Experiments on Persian Weblogs

Kyumars Sheykh Esmaili, Mohsen Jamali, Mahmood Neshati, Hassan Abolhassani and Yasaman Soltan-Zadeh
Computer Engineering Department
Semantic Web Research Laboratory
Sharif University of Technology, Tehran, Iran
{shesmail,m_jamali,neshati}@ce.sharif.edu, abolhassani@sharif.edu, y.soltan-zadeh@rhul.ac.uk

***Abstract*— Nowadays users of the Web are encouraged to generate content on the Web by themselves. In fact weblogs are one kind of social networks and they are one of the most important components in Web 2.0. There are a lot of Persian bloggers on the Web. In this paper we have tried to collect their blogs, produce some general statistics about them and have prepared a test bed for further research on weblogs in general and Persianblogs specially.**

I. Introduction

Social network analysis deals with mapping and measuring relationships and associations among people, groups, organizations and every other entities that can process information and knowledge. Nodes in such a network represent people and groups while edges show relationships among them. Social network analysis consists of visual and formal analysis of human relationships. The Web and its pages are a kind of social network, where pages are nodes and links between them are relationships. Also with the appearance of new generation of Web, known as Web 2.0, with Blogs and Wiki as its main components, the importance of social network analysis has been increased. A weblog or blog [1] is a personal page maintained by its owner as a single author which is updated based on his opinions in the chronological order. A blog has also a number of links to other ones. There are a variety of subjects for blog contents such as diary, photos, news and links to the other pages.

Since Persian language has special characteristics (such as encoding, font, right to left, etc.) creation of Persian blogs demands special facilities and as a result, the number of Persian blogs was very small before the appearance of Persian blog hosts which are specifically designed for Persian language. Because of this, famous sites like Technorati which works in blogs field has little work and statistics about Persian blogs. Fortunately currently there are a number of hosts for Persian language like Blogfa, Mihanblog, Parsiablog, Persianblog, and Blogsky. Now the interest among Persian natives for blogs is considerably high so that Iran has the 9th rank [1] in the world for the number of blogs. In this research we have used Persianblog [2] which is the largest and oldest Persian blog host including more than 50% of Persian blogs.

In this article we report our activities on Persian blogs. We applied a number of famous algorithms on them and analyzed the results. These activities include locating and gathering of the blogs, applying statistical analysis on them, and finally creation of a test bed for further activities.

The rest of the paper is organized as follows. Section 2 explains the blog gathering process and some general statistics of the gathered pages. Results of applying ranking algorithms are discussed in section 3, and section 4 describes the programming interface for this data set. Finally there are some conclusions of the results and future works.

II. General Statistics

In this section we explain in detail the process of gathering the pages and then we illustrate extracted statistics based on them. To find pages we have implemented a specific crawler. The crawler processes and gathers pages having http://weblogName.Persianblog.com or http://www.weblogName.Persianblog.com url patterns. On finding a page, the crawler first stores it and then applies the same work on its outlinks in breadth first order. For further processing, the results are stored in a MySql database. Since the blog graphs are not usually strongly connected, it is necessary to have a considerable number of blogs as the crawling seed. To do so we have selected a number of blog pages randomly from the user list of Persianblog. Normally,

[1] http://Persianweblog.com/articles/show.aspx?id=27

because of sparsity of blogs, there are a number of single blogs (those with no links to other blogs). To gather them we have processed all of the entries in the user list. Inter-blog links can be categorized into two groups. The first group includes links which directly link to the homepage address of a blog (in fact to those blogs that are among user preferences). The second group contains links which link to a specific note of a site, for example 'http://weblogName.Persianblog.com/#postNumber' or 'http://weblogName.Persianblog.com/date_weblogName /archive.html/#postNumber', such links are not showing a permanent preference but are temporarily links and therefore we ignore them. After a full crawl 106,699 blogs were discovered. There are 215,765 links between them which mean an average of 2.022 outlinks for each blog but the variance is high. According to the table I almost 45% (48,603 ones) of blogs are single ones, (they have no outlinks or inlinks). The frequency column in table I shows the number of components having the specified size. Also it is notable that 48% of the non-single blogs constitute a large single connected component with 20,8213 links and a ratio of 4.04 edges for nodes . The rest of blogs which are around 6% constitute small sized components.

No.	Size	Freq.	No.	Size	Freq.
1	51535	1	13	11	7
2	58	1	14	10	7
3	27	1	15	9	16
4	26	1	16	8	18
5	25	1	17	7	22
6	21	1	18	6	49
7	19	1	19	5	59
8	17	2	20	4	140
9	16	3	21	3	366
10	15	4	22	2	165
11	13	5	23	1	48603
12	12	2			

TABLE I

CONNECTED COMPONENTS IN PERSIANBLOG

III. RANKING THE BLOGS

In this section ordering of blogs with different ranking algorithms is explained. Since the importance of blogs outside the large single connected component is low (their size compared to large single component is very small), algorithms are only applied to the biggest component.

A. Rankings based on inlinks

It should be noticed that there exist a few anomalies in inlinks. For example http://vahidreza.persianblog.com has around 16,000 inlinks, but it's just because its author is the designer of a frequently used template and has embedded the link to his blog in the template. Since such inlinks are dummy, we do not consider them in the ranking algorithm. Table II shows top ten blogs ordered by their inlink count.

As noted before the pages outside Persianblog are not processed and we only keep the number of links from Persianblog pages to them and use such statistics to produce ranking for them. There exists 87,359 links from Persianblog to outside pages which consists of variety of pages . The ratio of inside Persianblog links to outside links is around 2.46 , therefore we can treat Persianblog as a separate social network.

Rank	URL	Number of In-links
1	fans	2925
2	delamgerefte	1896
3	link	1269
4	macromedia	1093
5	ghazalemoaser	264
6	mojganbanoo	231
7	rsaeedirad	212
8	iran-egold	205
9	varan	201
10	javascripts	198

TABLE II

RANKING OF BLOGS BASED ON THEIR INLINKS

List of 30 outside pages sorted by the number of links from Persianblog to them is shown in the table III. Based on the results, interesting analysis is deducible:

- Persian portals and the sites discussing on blog news and facilities to create them has highest rank
- Pages providing statistical facilities come in the second order (ranks 11 to 15 except 13).
- The last ranks in the table belong to the news web sites (BBC, IRIB, Baztab, Sharghnewspaper, ISNA).

It is necessary to mention that we ignored links to general web sites like Google and Yahoo because those links are not so valuable in our analysis.

B. PageRank Ranking

PageRank is presented by Page and Brin [3] to have an ordering algorithm for web pages. As noted in [4] calculations of this algorithm is done offline and is maintained as a stored value for each page. The value of this rank for each page is query independent and is calculated as:

$$R(A) = \sum_{B \to A} R(B)/outdegree(B) \qquad (1)$$

It is notable that the convergence of the algorithm is rather slow for Persianblog pages. It converged in 50 iterations. Table IV shows some pages with their associated PageRank value in different iterations. Of interesting points are the differences between this ranking and the ranking based on inlinks. It means that the linking patterns of bloggers are not homogenous and there is a high possibility for existence of many small sized communities.

C. HITS ranking

HITS algorithm was suggested by Kleinberg [5]. One of its applications is for exploring web communities related on a specific topic. For this purpose the algorithm introduces two different concepts: Authority pages which have useful information for the topic, and Hub pages having high number of links to authority pages. There is a dual relationship among these two types of pages. It means that a page is a good Hub if it has links to good authorities, and a page is good Authority if it is linked from good Hubs. These definitions are formulated as below:

$$Hub(A) = \sum_{A \to B} Authority(B) \qquad (2)$$

$$Authority(A) = \sum_{B \to A} Hub(B) \qquad (3)$$

As mentioned in [4], unlike PageRank the computation of this algorithm is online and is dependent to the query. In this experiment Hub and Authorities are calculated in a general form, without considering a specific topic or query. Table V shows a portion of the results.

One of the interesting points is the convergence speed of this method, less than 20 iterations, compared to PageRank.

If we assume the Authority values as page ranks, then the results of this algorithm is somehow similar to the ranking based on inlinks (4 commons out of 10 first blogs) but it has no similarity to PageRank. If we compare the Hub values to list of blogs having most values of outlinks, there is not any specific similarity.

Rank	URL	Number of Out-links
1	almofid	243
2	o0	241
3	hamgh	231
4	saberkarimi1	224
5	nale	212
6	little-king	188
7	saadedel	187
8	bingbang	185
9	firend2	181
10	behrokh1	174

TABLE VI
RANKING OF BLOGS BASED ON OUTLINKS

IV. TEST BED

We have compiled data gathered in this research as a standard test bed for future researches. In this test bed the following information exists:

- List of all crawled blogs
- List of links between nodes in this graph
- List of all connected components
- Calculated ranks for largest connected component based on inlinks, PageRank and HITS

To facilitate access to such data we exported the values from MySql to Microsoft Access in a mdb file format which can be processed without the need for a specific driver. We've also implemented an API to use the facilities we prepared for blogs (such as blog's inlinks, outlinks, rankings and etc.). The API is available at (http://ce.sharif.edu/∼shesmail/Persianweblogs). In articles like [6], and [7] for the compression of web graphs interesting techniques have been introduced, but because the url patterns for our problem area is fixed there is no need for such compression techniques. For each blog we only store the weblogname as url in the database.

There are many new research possibilities on this test bed. For example in [8] this test bed

Rank	URL	In-Links	Rank	URL	In-Links
1	http://www.Persianweblog.com	5174	16	http://www.Persiantalk.com	647
2	http://weblog.gardoon.com	5015	17	http://www.dev.ir	627
3	http://www.balmasque.blogspot.com	4898	18	http://www.tebyan.net	594
4	http://www.Persianyahoo.com	4044	19	http://www.sharghnewspaper.com	593
5	http://pb.Persianweblog.com	2235	20	http://www.eshgh.ir	569
6	http://www.sharemation.com	1918	21	http://www.isna.ir	561
7	http://www.yourname.com	1761	22	http://www.e-gold.com	554
8	http://www.bbc.co.uk	1467	23	http://www.baztab.com	514
9	http://www.irantemp.com	1306	24	http://www.Persianpixel.com	465
10	http://explorer.blogsky.com	1212	25	http://www.lostlord.com	460
11	http://stats.netsups.com	1001	26	http://www.naghmeh.com	446
12	http://www.nedstatbasic.net	994	27	http://www.bloglet.com	433
13	http://mazash.blogspot.com	925	28	http://www.irib.ir	420
14	http://v1.nedstatbasic.net	776	29	http://www.parseek.com	365
15	http://www.pagerank.net	760	30	http://www.linkestan.com	364

TABLE III

FAMOUS EXTERNAL SITES

Rank	URL	PR(20)	URL	PR(30)	URL	PR(40)	URL	PR(50)
1	iranreform	1	iranreform	1	iranreform	1	iranreform	1
2	faryadebeseda	0.489	faryadebeseda	0.493	faryadebeseda	0.495	faryadebeseda	0.496
3	mastegoleyas	0.450	mastegoleyas	0.454	mastegoleyas	0.454	mastegoleyas	0.454
4	sharpmusic-chod	0.440	raze-nahofte	0.440	raze-nahofte	0.440	raze-nahofte	0.440
5	raze-nahofte	0.437	sharpmusic-chod	0.387	valse1	0.369	valse1	0.369
6	sharpmusic-musicw	0.377	valse1	0.369	yadebaran	0.368	yadebaran	0.368
7	sharpmusic-events	0.377	yadebaran	0.367	vahy	0.350	vahy	0.351
8	sharpmusic-designer	0.377	vahy	0.350	ranginkamaan	0.350	ranginkamaan	0.350
9	sharpmusic-classical	0.377	ranginkamaan	0.350	linkestaan	0.350	linkestaan	0.350
10	sharpmusic-roundtabl	0.375	linkestaan	0.350	shahidan	0.350	shahidan	0.350

TABLE IV

PAGERANK VALUES FOR DIFFERENT ITERATIONS.

is used to design and implement a blog recommender system. The test bed is available at http://ce.sharif.edu/∼shesmai/Persianweblogs.

V. CONCLUSIONS

The primary goal for this research was to provide essential tools and facilities for researchers interested in new generations of social networks. Secondly it provides means to do some initial researches on the data. For example in this paper analysis of hyperlink analysis algorithms are discussed. Recommendation is another possible application. As the last goal we can mention about researches of social aspects. In fact with the provided test bed it is possible to test various hypotheses.

As mentioned in this research we only used the pages in Persianblog. We intend to include other Persian blog pages in our future works. The pages are in two groups:

Rank	URL	Authority	Hub	URL	Authority	Hub
1	3kseke	0.0132	1	fans	1	0.0175
2	hoviyat-i-gomshodeh	0.0017	0.9004	ghazalemoaser	0.1554	0.0205
3	delltang	0.0018	0.8976	varan	0.1274	0.2647
4	daryaagarbashad	0.0001	0.8971	mojganbanoo	0.1257	0.0985
5	kashkool2	0.0043	0.8952	mostasharnezami	0.1123	0.2687
6	yaali110	0.0009	0.8890	rsaeedirad	0.1067	0.2343
7	hezareh3	0.0017	0.8866	ghazaleemrooz	0.1051	0.0062
8	iresa1369	0.0023	0.8848	mfaraji	0.1035	0.2131
9	mosaferezaman7	0.0022	0.8840	nirvana	0.0970	0
10	javabet	0.0003	0.8780	ololon	0.0910	0

TABLE V

LISTS OF BEST HUBS AND AUTHORITIES IN PERSIANBLOG.

blogs hosted in hosts specific to Persian blogs. There is a small number of such sites and it is possible to apply the same methods discussed in the paper to process them. The second group is those blogs hosted in general hosts. We intend to use two types of information for finding Persian blogs in such sites. One is the encoding used in site and another one is the links from first group of blogs to them. Another extension we have in mind is the usage of contents of the pages and summarizing such information.

VI. AKNOWLEDGMENTS

Some parts of crawling operations are developed by students of Modern Information Retrieval course. The authors thank of Iman Sadghi, Siavash BenAbbas, and Morteza Alamghir.

REFERENCES

[1] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura, "Automatically collection and monitoring of japanese weblogs," New York, USA, 2004.

[2] Persianblog. [Online]. Available: http://persianblog.com

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998. [Online]. Available: citeseer.ist.psu.edu/brin98anatomy.html

[4] M. R. Henzinger, "Hyperlink analysis for the web." *IEEE Internet Computing*, vol. 5, no. 1, pp. 45–50, 2001.

[5] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999. [Online]. Available: citeseer.ist.psu.edu/kleinberg99authoritative.html

[6] P. Boldi and S. Vigna, "The webgraph framework i: Compression techniques," 2003. [Online]. Available: citeseer.ist.psu.edu/boldi04webgraph.html

[7] S. Vigna and P. Boldi, "The webgraph framework ii: Codes for the world-wide web." in *Data Compression Conference*, 2004, p. 528.

[8] K. S. Esmaili, M. Neshati, M. Jamali, J. Habibi, and H. Abolhassani, "A link structure based weblog recommender system." in *Submitted to WWW2006 Workshop on Weblogging Ecosystem, Edinburgh, Scottland, May 2006.*