Statistics of the Persian Blogosphere, Project "Didish"

Arash Kamangir (Abadpour)
kamangir.net, persian.kamangir.net, abadpour.com

The Persian blogosphere is a vibrant environment where hundreds of active blogs offer an alternative source of information. Analysis of this dynamic maze of blogs and social networks can offer invaluable insight into the mindset of the Iranian youth and how they see the world. In a society where formal communication mediums, such as newspapers and TV stations, are state–controlled, the blogosphere has become the vehicle for discussion and interaction between the people. Project "Didish" aims at discovering the trends and the hot spots in the Persian blogosphere through analyzing links shared by Persian bloggers, among the many tools it makes use of. This short report discusses the utilized methodology as well as some insight into the implementation of Didish accompanied with some sample results. The report also discusses some of the extensions of the project which analyze social networks as well as feeds.

Contents

About the author: Arash Kamangir is the pen name of Arash Abadpour, a graduate student of Electrical and Computer Engineering in University of Manitoba. Kamangir blogs at kamangir.net and persian.kamangir.net. in English and in Persian, respectively. His blogs are amongst the most visited in the Persian blogosphere (with a page view of 4000 per day as of October 2008). He does freelance research on the Persian blogosphere. For more information about his work visit the resume provided at `http://abadpour.com`.

Arash Abadpour's work on the Persian blogosphere is done in his free time and is not affiliated with where he studies as a full–time Ph.D. candidate.

1. INTRODUCTION

In the early days, the Internet was populated by static content, namely conventional HTML pages. These files carried text and images and were uploaded manually to a server. With the rapid growth of the online world, dynamic content conquered the web. As opposed to static web pages, a dynamic website constitutes of a piece of code which generates the pages, normally on the go.

Content Management Systems (CMS) are web–publishing tools which are widely used by bloggers. A CMS system, such as Wordpress or Blogger, is a software which enables a person with minor technical background to publish their content online. Many such tools are installed on commercial servers and the service is offered either for free or in exchange for money. Examples of these blog hosts are `blogspot.com` and `wordpress.com`. Some CMSs are open source and can be obtained for free and subsequently installed on other servers. For example, `wordpress.org` provides the famous Wordpress CMS for free.

A major contribution of many CMS–based implementations is that content and layout are separate, meaning the blog content is not contained in the template used for a particular presentation. Therefore, the same content is accessible in different forms, with the raw content–only form often called a "feed". In fact, any decent blogging CMS provides a feed, through subscribing to which, one is able to follow a blog without logging into the address in which it is presented in a template. As an analogy, the difference between browsing a blog and checking its feed is similar to watching CNN or receiving email alerts which contain the whole content from `cnn.com`.

Feed addresses have other applications as well. For example, it is not practical for a person to survey a thousand blogs everyday, but through using "feed aggregator" utilities one is able to check a large collection of feeds for new content in a matter of a few minutes. Feed has found applications in other services, including link–sharing, as well.

Google Reader, a widely–used feed aggregator lets users "share" links they find interesting from the blogs they are subscribed to. Other websites, such as `delicious.com`, let users "tag" content they would like to share with others. All these services provide the user with a feed through which one can inform others of their "shared links".

Project "Didish" follows a long list of "feeds to shared links" and then analyzes the results. Here, we go through the methodology used by Didish and present some of its extensions.

2. METHODOLOGY

In Persian, "Didish?" means "Have you seen it?". Project Didish constitutes of several building blocks, a brief description of their structure and their interface with each other is given here.

As of November 8th, 2008, Didish follows 1,048 active feeds to links shared by Persian bloggers. Each one of these feeds will be called a "source", for convenience. These sources include Delicious feeds, Google Reader feeds, and others (see Figure 1 for the shares). The process which handles these sources and generates the reports is schematically presented in Figure 2. Here, we briefly review this flowchart. To

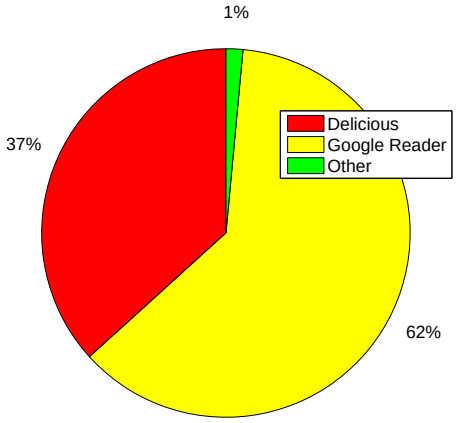

Fig. 1. Sources monitored in project Didish.

have an idea about the volume of information collected by project Didish, note that since January 2008, 272,648 items have been aggregated from the sources Didish monitors.

The first building block of project Didish is a feed aggregator. This aggregator monitors all sources for new content at least once every day. The links, accompanied by the date of submission, the name of the source which has shared the link, title of the item, and other related pieces of information, are stored in a database. Once every week, the utility DidishExtract is launched on a personal computer. This tool, which is a console–based Delphi application, goes through the database and collects the links for further processing. DidishExtract also carries out pre–processing tasks such as conversion from Feedburner url's to real addresses, etc.

The information collected by DidishExtract is saved on a local machine and then analyzed for generating several reports. The first report lists the one hundred web sites content from which has been shared most frequently by the sources. For convenience, the main address of any blog or web site is called a "domain". A similar report presents the one hundred servers where the linked domains are located on. This way, one is able to determine the share of different blogging platforms, such as `blogspot.com`, `wordpress.com`, `blogfa.com`, etc, in the traffic in the Persian blogosphere. These two reports constitute many hyperlinks, connecting sources and domain back and forth, making one able to look at the report from different perspectives. These reports can be found at `http://didish.kamangir.net/report` and are updated once every week.

Similar to the two lists mentioned in the above, the ten topmost domains and the ten topmost servers are also visualized as pie charts. A Sample report will be discussed thoroughly in Section 3.1.

Through collecting the information at different points of time, it is possible to analyze the trends in the blogosphere. These trends determine how the share of one particular domain changes over time and how much stable the rankings are. Section 3.2 discusses the trend curves. The database can also be used for keyword–tracking (Section 3.3) and custom analysis of a particular domain (Section 3.4).

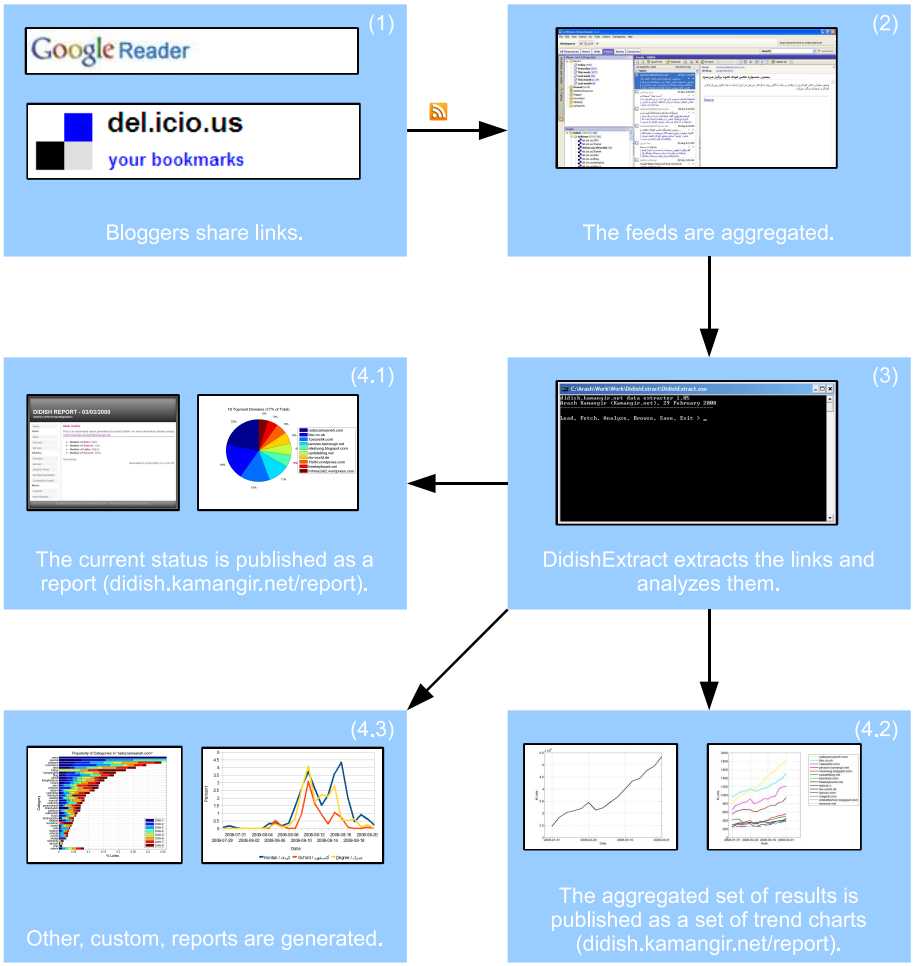

Fig. 2. Flowchart of Didish.

Different blocks of Didish take advantage of pieces of code, either available for free on the web or developed by the author. Omea Reader and Octave are examples of freeware applications utilized in the project. Other pieces of code functional inside the project have been developed by the author taking advantage of free environments of Turbo Delphi, PHP, etc.

3. RESULTS

This section discusses some of the scenarios through which the data collected by Didish can be put into context. These are only examples of the many other possibilities. Here, first the regular weekly reports (Section 3.1) and trend graphs (Section 3.2) will be discussed. Then, the more customized keyword–based analysis (Section 3.3) and custom domain–oriented analysis (Section 3.4) will be presented.

3.1 Weekly Reports

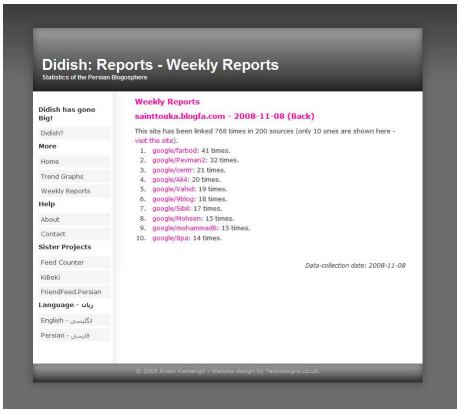

Fig. 3. Sample page in the web–based reports published by project Didish.

The most straight–forward type of report generated by project Didish is the weekly portray of the Persian blogosphere. These reports indicate the websites/blogs, content from which has been shared more frequently in the sources. Domain analysis, source analysis, and graphical representations are included in these reports as well. Figure 3 shows a sample page in such a report. Here, we go through a sample report and mention some highlights.

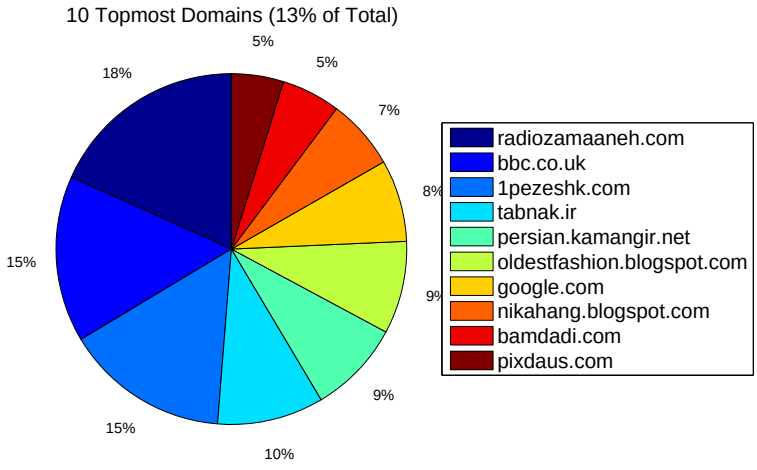

Fig. 4. Ten topmost domains, from a sample weekly report by Didish.

As mentioned before, a current–status report includes a pie chart which shows the shares of the ten topmost domains in the total number of links. Figure 4 shows the corresponding pie chart, for the report analyzed here. This chart denotes that 13%

of all links were from ten domains. This result is based on the analysis of 272,648 links, collected from 1,048 sources and located on a total of 20,862 domains. This result indicates that from the ten topmost domains, only four are formal sources (Radio Zamaneh, Persian BBC, Tabnak, and Google News). The six others are five personal blogs written by Iranian individuals (1Pezeshk, Persian Kamangir, Oldest Fashion, Nikahang Kowsar, and Bamdadi) as well as the image–sharing website Pixdaus. The significant rule of Persian blogs, compared to formal news sources, is clear in this figure.

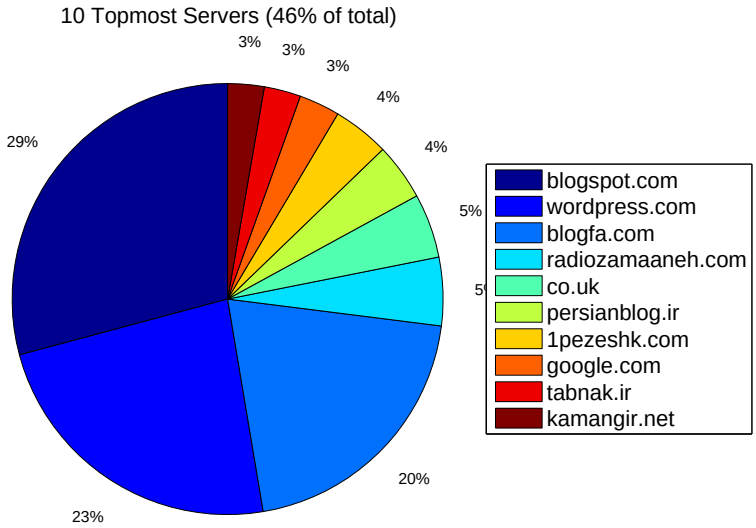

Fig. 5. Ten topmost servers, from a sample weekly report by Didish.

The set of ten topmost servers and their shares of links in a sample report is presented in Figure 5. As seen here, the blogging service `blogspot.com` is the first. The next servers are `wordpress.com` and `blogfa.com`. It is worth to mention that the two Persian blogs 1Pezeshk and Persian Kamangir have been able to produce traffic comparable to the whole Persian blogging service Persian Blog.

Not only it is important to know how many times a link has been shared from a domain, the number of sources which have shared at list one link from each particular domain can be considered as another measure of "popularity". Figure 6 shows the ten topmost domains based on this measure. This list includes the seven Persian blogs of 1Pezeshk, Persian Kamangir, Bamdadi, Nikahang Kowsar, Oldest Fashion, Free Keyboard, and M. H. Mazidi and the three formal sources of Radio Zamaaneh, Persian BBC, and Tabnak.

As many of the sources are operated by a blogger whom has a domain assigned to him/her, it is possible to draw a connection graph which exhibits how bloggers are connected in terms of sharing links from each other. To do so, from the one hundred topmost domains, the ones which do share links are picked. The result is then used for drawing a directed graph, where each node is a domain and the

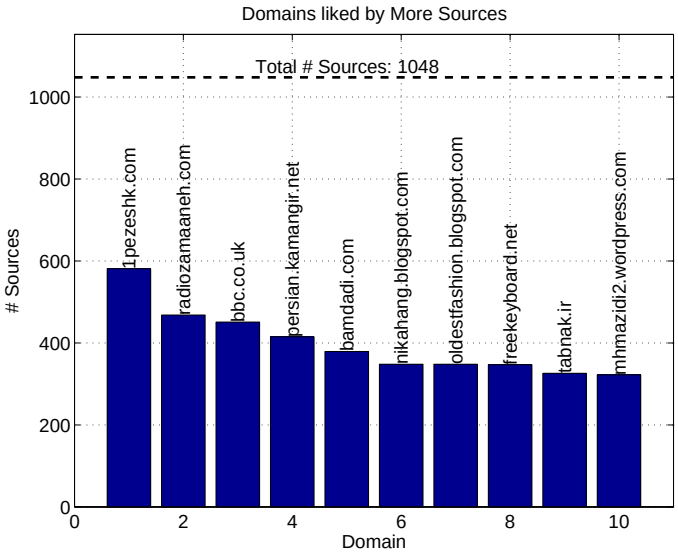

Fig. 6. Ten topmost domains, in terms of number of sources linking to them, from a sample weekly report by Didish.

Fig. 7. Connection graph of the Persian blogosphere, from a sample weekly report by Didish.

intensity of each edge determines the relative number of links the head of the edge shares from the end of the edge. Figure 7 shows this graph for a sample report.

3.2 Trend Graphs

Trend graphs constitute another group of reports generated on a regular basis by Didish. Based on the data collected since the start of the project, Figure 8–(a)

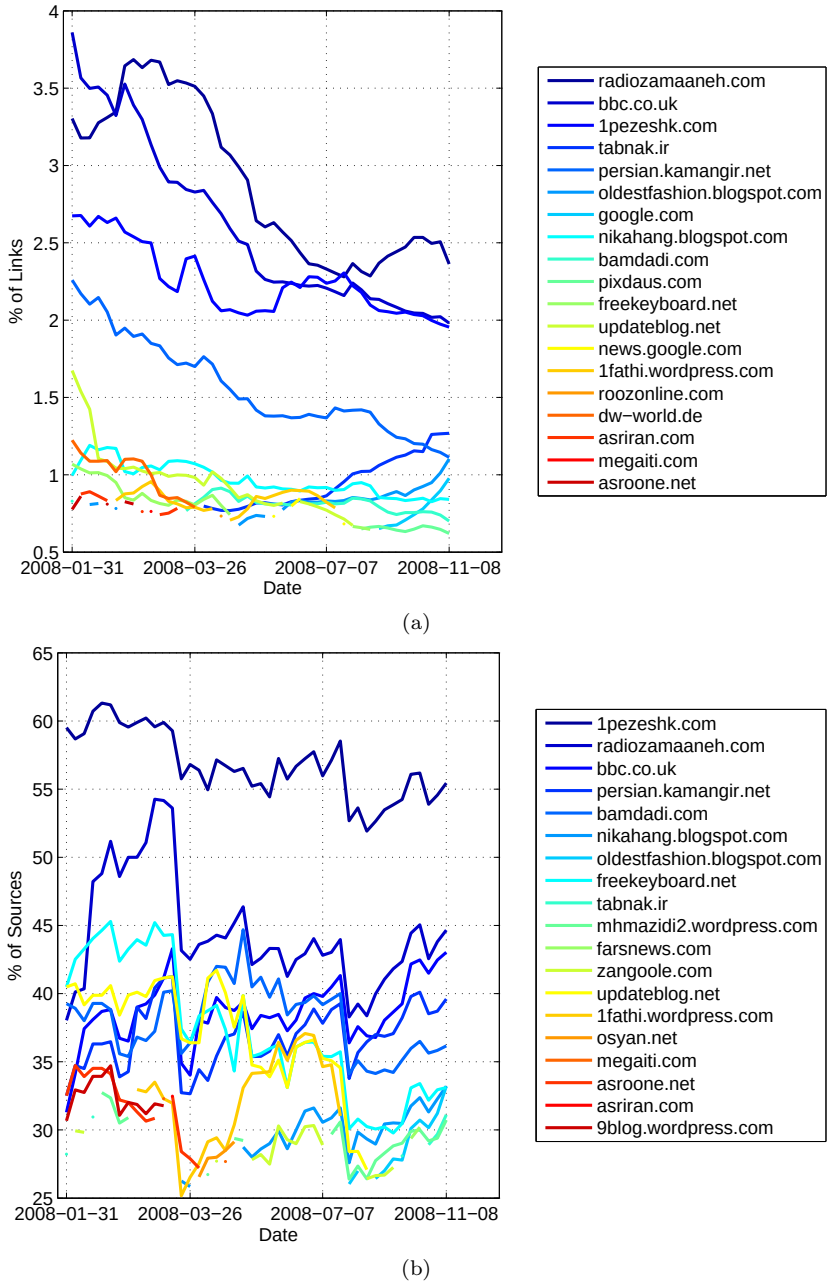

Fig. 8. Curves from a sample trend report. (a) Share of topmost domains in the total number of links over time. (b) Percentage of sources which share links from the topmost domains over time.

presents shares of the topmost domains in the total number of links over time in a sample report. This curve shows that there are three major sources in the Persian

blogosphere, from which links are shared, namely Radio Zamaaneh, Persian BBC, and the personal blog 1Pezeshk.

Figure 8–(b) shows the percentage of sources which share links from each one of the topmost domains in different weekly reports. While 1Pezeshk is steadily linked at in over half of the sources, other domains such as Radio Zamaneh and BBC Persian show figures around 40%.

3.3 Keyword Search and Topic Analysis

As the Didish database contains the link and title of every item shared through any of its sources, it is possible to use the aggregated data in order to collect information regarding a particular issue of interest. For example, in the piece which was published in Radio Zamaneh on August 2007[1], the reaction of the Persian blogosphere to the controversy surrounding the Ph.D. degree of the Minister of Interior of the time, Ali Kordan, was analyzed. This analysis contained daily traffic of the related items as well as the highlights. The following is an excerpt from that report.

> Figure 1 shows the daily percentage of items shared from July 29th till August 20th of this year which contained one of the keywords ‘‘Degree’’, ‘‘Oxford’’, or ‘‘Kordan’’ in their title. In order to draw this figure, first, for each day in the period, the total number of items which had one of the keywords in their title has been calculated. These numbers are subsequently divided by the total number of items shared in the corresponding day. Figure 2 shows a similar curve for the percentage of items which had any of the keywords in their title.

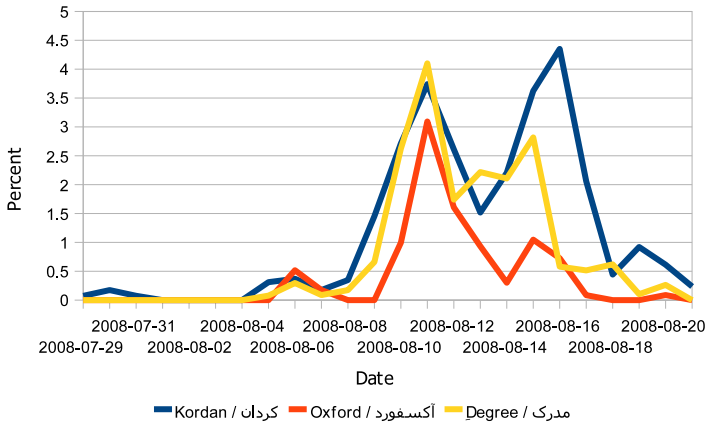

Figure 1.

> All four curves, the ones shown in Figures 1 and 2, exhibit peaks on the 11th and the 16th of August.

[1]Original Persian text: `http://zamaaneh.com/blog/2008/08/post_114.html`, English translation: `http://kamangir.net/?p=4743`, published on November 7, 2008

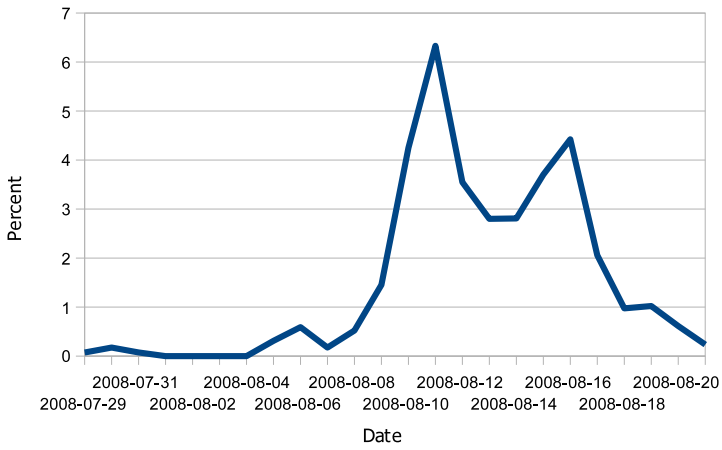

Figure 2.

The peak on the 11th coincides with the day after alef.com first published reports regarding mistakes in what was claimed to be a degree issued by Oxford University. On the same day, Alef also published a statement by Oxford University in which the degree was officially discredited. The Persian blogger Jomhour commented on the news by writing a post titled ‘‘Will the genius minister discredit the Oxford minister?’’ The next peak, namely on the 16th, corresponds to the date in which the Persian blogger ‘‘Big Sleep’’ wrote a post titled ‘‘I will file a complaint against Kordan’’.

Currently, content of the items has to be fetched in manually. In the future, similar automated analysis on the content of the posts will be possible as well.

3.4 Custom Analysis

The database of project Didish can be, and has in fact been, used for tracking the interest of users in any particular domain. The analysis can also be more specific to particular categories or a particular time span, subject to the availability of such information in the url's produced by the regarding CMS.

After a request was made by the director of Radio Zamaneh, an operation was launched in order to provide the governing board of the organization with detailed analysis of their user base. The operation included a custom designed survey as well as reports generated based on the information contained in Didish databases. The report has not been finalized yet and major parts of it are private to Radio Zamaneh. Some results, including the one shown in Figure 9, are based on the information Didish has aggregated through public sources. This figure is the result of the overlay of many graphs which show the popularity of the content published in Radio Zamaneh within different categories during the time span of January to August 2008.

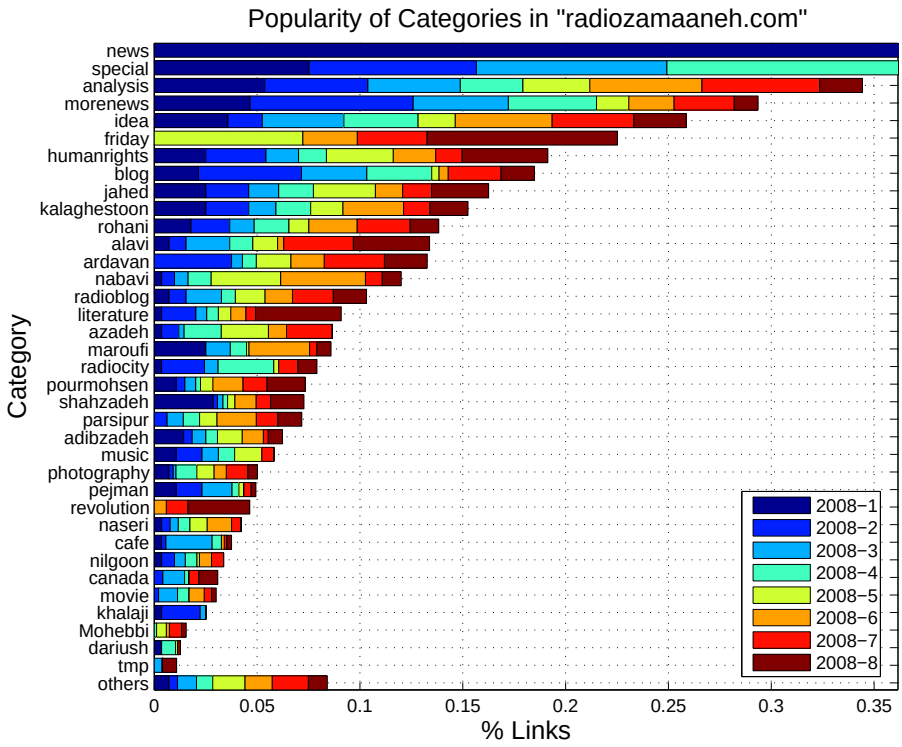

Fig. 9. Analysis of the popularity of the content published in different categories during the specified time span in one particular website.

4. EXTENSIONS

Here, some of the extensions to Didish are discussed. These are blocks which work alongside the flowchart shown in Figure 2. They both acquire information from Didish databases and contribute to them.

4.1 Social Networks

Social networks provide invaluable information voluntarily shared by web users. Among the many running services of this kind, `friendfeed.com` and `delicious.com` have gained the interest of Persian bloggers. Didish already has a wing which automatically tracks the activities of the Persian users of Friend Feed. Plans for a more active analysis of Delicious are under way. The results of these attempts lead to the discovery of more link–sharing feeds as well as a better understanding of the dynamics of the Persian blogosphere. Limited work on other social networks, such as `facebook.com`, has been carried out. More work on these utilities is scheduled for near future.

4.2 Feed Counter

Many Persian bloggers make use of the Google–owned service `feedburner.com` to have better control over the feed of their blog. In the new Web 2.0 paradigm, feeds are a major player, and Feed Burner not only provides the publisher with invaluable feedback about the readership of their content, but also it paves the way for advertisement and other commerce–oriented plans.

As Didish is actively involved in link–sharing which passes through Feed Burner services, a separate wing has been developed at the very early stages of the project in order to track and analyze the readership of feeds in the Persian blogosphere. Titled "Feed Counter", this custom made application takes advantage of the API provided by Feed Burner in order to carry out different tasks, including reports of "the most popular feeds in the Persian blogosohere"[2], tracking of growth of the feeds, and providing "Feed Growth (Feed 2.0)" feeds for the bloggers[3], among other applications.

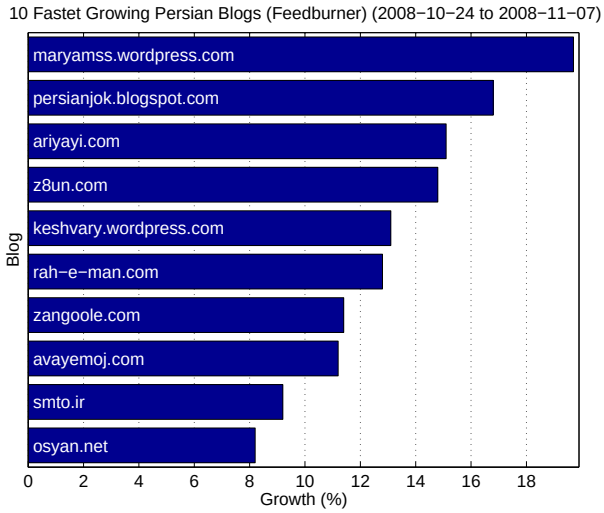

Fig. 10. Sample growth report generated by Feed Counter.

Figure 10 shows a sample "growth report" generated by Feed Counter. The blogs listed here have exhibited the highest growth in the number of their subscribers in the first week of November 2008, compared to the last week of October 2008.

4.3 Profiler

In order to correlate all the information collected through project Didish, a separate utility has been launched. Named "Profiler", an inclusive database of all the individual players in the Persian blogosphere who have been discovered through the analysis of the many social networks under surveillance is being created.

[2]Published almost every week on `http://feedcounter.kamangir.net`.
[3]Accessible at `http://feedcounter.kamangir.net/feed2.php`.

Recently, a limited presentation of the Profiler was given in a talk in Toronto (in September 2008). Current structure of the Profiler enables it to provide comprehensive answers to questions regarding the nodes and the connections in the Persian blogosphere.

For more information about Didish please visit `http://kamangir.net/statistics-of-persian-blogosphere/` or send an email to `arash@kamangir.net` or `arash@abadpour.com`.