Computers in Human Behavior 57 (2016) 195–207

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

irBlogs: A standard collection for studying Persian bloggers

Abolfazl AleAhmad*, MohammadSadegh Zahedi, Maseud Rahgozar**, Behzad Moshiri

Database Research Group, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University College of Engineering, University of Tehran, Iran

ARTICLE INFO

Article history:
Received 3 July 2015
Received in revised form
8 November 2015
Accepted 19 November 2015
Available online 25 December 2015

Keywords:
Persian bloggers
Social networks
Test collection
Weblog retrieval

ABSTRACT

A large number of internet users share their knowledge and opinions in online social networks like forums, weblogs, etc. This fact has attracted many researchers from different fields to study online social networks. The Persian language is one of the dominant languages in the Middle East which is the official language of Iran, Afghanistan and Tajikistan; so, a large number of Persians are active in online social networks. Despite this fact, very few studies exist about Persian social networks. In this paper we will study the characteristics of Persian bloggers based on a new collection, named irBlogs. The collection contains nearly 5 million posts and the network of more than 560,000 Persian bloggers which assures the reliability of the results of this study. Some of the analyzed characteristics are: the similarities and the differences between formal Persian and the language style that is used by Persian bloggers, the interests of the bloggers and the impact of other web resources on Persian blogosphere. Our analysis show that IT, sports, society, culture and politics are the main interests of Persian bloggers. Also, analysis of the links that are shared by Persian bloggers shows that news agencies, knowledge-bases and other social networks have a great impact on Persian bloggers and they are interested to share multimedia content.

1. Introduction

Persian is one of the dominant languages in the Middle East which is spoken in Iran (natively known as "فارسی" Farsi or "پارسی" Parsi), Afghanistan (officially known as "Dari"), Tajikistan (officially known as "Tajik") and some parts of other countries like Iraq, UAE, Bahrain, etc. Millions of Persian speakers use online social networks to share their knowledge or express their opinions in this popular media. Persians are ranked 9th in the world based on the number of bloggers (Megerdoomian & Hadjarian, 2010). The statistics of Alexa (Alexa 2015) show that 4 of 10 most popular Iranian websites are Persian weblog service providers. Although these websites are among the most visited in Iran, very few researches have studied them. One of the main reasons for this fact is the lack of a standard dataset which is a prerequisite for studying theories or developing algorithms in this field.

While blogs share some similar features with common webpages, they have some distinct characteristics; for example, there exist noisy data in blogs content and the language of bloggers is informal and conversational. This paper studies the characteristics of Persian bloggers by use of a standard collection of 560,000 + Persian weblogs, named irBlogs. In other words, first we will discuss the creation process of a standard dataset for studying Persian blogosphere and then the characteristics of Persian bloggers are analyzed based on the dataset. The main characteristics that will be analyzed are as follows:

- The subjects that Persian bloggers are interested to write about them
- The properties of the bloggers network (e.g. in-degree and out-degree of the network)
- The language style of Persian bloggers (e.g. the POS tag and other properties of the most frequent words that they use).
- The relation of Persian blogosphere with other web resources. In other words, we will analyze the links shared by Persian bloggers to answer the questions like: Which websites have more impact on Persian bloggers? Which type of web resources are shared by them?

The rest of this paper is organized as follows: Section 2 is a review of related works, Section 3 discusses the development process

* Corresponding author. School of Electrical and Computer Engineering, University College of Engineering, North Kargar St., Tehran, P.O. Box: 14395-515, Iran.
** Corresponding author. School of Electrical and Computer Engineering, University College of Engineering, North Kargar St., Tehran, P.O. Box: 14395-515, Iran.
E-mail addresses: aleahamad@ut.ac.ir (A. AleAhmad), s.zahedi@ut.ac.ir (M. Zahedi), rahgozar@ut.ac.ir (M. Rahgozar), moshiri@ut.ac.ir (B. Moshiri).

http://dx.doi.org/10.1016/j.chb.2015.11.038
0747-5632/

of the dataset which will be used for the analysis of Persian bloggers, then Section 4 analyses the characteristics of Persian bloggers and finally the paper is concluded in Section 6. irBlogs contains 45 standard topics together with nearly 24,000 relevance judgments. This fact makes it suitable for the development of blog retrieval algorithms to answer the information need of internet users from Persian blogosphere. So, some future research subjects will be introduced about this subject in Section 7.

2. Related works

There exist a number of standard collections for studying non-Persian weblogs; Blos06 and Blogs08 (Ounis, Macdonald, & Soboroff, 2008, Macdonald, Ounis, & Soboroff, 2009) are two widely known collections for this purpose. Blogs06 was used in TREC conference form 2006 to 2008 which is an 11 week snapshot of 100,000 blogs from late 2005 to early 2006. After that, the authors of Blogs06 tried to create a larger collection, named Blogs08, that contains more than 1 million blogs crawled from January 2008 to February 2009. These two collections are widely used in many other researches such as: (Lee, Na, & Lee, 2012; Macdonald & Ounis, 2008; Nunes, Ribeiro, & David, 2009; Weerkamp, Balog, & de Rijke, 2011).

But very few researchers have studied Persian social networks till now. A project in Rand research institute analyzed people's opinion about presidential election of Iran using Persian tweets (Elson, Yeung, Roshan, Bohandy, & Nader, 2012). They investigated more than 2.6 million Persian tweets of 124,000 users from 17 January 2007 to 28 February 2010. Their results show that Iranian political events can be analyzed using the opinions that Iranians share in Persian social networks.

In Alavi (2005) a few sample Persian weblogs are selected and analyzed from a sociological perspective. The authors conclude that because of the autonomous writing feature of weblogs, they are a good representation of Iranians society. Kelly and Etling investigated Persian weblogs from political and cultural perspectives (Kelly & Etling, 2008). After clustering Persian weblogs, they suggest that Persian weblogs can be categorized into 4 independent large categories: political, religious, literature and miscellaneous. They evaluated Persian weblogs as the 4th largest weblog space of the world. Also, in Golkar, (2005) the authors analyzed Persian weblogs from a political perspective and conclude that Persian weblogs are a suitable space for Iranians to express their opinion.

Some works have studied the language style of Persian bloggers (Megerdoomian, 2000, 2008, 2010; Megerdoomian & Hadjarian, 2010). The authors of Megerdoomian and Hadjarian (2010) investigated morphological properties of conversational Persian in weblogs. They could detect many new words that are devised by Persian bloggers. They state that the new words are either:

- Borrowed from other languages such as English or French
- Created by combining English and Persian words and postfixes
- Created by changing already existing Persian words

In Megerdoomian (2010), the author analyzed Persian weblogs morphologically and argue that Persian weblogs contain different kinds of formal and conversational texts; even the formal language style used in the weblogs is different from the commonly-used Persian formal language style.

3. Creation of a standard dataset

This section describes the development process of irBlogs. The main steps of creating the collection are: crawling weblogs, preprocessing the crawled webpages, extracting posts and blog features, extracting the bloggers' network and finally converting the collection into standard XML format. Each step will be discussed in the following subsections respectively.

3.1. Crawling weblogs

A crawler is an intelligent program that starts from a list of initial seeds, navigates the web network to find webpages and downloads them. It repeats this process automatically based on some predefined strategies until an ending condition is satisfied.

In order to crawl a good sample of the Persian blogosphere, a focused crawler is implemented and configured to crawl Persian weblogs only. An important issue before starting a crawler is to choose a set of high quality seeds. For this purpose, a list of 180,000+ Persian weblog URLs is used that has been already prepared carefully in AleAhmad and Habibian (2011). The crawling process continued for 10 days during which more than 900,000 unique Persian weblogs were detected.

After processing the crawler's log, some invalid blog URLs were detected and removed. Table 1 depicts the main reasons that caused the invalid blogs.

The banned weblogs (last row of Table 1) are crawled again using an internet connection that is not filtered by the government. Finally, nearly 200 K invalid weblogs are removed from the final result of this step.

3.2. Preprocessing

In this step, all HTML files that were downloaded in the previous step are verified and invalid webpages are eliminated (e.g. those with no content). Also, as the collection will be finally stored in XML format, all invalid XML characters are deleted. Other preprocessing activities like code page detection and Unicode conversion are also carried out in this step.

3.3. Creating the final collection

After extracting the bloggers network and their posts, the network is stored in form of two text files. Also, the posts are stored in standard XML format. Fig. 1, depicts the DTD schema of the XML files that are used to store different features of the weblogs:

Table 2 describes each XML element of the DTD.

Also, a sample document of the collection is depicted in Fig. 2 to clarify the structure of the collection:

3.4. The properties of irBlogs

Table 3 shows some important statistics about irBlogs.

Also, Fig. 3 and Fig. 4 depict some useful statistics about the number of bloggers that exist in irBlogs from each blog server:

The above charts are consistent with the statistics of the most visited websites of Iran (Alexa 2015). Blogfa, MihanBlog and PersianBlog are top 3 most visited weblog hosts according to Alexa; also, most of the weblogs in irBlogs are gathered from these hosts (see Table 3).

4. The characteristics of Persian bloggers

Sections 3 introduced the creation process and some useful statistics about irBlogs dataset. In this section, we will analyze some important characteristics of Persian bloggers based on the dataset.

4.1. The bloggers network

Figs. 5 and 6 depict the in-degree and out-degree distributions