in Persian weblogs till the date of writing this paper.

3 Data

In this section we will describe the characteristics of our dataset. To choose target data, we needed to choose a weblogging host. Persianblog [4] is one of the most popular Persian blogging system providers. It actually was the most popular at the time of gathering data among hosts such as, Mihanblog⁵, Blogfa⁶, blogspot⁷,

⁵ www.mihanblog.com

⁶ www.blogfa.com

⁷ www.blogspot.com

and some others. From all Persianblog weblogs, we chose those with archives. A crawler was implemented to crawl and gather required data. Unfortunately RSS format was not available for many blogs so, in the next step we used an HTML parser to extract comment links, and posts information including post names, dates, times, descriptions, content, etc. After downloading all comments, we implemented an XML convertor and changed the files into our designed XML format. In our data, there are almost 80,000 XML files, each for monthly archive of every blog. Related information about data is available at [2].

3.1 Corpus Size

The data, contains archives of more than 22,000 weblogs in a 15 months period from March 20, 2005 to June 20, 2006. There are nearly 347,800 posts with a total number of 1,258,000 comments. This leads to an average of 3.6 comments per post. This average is comparably higher than that of previous corpus [12,17], for which the average value was 0.3, 0.9 respectively. Lets go over some of these corpus statistics in table 1.

Table 1. Basic analysis on corpus size

Number of Weblogs	22,306
Number of Posts	348,700
Number of Posts with Invalid IDs	492
Number of Correct Comments	1,257,561
Number of Incomplete Comments	89,349
Number of unavailable Comments due to crawl error	385
Accuracy	93%
Number of Commented Posts	339,884 (97.5%)
Number of Uncommented Posts	8,816 (2.5%)

There are errors with our XML files in comparison with the main HTML files. As stated in table 1 there is a 7 percent error in comment files. This means that in average 7% of all HTML comment files were not successfully converted to XML format and are ignored. This might have happened due to some unstructured and old pages of blogs which made our HTML parser halt.

4 Links and Graphs

Here a piece of theory which is of help is the notion of *typed graph*⁸. Lets assume that labels assigned to nodes are chosen from a finite alphabet Σ . Let $\lambda \notin \Sigma$ be a null character, and $\Sigma_\lambda = \Sigma \cup \lambda$. A *typed graph* is denoted by $G(V, E, T)$, where E is of type $E : V \times V \rightarrow T$. Labels in T are chosen from $\Sigma_\lambda = \Sigma \cup \lambda$.

⁸ There is no consensus in mathematics on this name

In such a graph an edge e between two vertices a, b with type t is denoted by: $e(a, b) : t$. With this definition in mind follow with the link types in our data.

In our data, four different types of links are considered.

- *Blog Roll Link* which is a hyperlink put in the side bar of blog page. These hyperlinks usually link to blogs or homepages of the blog maintainer’s friends, and are not updated very often.
- *Post Link* also we call it an *entry-to-entry* link, is a hyperlink put in the content (body) of a post. It might *probably* point to a related material or post, what so ever.
- *Comment outLink* is a hyperlink put in the content (body) of a comment.
- *Comment inLink* is a hyperlink put in the footer of a comment and points to the blog, homepage or email address of the comment leaver.

For each set of links, described above, we made a typed graph.

In each graph, vertices were representing blogs. For blog roll links types were null characters (λ), but for others, the type was chosen to be the date and time of the link. These graph made future investigations very simple.

5 Basic Analysis

After having dataset at hand, we came up with some basic and preliminary analysis. There were several interesting investigations. In this section we will discuss some of them. The primary analysis was on comments. As a first step we wanted to know how many percent of comment leavers were inside our blogosphere. This had not been not possible to investigate unless we ignored comments which had no inLinks (See Sec. 4). As given in table 2, even though there are fewer people who leave comments and have blogs in the blogspace, this minority leave the majority of comments (66%).

Table 2. Basic analysis on corpus size

Comment writers with blogs inside blogspace	58%
Comment writers without blogs inside blogspace	42%
Comments left from blogs inside blogspace	66%
Comments left from other links	34%

5.1 Comments in Weekdays

It was very interesting to see the behavior of bloggers in commenting others due to weekdays. For this one, we chose only comments in the corpus with inLinks (See Sec. 4). Number of weekly comments were extracted and separated by weekday. The diagram is shown in Fig. 1 a. There we got interesting clues about the data. It is clear in the diagram that the line for Friday underlies other lines most of the times. The next option is Thursday. This was the clue to extract the second diagram on weekends. As it is seen in Fig. 1 b, Fridays with 12% of all

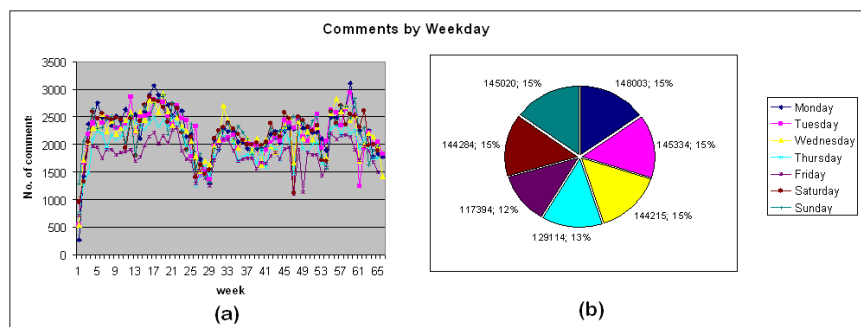


Fig. 1. Number of comments in each weekday of each week

comments contains the least amount. Thursday is the first runner up with 13%. Other days share an approximately equal number of comments. This could be due to weekend time in Iran, which is Thursdays and Fridays.

5.2 Outliers

In Fig. 2 some outliers are tagged. The chart shows the number of comments in each day of studying the blogspace. As you see new year holidays, and the beginning of the new academic year is a reason of low comments, and presidential election, is a triggering event. However we are satisfied with these investigations at this level and we are not going to extract a model for motivations in this paper.

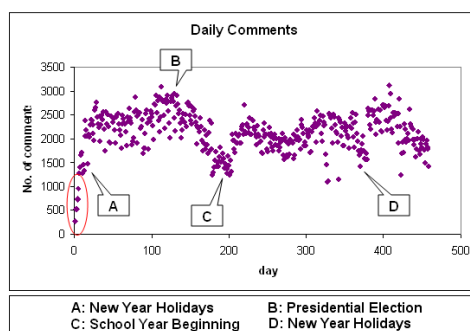


Fig. 2. Chart shows number of comments in each day, with tagged outliers.

5.3 Fall in Persianblog

Just like many other blog spaces, Persianblog has/had some *professional* bloggers. These bloggers are called professional for that, they update very often, and usually maintain the most popular blogs with the most number of visitors and links. These bloggers are not so many (hardly more than 400 people), and are known by most of the bloggers. As we talked to some expert bloggers, they admit that during the year of our study, Persianblog came with a *fall*. By the term

fall, we mean, that some professional bloggers left persianblog and ceased writing in persianblog and chose some other blog hosts. This was due to failures in persianblog efficiency, availability, support and policy. Especially when the host was sold to another company.⁹ The leaving of these professional bloggers may not change the number of comments in times, because that number is too large in comparison with comments of these blogs. But, terminating their blog, they actually affected the comment graph in another way; Betweenness Centrality. Actually we looked for an answer when we draw the chart in Fig. 3. For each month i we made a graph of inLinks (See Sec. 4) of that month and the months before that. And the decreasing chart was reached. According to [11]: “betweenness centrality views an actor as being in a favored position to the extent that the actor falls on the geodesic paths between other pairs of actors in the network. That is, the more people depend on me to make connections with other people, the more power I have.”, and powerful people were terminating their blogs and leaving Persianblog little by little. The end of the chart actually shows the attempts of persianblog stockholders to advertise and bring motivations.

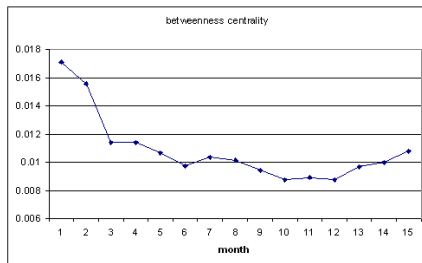


Fig. 3. Betweenness centrality of comments inLink graph over time.

5.4 Core of Comment Graph

The professional bloggers, whom we talked about in previous section, are on an unanimous agreement that most of the professional bloggers get to know each other, read each others' blog, and comment each other. This actually means that, these people are somewhat connected. To verify that in our data, we made a graph of comment inLinks (See Sec. 4). First of all we ignored all comments which were not from blog holders in persianblog cause we wanted to know how bloggers are connected. In the second step, we used a threshold for the sum of indegree and outdegree of each node, to clean the data. Then from the available nodes, we chose all pairs A, B if A had at least commented B once, and B had at least commented A once as well (mutual-acquaintance). In summary, we made the core graph based on the mutual-acquaintance with the degree threshold of 200 (See Fig. 4).

The edges in the graph of Fig. 4 show the mutual-acquaintance relation between nodes and are directed in both directions. In this graph there are 1322

⁹ A related article is available at: (Accessed, September 2006)
<http://www.sharghnewspaper.com/850128/html/media.htm#s396027>

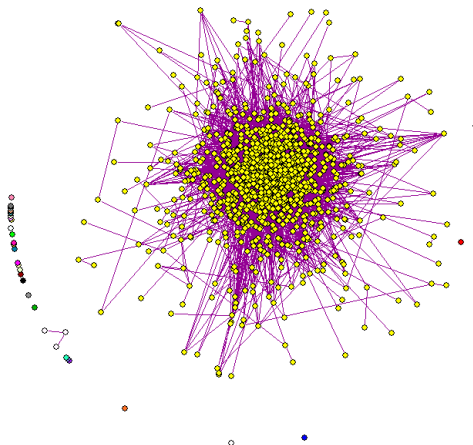


Fig. 4. Comment core graph with degree threshold 200.

vertices which form 33 Strong components. The major component contains 1288 vertices. There is another component with 3 vertices, while others are solitary nodes. This is a clear verification of connections in persianblog.

6 Comment Distribution

Before going on, let's give a definition that will help us in this section. We define $c_{i,j}$ to be the total number of comments left in the j^{th} day after their posts, which are in i^{th} day. So by definition, $c_{1,0}$ is the number of all comments of first day's posts, left on the post day. And $c_{3,10}$ is the number of all comments left ten days after their posts which are posted on the 3rd day.

With this definition in mind, please go back and have a look at Fig. 2. There is a clue in there. Pay attention to the four first points circled in red. These points show the number of comments in the first few days of the new year. They have a very low value. but this much low magnitude is not for holidays or any similar reason. There is another reason behind it.

Comments left for a post, come in days after it. Even a post may have comments after 30 days or more. This means that, for each day in the chart of Fig. 2, the total number of comments in that day, consists of comments for posts of that day and some comments for posts of previous days. Clearly speaking we have:

$$\text{Number of comments in } i^{\text{th}} \text{ day} = \sum_{1 \leq j \leq i} c_{j,i-j}$$

Yet for our data, we do not have the archives before our first day, so the total number of comments in first day, say, $c_{1,0}$, only contains the total number of comments left for posts of the first day on the same day.

This introduction was the main clue for us to look for a model. This model should provide us with a reasonable formula to compute the number of comments of all posts for some days after they are posted. Fig. 5 a, is a chart that for each day d ,

shows the total number of comments, left d days after their corresponding posts, divided by total number of comments left on their post day. that is, $\frac{\sum_k c_{k,d}}{\sum_k c_{k,0}}$.

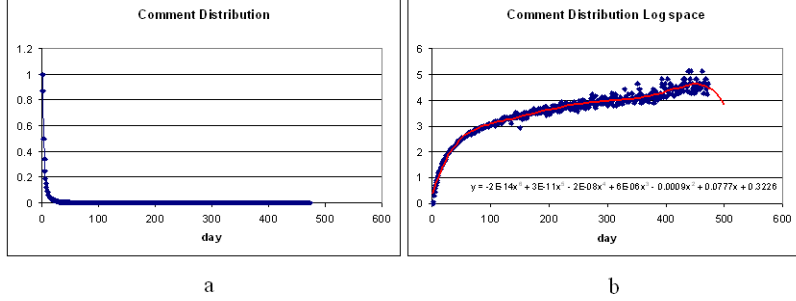


Fig. 5. Distribution of comments. (a) shows $\frac{\sum_k c_{k,d}}{\sum_k c_{k,0}}$ for each day. (b) shows $(-\log)$ of a .

In Fig. 5 b. you see a negative log scale chart of points in Fig. 5 a. In the next step we interpolated a polynomial function till 400 days in Fig. 5 b. We looked for the fittest polynomial with lowest degree, and 6 seemed to be the most appropriate one. This interpolation was done up to 400 points and so the function is valid for 400 days. The result is given as the following function:

$$f(x) = -2E-14x^6 + 3E-11x^5 - 2E-08x^4 + 6E-06x^3 - 0.0009x^2 + 0.0777x + 0.3226$$

So to find $c_{i,j}$ we should have,

$$c_{i,j} = c_{i,0} \times 10^{-f(j)} \quad (1)$$

The power-law model described for $c_{i,j}$ is based on the assumption that, comments after d days only depends on the number of comments in the first day. However this assumption may not be correct. Lets put it this way, There are some other parameter affecting number of comments, the post content, blog readers, etc. But for this step, we assume that all these factors somehow affect the number of first day comments. This presumption however may have more error.

As stated, There are several errors associated with current model. The error of the model itself, together with the error in data, and interpolation. But another error roots in the presumption made that comments in each day is only a function of comments of first day.

For the presumption error, we draw the actual graph of all comments for all posts of a day d , say C_d . This is equal to $\sum_{0 \leq k} c_{d,k}$. Besides we extracted all first day comments of all posts of the day d , equal to $c_{d,0}$ and using the model, we found an estimate for C_d , say \hat{C}_d (See Fig. 6).

To find the error, we computed relative error in each day of study. Actually for each day d we computed $e_d = \frac{|C_d - \hat{C}_d|}{C_d}$. The estimated value of e_d is, $E(e_d) = 14\%$

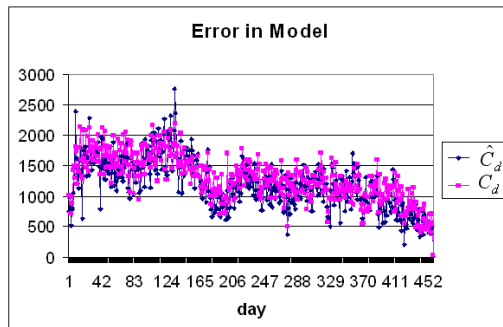


Fig. 6. C_d and its estimate \hat{C}_d over 458 day study

and the variance of error is $\sigma^2(e_d) = 0.012$. Now we can say that, the presumption that lies in the equation 1, that *the number of all comments left in the j^{th} day after their posts of a day, is only dependant to the number of comments of those posts in their first day*, has an accuracy of 86%. It should be taken care, that with this model, we are not talking about individual posts, because the model is extracted from cumulative comments of all posts in each day. Comments for individual posts may obey a more complicated model with many parameters, finding which could be a nice future work.

7 Conclusion and Future Work

The main goal of this paper was to introduce a new dataset on Persian weblogs. Alongside we provided detail statistics and analysis on the data that could be useful to know the characteristics of Persian weblogs. In addition we provided a simple, yet interesting, model for distribution of comments in Persian Blogs.

There are several directions for future work. First, further studies are inevitable to compare this blog space with others such as live Journal, Blogspot etc.

Second, there are certain events, as it was shown, that affect the behavior of bloggers. Some of these events are temporary, as new year holidays or presidential election and some are permanent, as for week- days. As an enhancement of our work, we would like to provide a model that describes bloggers' responses to these events. Third, as stated in Sec. 5.3, many bloggers quit after a short time. We are interested to find a mathematical model for this phenomenon and answer questions such as : Is there any model to describe the terminating blog fact? Do events affect the presence of a blog? What are the motives and reasons behind such termination? and other similar interesting questions.

Fourth, the dataset contains some other features that are not stated in this paper. Smiley emotions are one of them, and we believe it is good resource for a future research, on the behavioral aspects and moods of writers.

References

1. The blog herald blog count october 2005: over 100 million blogs created. available online: www.blogherald.com/2005/10/10/the-blog-herald-blog-count-october-

- 2005.
2. Blogscience. available online : www.blogscience.org/data.html.
 3. Globe of blogs - browse by weblog location,. available online: <http://www.globeofblogs.com/?x=location>. Accessed September, 2006.
 4. perisanblog. available online : www.pesianblog.com.
 5. L. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the WWW2005 Conference's 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*, 2005.
 6. E. Adar, L. Zhang, L. Adamic, and R. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*, 2004.
 7. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
 8. R. Blood. How blogging software reshapes the online community. *Communications of the ACM*, 47:53–55, 2004.
 9. S. Brin and L. Page. The anatomy of a large scale hypertextual web search engines. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
 10. M. Gumbrecht. Blogs as “protected space”. In *In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, at WWW 04: the 13th international conference on World Wide Web*, 2004.
 11. R. A. Hanneman and M. Riddle. *Introduction to social network methods*. Riverside, CA: University of California, Riverside (published in digital form at <http://faculty.ucr.edu/hanneman/>), 2005.
 12. S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *The 37th Annual Hawaii International Conference on System Sciences (HICSS04)*, 2004.
 13. J. M. Kleinberg. Authoritative sources in hyper-linked environment. *ACM*, 46(5):604–632, 1999.
 14. R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47:35–39, 2004.
 15. R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of the 12th International World Wide Web Conference*, pages 568–576, 2003.
 16. Y. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. Discovery of blog communities based on mutual awareness. In *Proceedings of the WWW06 Workshop on Web Intelligence*, 2006.
 17. G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*, Edinburgh, Scotland, May 2006.
 18. T. Nanno, Y. Suzuki, T. Fujiki, and M. Okumura. Automatic collection and monitoring of japanese weblogs. In *In WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.
 19. M. Oka, H. Abe, and K. Kato. Extracting topics from weblogs through frequency segments. In *Proceedings of the WWW06 Workshop on Web Intelligence*, 2006.
 20. K. SheykhEsmaili, M. Jamali, M. Neshati, H. Abolhassani, and Y. Soltanzadeh. Experiments on persian weblogs. In *Proceedings of the WWW06 Workshop on Web Intelligence*, 2006.
 21. E. M. Trevino. Blogger motivations: Power, pull, and positive feedback. In *Internet Research 6.0*, 2005.