BlogDisc: A System for Automatic Discovery and Accumulation of Persian Blogs

Kyumars Sheykh Esmaili
Web Intelligence Laboratory
Sharif University of Technology
Tehran, Iran
shesmail@ce.sharif.edu

Hassan Abolhassani
Web Intelligence Laboratory
Sharif University of Technology
Tehran, Iran
abolhassani@sharif.edu

Zeinab Abbassi
Web Intelligence Laboratory
Sharif University of Technology
Tehran, Iran
z_abbasi@ce.sharif.edu

Abstract — One of the important elements of the new generation of the Web is the emergence of blogs. Currently a considerable number of users are creating content using blogs. Although Persian blogs have a short history, they have improved significantly during this short period. Because of fundamental differences between Persian and other languages, limited work has been done to analyze Persian blogs. In this work, a system named BlogDisc for automatic discovery and accumulation of Persian blogs is developed. This system uses content as well as link structure of the blogs. As an important part of this research, we propose an algorithm to recognize blogs that are not hosted on special blog hosts.

Keywords: Blogosphere, Weblog, Blog Discovery, Web Graph

I. INTRODUCTION

Nowadays, Blogs (also know as Weblogs) are known as an important element of the Web. Using them, users can produce content on the Web very easily. Informally, a blog is a personal page maintained by a person and is updated based on its author's thoughts regularly and in the order of time. Normally a blog also has links to other blogs. On the other hand the content of a blog can be highly varied: links, explanations about other blogs, photos and also diaries and news. A more formal definition for a blog is given in [4].

According to that definition, a blog is a web page which has dated posts sorted in reverse chronological order. Some other characteristics of them are as follows: containing the author's profile in its margins, using blogging tools and templates, personalized content, being read by special users, and membership in some active communities.

The first blog appeared in 1996 but they became popular in 1999 when blogger1 and other easy to use tools were announced. In the year 2002 an article in the Newsweek2 magazine estimated that there are half million blogs. A couple of electronic magazines' articles in year 2000 brought many attentions to them [5]. As reported by the Technorati3 site currently there exist more than 55 million blogs. They build a huge source of information and hence their analysis becomes very important. Generally two reasons for systematic study of the blog space can be mentioned:

• Social aspects: blogs have structural differences from normal web pages because they display a hierarchy of messages, like news groups, but each is written by a single person. However, there are more important structural differences: blogosphere culture is highly concentrated on small community interactions between a limited number of bloggers. Members of such informal societies put links to other blogs in theirs and often they leave some comments on each other’s posts. Often the thread of responses is very high in a short period but then suddenly it stops. Naturally we can study the structure of such highly dynamic societies and model them.

• Technical aspects: popular studies on the Web and its graph are based on snapshots and static pictures of it that are obtained by crawling. Those studies are concerned with what happens during long periods of time. Some of them are responding to this concern by generating and analyzing of sequences of snapshots [3,4,9]. Developing tools and methods to analyze such sequences demands time-consuming efforts. On the other hand, blogosphere has a technical advantage: for ordinary web pages if a new crawling is done it is impossible to find exact time for changes but since each entry in a blog has date and time associated with it, it is possible to provide a more exact image of blogspace evolutions [5].

Because of special characteristics of Persian language (right to left direction, encoding, fonts, etc.) creation of Persian blogs requires sophisticated tools. Therefore before the appearance of specific hosts for Persian language, the number of Persian blogs was too low. Fortunately at the present time, there are a considerable number of hosts for Persian bloggers. Nowadays Persian speakers show a lot of
interest in blogging such that currently Persian is among 10 most active languages in blogging\(^4\).\(^5\)

This research aims at discovering, gathering and analyzing Persian blogs. As an important part of this research, a system for automatic discovery and gathering of Persian blogs named BlogDisc is introduced that will be explained in this paper.

The rest of this paper is organized as follows. In Section 2, a review of the related work in automatic discovery of blogs and in particular Persian blogs are discussed. Then in Section 3 a general categorization for Persian blogs is introduced which has an important role in our algorithm. In Section 4, the main algorithm of BlogDisc is introduced. The implementation of this algorithm and some of the results are discussed in Section 5.

II. RELATED WORK

A. Persian blogsphere Analysis

Authors in [7] have gathered Persian blogs on a specific blog host – PersianBlog which is the largest and oldest host for Persian blogs. After crawling they have found 106699 blogs and 215765 links between them. Therefore, the average number of links in a blog is 2.022 but the variance is high. Near 45% of them are singular blogs having no outlinks. Also around 48% of them (51535 blogs) construct a huge single connected component having 208231 links which means there is 4.04 links per page in it. All of the links from Persian blog to outside it is about 87359 links covering a wide variety of destination pages. Blogs constitute a social network because the ratio of inlinks to outlinks is about 2.46. The authors compiled the gathered blogs in a test collection containing following information:

- list of all crawled blogs
- list of links between nodes in these blogs
- list of connected components
- ranks of blogs in the largest component (based on number of in-links, HITS and PageRank scores)

Also in [8] based on the same test collection, recommender techniques in Persian blogsphere have been evaluated.

B. Automatic recognition and gathering of blogs

In automatic blog discovery the only reported works are related to Japanese blogs [4,6]. In the papers, authors named the sites that are not hosted in special blog hosts and also does not use special tools for blogging as “web diaries”. In fact these two papers explain the differences between blogs and also provide techniques for automatic recognition of “web diaries” for Japanese blogs. In their view, a blog is a personal web page which has order for dated entries and contains its author thoughts and beliefs. The main approach introduced in the papers for automatic recognition of “web diaries” uses following modules:

- Crawling module: it tries to collect candidate blog pages from the Web. To do so it: 1) crawls the Web, 2) uses information of Blogs directories and their ping servers, and 2) uses the links between the candidate blog and pages which are recognized as blog by recognizer module.
- Recognizer module: it tries to select only pages that are blogs. Such recognition is based on the characteristics of blogs. A page is recognized as a blog if and only if a sequence of entries that are written for a given date can be extracted from it. Also entries should satisfy following conditions:
  - They should have a date and the date should appear in top of them
  - Dates in their entries should be consistent (e.g. 01-Jan-2003 and 2003/1/1 are considered to be inconsistent) and were ordered in ascending or descending order.
  - Post date should be in a chronological order

III. CATEGORIZATION OF PERSIAN BLOGS

It is possible to categorize Persian blogs to the following subcategories:

A. First group

This group includes the blogs that are hosted on Persian blog hosts; they are well-defined blogs which it means we are sure that they are blogs and ordinary web pages. The positive point of this group is that blog URLs have a specific pattern which facilitates discovery and recognition of them. The other positive point is that they can be used as starting points of crawling to find blogs in group 2 and 3.

The main Persian blog hosts which provide blogging tools are

- Persianblog\(^6\)
- Blogfa\(^7\)
- Mihanblog\(^8\)
- Blogsky\(^9\)
- Parsiblog\(^10\)

\(^4\) http://www.timesonline.co.uk/article/0,,1068-1957461,00.html
\(^5\) http://technology.guardian.co.uk/online/weblogs/story/0,1377538,00.html
\(^6\) http://www.persianblog.com/
\(^7\) http://www.blogfa.com/
\(^8\) http://www.mihanblog.com/
\(^9\) http://www.blogsky.com/
\(^10\) http://www.parsiblog.com/