Journal of Web Engineering, Vol. 14, No.1&2 (2015) 215-233

AN ONLINE SYSTEM FOR NOTIFICATION OF CHANGES TO BLOGGING SPACE TO ACHIEVE INFORMATION DOMINATION

MEHDI NAGHAVI

School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
naghavi@iust.ac.ir

MOHSEN SHARIFI

School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
msharifi@iust.ac.ir

April 17, 2013
November 17, 2014

Exponential growth of information in the Cyberspace alongside rapid advancements in its related technologies has created a new mode of competition between societies to gain information domination in this critical and invaluable space. It has thus become quite critical to all stakeholders to play a leading and dominant role in the generation of information and monitoring of voluminous information uploaded to this space. Dominance in monitoring of large amount of information in cyberspace requires real-time monitoring using new techniques and approaches instead of traditional techniques. Concerned with the latter case, we limit our focus in this paper on Blogs as an important part of the Cyberspace and propose a novel notification system for quick reporting of changes made to Blogs. This is achieved by restricting the search for changes to high volumes of Blogs only to changes to the abstracts of Blogs derived from Blogs. We show that this system works favourably compared to systems that require cooperation and synchronization between information providers.

Keywords: Abstract Acquisition, Online Web Notifier, Social Networks, Cyberspace, Blog
Communicated by: G.-J. Hoban & J. Freire

1 Introduction

According to a Web server survey carried out by Netcraft, there are almost 240 million registered Web domains [1], about 130 million of which are active [2]. Reports show that Google search engine has access to one trillion unique addresses [3]. Although there is no exact formal statistics on the number of existing pages on the Internet, but the number of pages indexed by the Google searcher, in January 2008, has risen to an estimated 30 billion pages and Yahoo's at 37 billion [4]. Investigations also show that the updating process of the search engines' information is declining. This is to say that search engines have been slow in processing this vast amount of information growth in the Web in spite of continuously getting strengthened in utilities, supports and computational resources and powers. For example, Google has updated nearly 83% of Web pages in year 2005, and only managed to update only 24% of Web pages by the end of year 2007 [5].

In this paper, we offer an online method, which gets aware of the changes made to the Blogs of a country or a specific language, with no need for exhaustive processing resources and bandwidth. The other researches like tracking and analysing the blogosphere [6], event detection and tracking in Social streams [7] and monitoring the changes created in Blogs [8] have been done in this field which have been offline or have used technologies like Ping [9] that many Web providers do not support. What we have concentrated on in this paper is on time and online acquisition of this information regarding the limitations of resources like processing power and bandwidth.

The rest of paper is organized as follows. In Section 2, the Cyberspace and its two main elements including social nets and Blogs are elaborated. In Section 3, notable related works are introduced. The challenges of blog investigation are discussed in Section 4. The goals of blog investigation and on time awareness are discussed in Section 5. In Section 6, the online system of Abstract Acquisition of Changes to Blogging (ACBOS) is explained. Section 7 presents our proposed ACBOS architecture and reports our evaluation of ACBOS based on real experiments. Section 8 finalizes the paper.

2 Cyber Space

Cyberspace is a combination of the words Space and Cyber, which has been extracted from the word Cybernetic. Cybernetic is a theory that determines the relationship between the man and the machine, and the machine and machine. This theory has been put forward by Norbert Winner in 1948 [10]. The word "Cyberspace" has been mentioned for the first time in a science fiction novel in 1984 [11] and now it is one of the expressions used very often in the Web Space. This term has been used as an equivalent of the term virtual space and encompasses many combinations like Cybercitizen, Cyber money, Cyber culture, Cyber trade, and different phrases of this kind. The importance of Cyberspace or Virtual Space increases day by day, so much so that the United States' defence ministry has established the US Cyber headquarter and by issuing an approach document about "Operation in Cyber Space", it has announced that it will do military operations in this space [12]. Pentagon has announced that Cyberspace is a war field like land air and sea, and has introduced it as a new war field.

Considering the Cyberspace equal to other war fields shows the importance of the Cyberspace field to United States. In such a space, recognizing this field more and more has been very important and can help to wisely convert threats to opportunities. Social media have vivid characteristics of the Cyberspace. For this reason, one of the best and the cheapest fields for informing and guiding common sense is social media. Based on the definitions and characteristics of these media [13, 14, 15, 16], we can categorize Web social media as in Table 1.

Social networks and blogs are the most important characteristics of Cyberspace. The widespread use of these networks indicates the tendency of people towards them. According to this article, more than 118 million blogs had been recorded in Iran since 2012; excluding social networks like Facebook, Twitter and LinkedIn. Many scientists had also become interested to use the rich information in social networks using applications in the fields of topic and event extraction, such as the extraction of opinions from Blogging [13.17], automatic summarization of sporting events [18], automatic summarization Blogging [19], and in particular cases such as extraction of hot trends and events from Blogging, micro-blogs and social networks [13,20, 21,22, 23,24,25,26,27,28] that is related to the work presented in this paper. Social networks and blogs will be elaborated further later.

Table 1- Categorization of Web social media

Media	Description	Examples
Blogs	Users' open space for comments and online information	Blogger Wordpress Blogfa PersianBlog
Social Networks	Common space to communicate with each other by posting information, comments, messages, images, etc. (online journals)	MySpace Facebook LinkedIn
Wikis	Users' common space to store, edit and view the content	Wikipedia Wikia Wikinews
Forums	Special space for discussions, which is one of the oldest form of online social media	StonedLizard DiscussionLounge
Microblogs	Combination of social networking and blogging with all instances and short messages	Twitter Pownce Jaiku

2.1 Online Social Networks

Online social networking is one of the fastest and growing phenomena in the Web space, which has created the potential for interaction between users. These networks offer different services, like interactive friendly relations, one way relations, online awareness of friends' locations, expressing opinions, private and public messages, multimedia albums, events and written, audio and video chats of users.

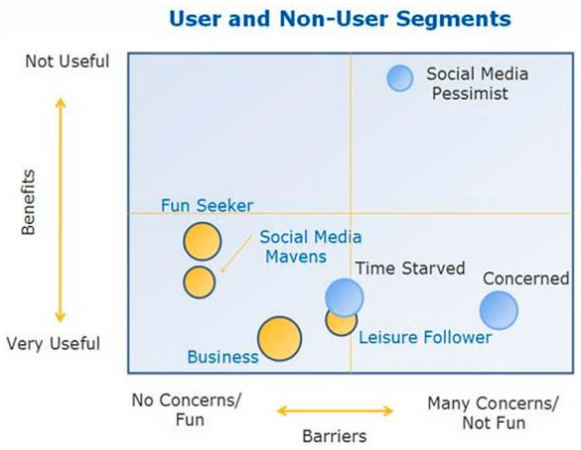


Figure 1. Categories of social networks users

The users of social networks may relate to these networks for different purposes. Anderson Analytic site has issued a categorizing model for users which can help to recognize different types of users of these networks [29]. Although Anderson study has been concentrated on the American users,

considering many application similarities, we can generalize it to other users too. In his research, users have been divided to seven groups in two categories of three social networks non-user segments and four social networks user segments. Four groups related to the members have been named maven users, leisure users, fun seekers users, and business users, and three social networks non-user segments under the name of time starved users, concerned and pessimist users. Figure 1 shows this classification [29].

2.2 Blogs

Blogs as a newly introduced phenomena to Internet have radically changed the Web space social relationships. The most important characteristics that distinguish blogs from other sites in Internet are their epidemic and instantaneous updating. Given these characteristics, investigation of blogs can yield valuable information.

Blog is a kind of site, which is often written by one person, containing notes, comments, informal news related to certain subjects, events explanations, or audio and videos subjects and their links. Majority of blogs let their visitors transfer their opinions and messages to each other using blog tools. These features have increased the dynamism of such sites differentiating them from traditional static sites [30]. Blogs were put forward and introduced in 1997 [31]. By the year 2009, more than 133 million blogs were indexed by the Technorati search engine, which is the search engine for blogs [32], and more than 900 new blog posts were registered every 24 hours [33]. Since 1997, these statistics have shown a very fast and unpredictable growth of blogs. Creating one blog in every second in 2006 [34] and increasing to 11 blogs in every second in 2010, shows the fast growth of blogs.

3 Related Work

Agarwal et al [6] have presented an analytical tool to help sociologists to track and analyse blogs in 2009. Their tool fetches and indexes blogs preparing them for statistical analysis. The analyser then counts the number of occurrences of each keyword in a given time period in the blogs and divides bloggers into *Active-Influential*, *Inactive-Influential*, *Active-Non Influential*, and *Inactive-Non Influential* bloggers. Using this tool, users can be informed on the occurrence of their interested words in new posts via email. The advantage of this system is that both acts perform data collection and data analysis. The weakness is that it does not work online and fetches a complete page instead of just the changes made to the page.

Discovering and distinguishing events in social networks is another task that has been studied by Sayyadi et al [7] in Microsoft in 2009. A graph of keywords is created for documentation and a cluster of keywords is allocated to each event. A keyword network is then constructed based on the cooperation of nodes for creating a common document. In this graph, low frequency words are filtered and the rest of words make up the nodes. Edges are formed based on the cooperation of nodes in creating a document. If keywords of nodes n_i and n_j are cooperating in creating a document, an edge $e_{i,j}$ is created. The edge will be omitted if the frequency of the co-occurrence of the keywords associated with its nodes becomes lower than a minimum threshold. Also, if the possibility of seeing the word k_i in the document, provided that the word k_j existed in the document, is less than the defined threshold,

the edge will be omitted. They have carried out their experiments on 18000 blog posts in a 2-month period.

Oh and her colleagues [17] have presented a system for categorizing political opinions in blog posts and analysing the results. The system is based on a classifying training model using a supervised learning algorithm. It uses the last blog posts that have been classified as liberal and conservative opinions as input data. By providing an initial model for finding and classifying political blogs, the classified results are shown to operators for analysis and further training of the system. They have used prefabricated addresses for recognizing blogs and fetching and classifying of blogs' information. In implementing the above mentioned tasks, the concepts of data mining, clustering and graph have been used.

Gill et al [35] in 2009 have investigated the personality and motivation of bloggers and blog topics. By documenting and referring to the Trait theory, they have dissected personality into a number of measureable factors or adjectives and have investigated the personality based on a five factor model. They show how discussed subjects and blog discussions are arranged. Their work helps to understand the blog subjects, writers and their interests better.

Pathak and Thakre [8] have created an intelligent method for monitoring the changes made to Web sites. They have considered the interests of users and classified the amounts of changes. They have tested their proposed method by a sample system called "WebMon". To consider user interests, they have given the users the right to weight keywords and then used the "Vector Space Model" (VSM) algorithm for estimating the amount of changes made. The algorithm takes in the site addresses given by the user and shows a list of changes made as a decimal digit between 0 and 1.

4 Blog Investigation Challenges

As noted earlier, by the year 2009, more than 33 million blogs have been identified by Technorati search engine. Considering the huge volume of blogs, studying and investigating them is a challenging task. Some of the critical challenges are elaborated below.

4.1. Notification Urgency

Temporally, blogs are always in a temporary state. Bloggers frequently post new subjects and readers frequently comment on new posts. Users expect to be informed about new posts as soon as the posts are uploaded to keep pace with fresh information posted. Ping technology, first introduced in 2001 [36], is a sample remote procedure call technique based on XML that blogs can use to inform a server machine, called the Ping Server, about new posts and comments on the posts [9]. How quickly this notification can be done remains a challenge though.

4.2. Impermanency and Freshness

Information posted to or commented in Blogs are mostly valid and of interest in a given often short time period, after which become stalled and lose their informative value. It is thus critical to be able to investigate Blogs quickly whilst posts and comments are still fresh and worthwhile.

4.3. Categorization

Document categorization is one of the foundational problems in Web information retrieval [37]. Categorization of blogs is a difficult and time-consuming task for both human and computers. Human investigation is hindered both by often the long time it takes to investigate blogs manually as well as by the need for a large number of skilled manpower. Mechanized investigation is also challenging because information in blogs do not have a uniform structure but instead their structures depend on the favoured styles of individual users of blogs that mostly write irregularly, discursively and illogically. Some have used the linguistic characteristics of the post names in blogs and the anchor texts to categorize blogs. However, tests have shown that categorization of blog text under one category has not been successful [38]. In addition, it has been difficult to put a blog in just one category as most blogs could well fit into more than one category making it difficult to choose between different categories for a given blog. Having said that, faceted classification [38] based on Ranganathan theory [39] has shown to be more suited to categorization of blogs.

4.4. Spam Pages and Links

Obtrusive pages named Spams are one of the obstacles to investigation of blogs. Based on Technorati's report, about 3000 to 7000 new spam blogs are created each day and this is growing at the rate of 11000 pages per day [40]. .In the best case, Web spam pages are a nuisance that provide undeserved advertisement revenues to page owners. In the worst case, these pages pose a threat to Internet users by hosting malicious content and launching drive-by attacks against unsuspecting victims [41]. Spam links are false connections with no meaning. Spam links are made dynamically from comments and answers. Spam links made from comments are capable to be created easily in the blogs. A spammer writes a simple agent that randomly visits blogs and posts comments that link back to the spammer's page [42].

4.5. Colloquial Writings

Using spoken and informal or conversational language in blogs creates a great variety in words and their meanings. Most blog posts and comments contain technical and foreign words with foreign spellings including lots of spelling and typing mistakes. These make it difficult to understand the precise meanings of the words and to investigate the blogs purposefully. It also makes it more difficult to find equivalents for words and reduces the role of the machine and increases the role of human in producing a glossary.

4.6. Blank Blogs

There are many blogs that have been created but left inactive. Investigating millions of blank pages wastes valuable resources like the bandwidth and processor time. A preventive technique is to develop a blank page bank and try to recognize them before processing them any further. However, given that most blog pages are blank at the time of their formation and that these pages get filled gradually, it is reasonable to be quite careful in labelling pages as blank pages.

4.7. Discrete Documents

Documents in social nets and blogs are always discrete while normal Web pages are linked to each other through subpages [43]. This lack of linkage in blogs is exacerbated when servers omit the links because the linked pages are less used. The lack of links prohibits the deployment of a continuous process for crawling and fetching of pages.

5 Blog Investigation Contribution

Based on what was mentioned in previous sections, blogs are valuable sources whose purposeful investigation can provide priceless information. Some useful contributions from blog investigations are as follows.

5.1. Community Interests

Internet users have grown in number as Internet technologies advance. By June 2010, the estimated number of Internet users was about two billions (1,966,514,816) [44], most of which were bloggers and blog readers. Finding the interests of blog communities provides great opportunities such as active participation in trades with the awareness of users' needs, purposefully issuing political viewpoints in line with users' tendencies, and getting informed about specific crimes on Internet.

5.2. Daily Trends

Today's communities are quickly vulnerable to different events propagated by new advanced media technologies including Web. Online investigation of blogs can keep community members in their interested subjects to get informed and aware of daily relevant significant information published on the Web.

5.3. Sensitivities

Sensing the pulse and common craze of a community from the contents of blogs is another application of blog investigation. Sociologists and politicians are particularly interested to be aware of sensitivities of their societies. With daily increases in the number and variety of Web users, Web space has become an inherent part of human societies. Getting informed about the sensitivities of this space can well provide valuable information on the heartbeat of human societies in the large.

5.4. Consequences

Taking Web users' community as an important part of human societies, their views are representative of the common sense of the human societies in large. It is critical to know the consequences of events on societies if proper contingency plans are to be sought beforehand or after events have occurred. Some example events, that strongly affect the societies and decision makers' awareness of the consequences of their occurrences and is deemed most critical, include natural disasters such as earthquakes, floods and hurricanes, unnatural events such as wars, social events such as elections, putting in action cross country regulations such as increasing or decreasing annual salaries or taxes, which affect everyone's life in societies. Knowing the consequences of such events greatly helps to

find proper solutions to hamper probable crisis. Web investigation can provide a great contribution to become aware of the common sense of societies in such cases.

6 Abstract Acquisition of Changes to Blogging (ACBOS)

Getting aware of Web changes fast and quickly would near us to the goals we mentioned in the previous sections. The Web writing scenery is a symbol of events that occur in the society. Getting informed about the changes made in the blog writings will guide us to new trends in the society. Our proposed blog changes online notifier system has been designed with this concern. To get online notifications on updates to all Blogs is a daunting task given time and computing resource constraints. We have thus opted to narrow our investigation of changes only to a certain country (Iran) and a certain language (Persian), as well as to a limited number of mostly favoured blogs in that country. We further discuss this design issue in this section before presenting the designed structure of the notification system.

6.1 Concentrating on a Specific Domain of Blogs

As mentioned in Section 1, based on Kunder estimation [4], the number of blog pages that have been indexed by one of the search engines by the year 2008, were 30 billion pages. Investigating this volume of information requires very expensive equipment and a high expense. If, in the whole Web space, we just concentrate on the blogs only, the practice domain would be limited from a 34 billon page domain to a 133 million page domain [33]. It is necessary to reduce the target domain into a few million blogs, by choosing a specific domain of the blogs, in order to be able to achieve our purpose, using accepted equipment and suitable architecture.

6.2. Blog Investigation Limitations

The rate of changes to blogs, the number of users of the blogs, deadlines for getting the results, and available processing power and network bandwidth are the limiting factors that can determine a reasonable architecture for our online notifier system. Investigating the target space in a specific and acceptable time limit is a necessity for blog online investigation. If the wanted result is ready after the definite time, the achieved information loses its value and applications and would be useless. Also because of limitations in supplying the abandon hardware, the processing capacity would be limited. The offered architecture should be able to consider the hardware and equipment limitations and should be offered in a way that we can install it, having the mentioned limitations. The bandwidth is also another case that is considered a limitation, because of imposing high expenses. The offered architecture should be in a way that it allows the system to work with low network bandwidth too and does not always require high bandwidth.

6.3. Limiting Blog Domain to Persian

Based on the report by the Technorati site, The Persian Language was among the first 10 languages of the blogs in the entire world [45]. And in 2008 the number of blogs in Iran has been estimated to be around 2 million [46].This statistic shows this important matter that blog writing has a lot of fans in Iran and we can, by concentrating on it, extract valuable information out of it. To investigate Persian

blogs, 81 cases of Persian blog providers have been identified and investigated. Identifying these providers by implementing a Map Reduce program, on the crawling results of 32 million Persian pages, which have been done on HBase distributed data base in the National Search Engine Laboratory, 21300 blog addresses were extracted. By refining these addresses, the main blog service providers were extracted, based on the final opinions of the adept human labour.

To investigate these providers and prepare the wanted statistics, the services offered by Alexa (*http://www.alexa.com*) have been used [47]. Also, in addition to using Alexa services, in order to extract the other needed statistics, the "site:" operator of the Google search engine has been used and finally a chart was prepared that showed the results of six factors for investigation of each different 81 Persian blog providers. These factors related to the *global rank*, *rank in Iran*, *reputation*, *reach*, *page views*/*user*, and *the number of existing pages in each service provider*. The world rank shows the rank of the blog in the whole sites in the world. The rank in Iran shows the traffic rank of the blog relative to whole sites in Iran. The reputation factor shows the number of other sites' references to the blogs. Reach factor shows the percentage number of blog users relative to the whole users in the world, and the page view/user shows the average number of visited pages by each user. In order to achieve the sixth factor that shows the approximate number of existing pages in each provider, the 'site:" operator of the Google has been used [48].

In Table 2 only the collected information of the first 10 service providers of the Persian blogs are presented. These 10 service providers are the first providers considering all factors except the page view/user factor.

Table 3 reports the results from compiling information related to 81 identified service providers, comparing the average number of the first 10 providers with the average numbers of the whole Persian blogs. By investigating these statistics, we can conclude that to collect information from the blogs, investigating the first 10 providers is enough, because from the point of view of the volume, it has covered 86 percent of the number of existing pages, and the other percentages also show that the other providers will not have any significant effects on the achieved results.

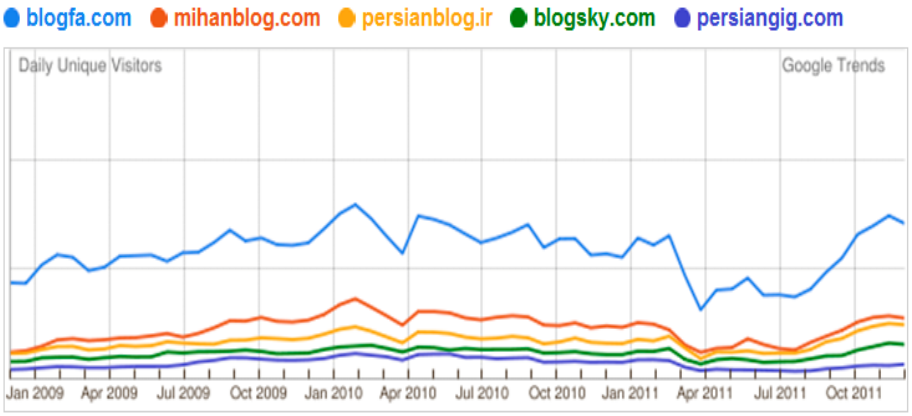
Figure 2 shows the number of visits to the first 5 Persian Blog providers from the beginning of 2009 until the end of 2011. These statistics are calculated using Google Trends [49]. These results are consistent with previous statistics that is provided in Table 2 about page views per user.

Table 2- The first 10 service providers of the Persian blogs

Blog Service Providers	Global Rank	Rank in Iran	Reputation	Reach	Page Views /User	No. of Pages (Million)
Blogfa.com	187	3	178508	0.5298	4.71	50
Mihanblog.com	427	5	47354	0.2716	3.25	18.7
Persianblog.ir	572	6	57087	0.2214	2.64	9.3
Blogsky.com	1045	13	35196	0.1352	2.56	3.8
Persiangig.com	2473	38	23941	0.0599	2.72	0.715
Parsiblog.com	2525	34	16549	0.0656	1.93	8.29
Iranblog.com	3213	55	9674	0.0481	2.58	1.1
Rozblog.com	4461	71	8307	0.034	3.06	5
Parsfa.com	5582	78	1705	0.0337	1.7	0.817
Loxblog.com	7939	123	5530	0.0207	2.9	4.53

Table 3- The first 10 providers versus the whole Persian blogs

Providers	Global Rank	Rank in Iran	Reputation	Reach	Page Views/User	No. of Pages (Million)
All Persian blogs	1076232	8481	5144	0.02	2.9	118
The first 10 providers	2842	42.6	38385	0.14	2.8	102



011Figure 2. The visiting number of the first 5 providers

6.4. Blog Updating

In order to estimate the required resources (namely, processing power and network bandwidth) for crawling updated pages, it is necessary to determine the number of these pages to predict the time duration of updating the blogs for fetching the pages in reasonable time. In order to do that and using the application of "site:' and Google time duration service, we extracted the statistics of the first 20 providers on 16 January 2012 [50]. The achieved results in 1 hour time duration, 1 day time duration, 1 week time duration, and 1 month time duration have been registered. Table 4 shows the results of this test for the first 10 providers. As it is shown, the total average of updated pages of the first 10 blogs in one day is 474586 pages. If we want to get aware of the changes online, we should be able to fetch this volume of pages daily. That means we must be able to fetch 5.5 pages a second.

In order to achieve the needed bandwidth, it is necessary to know the volume of each page. The curve in Figure 3 shows the result of the examination of investigating the volume of the Persian blogs' pages. In this examination, we only considered the 4 first updated pages of the blogs. We must consider that the total number of pages of the providers is 81.4 million pages that makes up 69 percent of the total number of blogs. The results achieved from the examinations is acceptable. Figure 3 shows that the volumes of the pages are not very varied and that the very high and very low page volumes occur very scarcely. Table 5 shows this examination.

Considering the above results, we can estimate the required bandwidth and storage space. Based on Table 5, the average volume of a Persian blog page is 40461 Bytes. So based on this, for fetching 5.5 pages in a second, we need 1.8 Mbps bandwidth. Based on Tables 4 and Table 5, the needed daily space for storing the blogs is also 19.2 Giga Byte.

Table 4- The number of updated pages

Providers	Pages updated				
	Hourly	Daily	Weekly	Monthly	Average
Blogfa.com	13300	384000	2500000	3420000	293586
Mihanblog.com	76	136000	871000	2250000	84313
Persianblog.ir	307	98000	474000	1120000	52604
Blogsky.com	1	13500	147000	407000	12023
Persiangig.com	0	61	9830	41700	714
Parsiblog.com	30	44	12300	61500	1143
Iranblog.com	0	41	385	17000	166
Rozblog.com	26	41000	296470	562600	25683
Parsfa.com	0	168	13500	114000	1474
Loxblog.com	7	271	34900	183000	2881
Total	13,747	673,085	4,359,385	8,176,800	474586

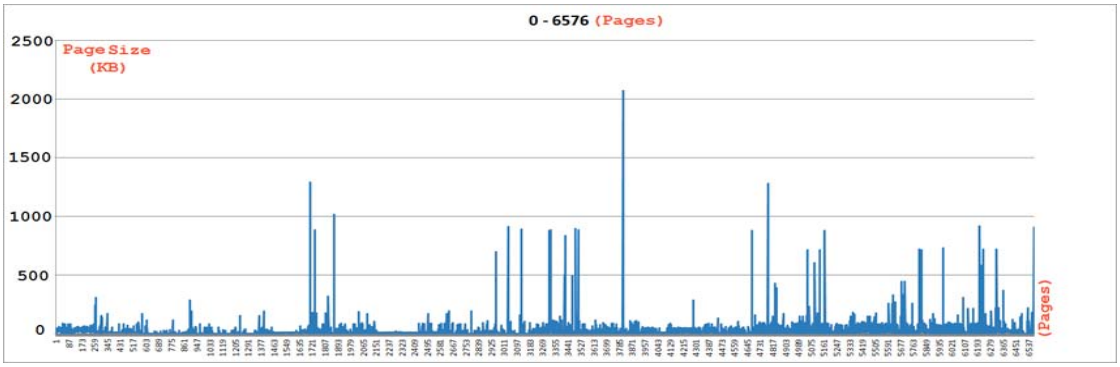


Figure 3. The volume of the Persian blogs' pages

Table 5- Specifications of the fetched blogs

Fetch interval Pages: From 19:19 on 01/26/2012 until 7:57 on 27/01/2012	
The number of fetched pages	6575 Pages
The number of zero-byte pages	3207 Pages
Total size of fetched pages	136273566 Bytes
The average size of each page	40461 Bytes
Maximum page size	2071479 Bytes
Minimum page size	704 Bytes

7 Architecture and Functioning of the Proposed Notification System

Based on features we arrived at in our arguments and experimentations reported in Section 6, let's now present the architecture and functioning of our proposed online notifier system for Abstract Acquisition of Changes to Blogging (ACBOS). Our notification system should be aware of changes to blogs. But, according to the results obtained in this paper, the number of Persian Blogs were estimated to be at the

range of nearly 120 million pages in 2012 and expected to increase yearly. On the other hand, the awareness of changes to blogs without cooperation of their hosts requires a large bandwidth and processing power. We have set the objective to design an architecture for our proposed online notifier system that could work with minimal network bandwidth and hardware resources in a timely online manner, without requiring the cooperation of hosts. This is one of the unique features of our proposed notifier system that distinguishes it from other notifiers such as *SPADE* [51], *StreamWeb* [52] and *SIS* [21].

Figure 4 shows the architecture of our notifier system. RSS Cloud, PuSH and RSS Gather units are responsible for gathering information. The RSS Cloud unit works based on the RSS Cloud technology [53]. This technology is an advanced technology based on RSS Protocol, which provides immediate awareness.

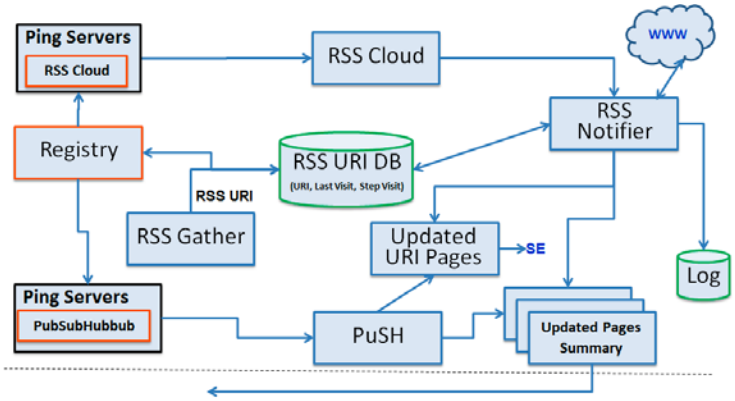


Figure 4. The architecture of ACBOS

The RSS Cloud technology, which is a Web service, is installed based on HTTP_POST, XML_RPC, or SOAP. This technology provides the potential of registering in a cloud for processes, so that it can make them aware of the updates [54]. The unit PuSH, which supports PubSubHubbub technology, is responsible for inquiry from providers that use this technology. The PuSH technology is like Cloud with the difference of sending the changed contents too. This protocol is a decentralized protocol that is completely open. When a new subject is posted, the publisher will inform a Hub and this Hub will send changes to all joints. In this case, customers do not need to ping the Hub to get new updated information [55].

The RSS Cloud and PuSH units need to cooperate with providers. The RSS Gather unit is designed for this purpose in the absence of providers' cooperation and can be notified from their updates. This unit uses the patterns of each host, analyzes their RSS, detects new or modified posts addresses, filters the received data by normalizing and classifying them, and finally sends the filtered data to other units. This information, including abstracts posts and updated pages' addresses are used for other applications such as Web monitoring. By analyzing RSS of providers, their patterns are extracted and placed in different categories. RSSs that are of a same category are fetched with the same pattern.

The main challenge of RSS Cloud and PuSH technologies is that they can be efficient only if blog service providers support them. If blogs do not support these technologies, it is not possible to use

them. Many blogs do not support these technologies. Based on the investigations into the first 10 Persian service providers, all the first 10 Persian blogs did not support this technology. So another solution should be presented for that group of blogs, which do not support this technology. If the polling method is used, based on the results of the tests done, the needed time for fetching the updated pages of the Persian blogs with 1.8 Mbps of bandwidth, is one day. A one day time for online awareness is intolerable and unacceptable.

To confront this problem, in the first look, we can consider increasing the bandwidth as a solution. In that case if we want to reduce the awareness time from one day to 10 minutes, we need about 260Mbps of bandwidth. Providing such a bandwidth only for getting aware of the blog changes is not lucrative and because of financial matters it may be impossible. A suitable solution is to be able to reduce the volume for downloading. RSSs crawling are the solution we have come up with. Considering the low volumes of RSSs, the time of fetching them can be much shorter than fetching the whole information.

To counter the time and bandwidth constraints, we carried out another examination using the same architecture as shown in Figure 4, but with different settings using a two quad-core processor machine with 2.5 GHz frequency, a 48GB RAM, a 12TB HDD, and a 2Mbps network bandwidth, running 16 threads. We fetched the blog changes of the first 4 providers of the blogs, which make up 69 percent of the blogs volume, in 27 days from May 17 to June 13, in 2012. The results of this examination are both shown in Figure 5 and tabulated numerically in Table 6. We wrote a program for fetching the changes and ran one copy on each of the 4 providers. Each run of the program connected to its assigned provider and fetched the changes in information through RSSs and after a minute pause, connected again to that very provider and repeated fetching changes. As stated in Table 6, during 27 days, we visited each provider 27964 times and downloaded their changes. An average of 83 seconds was spent for each fetching. Considering that after each fetching we had a “1” minute pause and reconnection to that very site, we can reduce “60” seconds from this estimated time, and get to the time “23 seconds”. In other words, we can fetch the changes of these blogs (that make up 70% of Persian blogs) in 60 seconds, and we can fetch the changes to all the Persian blogs in 33 seconds. This is an approach for investigating the blogs and monitoring them online.

Table 6- Profile of visits of 4 first providers

	blogfa	blogsky	mihanblog	persianblog	Average
Number of visits	28157	27795	30082	25823	111857
Max. number of updated pages per visit	231	100	247	110	247
Total number of updated pages	814457	5180	76088	77650	973375
Average per visit	29	3	3	3	9

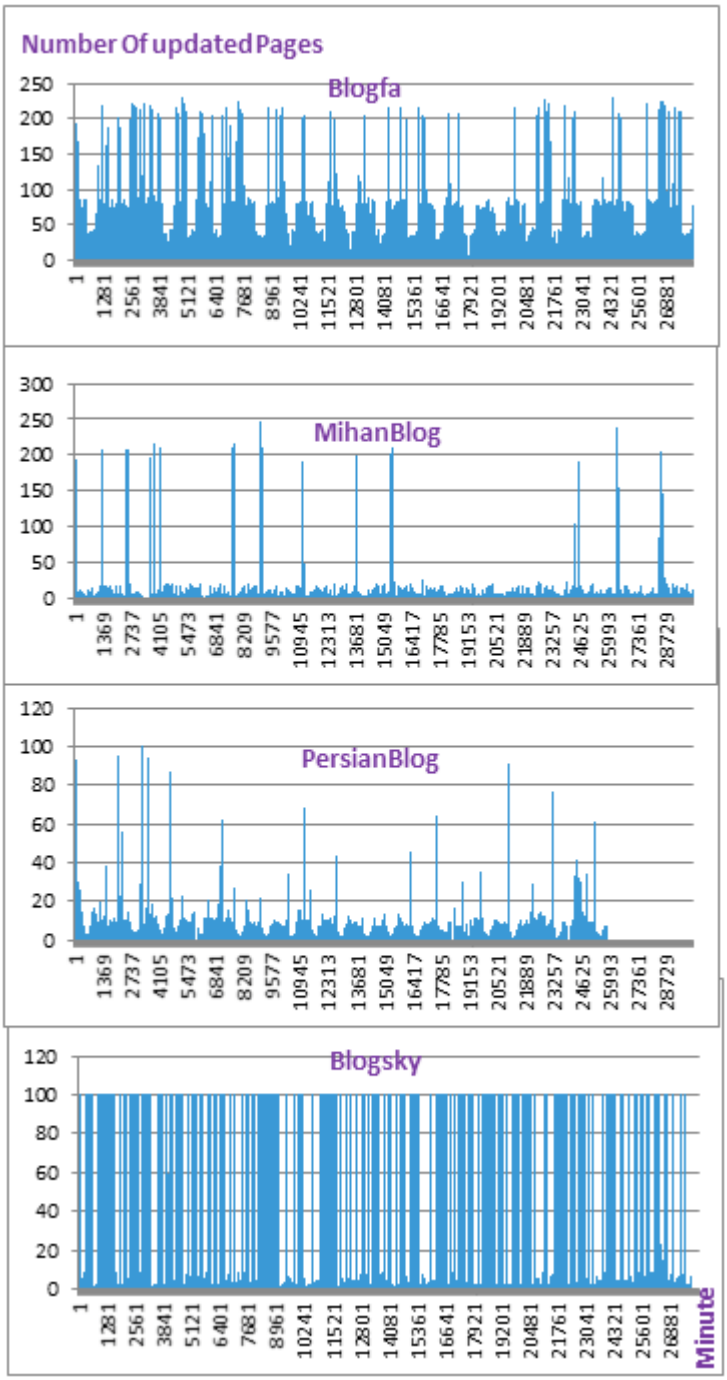


Figure 5. Updates of the first 4 blogs in 27 days (vertical axis denotes the number of updated pages and horizontal axis denotes the time to fetch pages in minutes)

Table 7- Number of fetches (updated pages)

In ***n*** minutes after the creation	Number of pages fetched	Percent of total
1	614	0.6
2	4948	5.6
3	10173	11.5
4	15964	18
5	21761	24.6
10	39279	44.5
15	45102	52
20	48019	54.4
25	50018	56.6
30	51307	58
35	51980	58.9
40	52407	59.3
45	52784	60
50	53159	60.2
55	53435	60.5
60	53694	60.8
360	57331	65
1440	63320	71.7

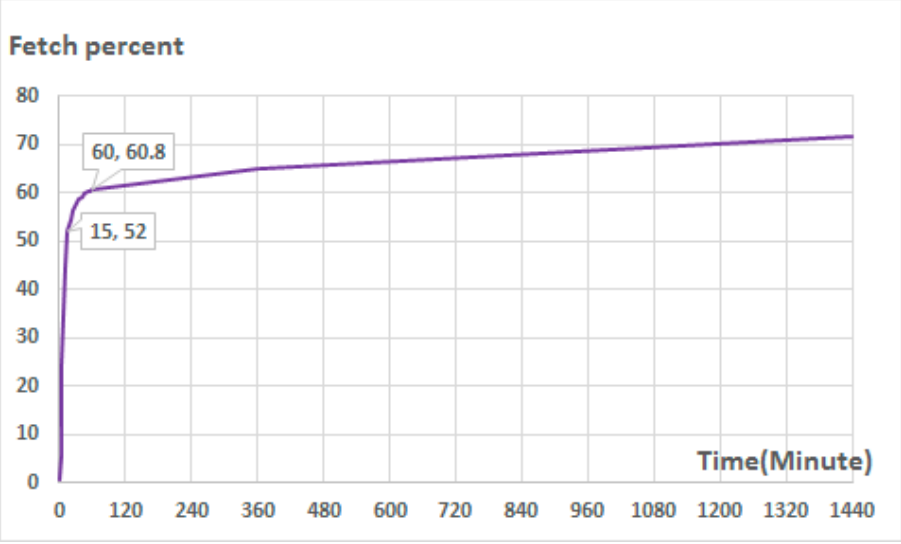


Figure 6 - Fetch-time diagram

Based on the offered architecture, after fetching a blog's RSSs by the RSS Gather unit, the RSS Notifier unit investigates all RSSs and extracts the addresses of the changed blogs and stores them in a chart. Some of the blog providers have got a page in which the latest updated blogs are registered. By processing this page, we can get these blog's addresses. So by periodically investigating this page, we can become aware of the changes in the blogs in a short time. The mentioned time period might be different compared to different providers and set based on the traffic behaviour of each provider. Time duration of the mentioned pages are estimated by relation (1).

$$StepVisit = \begin{cases} \frac{StepVisit}{2}, & UpdateFlag = 1 \\ StepVisit + NextVisitDelay, & UpdateFlag = 0 \end{cases} \quad (1)$$

In the relation (1) *StepVisit*, the investigation time duration is based on seconds and *UpdateFlag* specifies that if a visited page is updated. If the page is updated *UpdateFlag* is set to 1 and otherwise it is set to zero. If during the visit, the page is updated, *StepVisit* is halved and thus next visit is sooner and crawling is faster. If the page has not been updated, the next visit will be *NextVisitDelay* seconds later than the current interval. According to the experiments done, 120 was suitable for the next visit delay.

To illustrate the operation of the system, another experiment was set up. Blogfa that has the most pages and the highest number of users amongst the Persian service providers, was studied in this experiment. Table 7 shows the results of this study from October 8 to October 11, 2014. In this experiment, we studied 88265 pages.

Figure 6 shows these results diagrammatically. As is shown in this figure, 52% of updates were fetched in the first 15 minutes. According to our investigations, there were two main reasons that caused 48% of updates to be received after 15 minutes. Firstly, it was observed that in some cases old data were fetched alongside updated information causing the information to be recorded wrongly. We reckon this symptom is attributed to Laboratory environment conditions and failure in non-stop operation of the system possibly due to the failure of communication systems such as network or DNS failure. Secondly, the time difference between local time and GMT time led record update time on blogs in two different base times causing some information to be recorded wrongly 3.5 hours later.

8 Conclusion

Social networks and blogs have turned into the most useable informal media so much so that most users refer to these kinds of networks rather than the formal networks. Statistics show that the popularities of these networks are very high and increasing in most countries. The very existence of such networks in the Cyberspace can be considered as providing a venue for vast opportunities as well as threats. If we have the readiness to use these opportunities, we can benefit and use them in our advantage. To this end, we presented an architecture for an online notifier system that can collect updates to Persian Blogs as an exemplar domain in world Cyberspace. We experimentally estimated the volume of Persian Blogs for the first time, amounting to 118 million pages, each page 40 KB in size on average. We found 10 top Persian Blogs and showed that they count for 86% of total Persian Blog. We designed our online notifier in such a way that it can work only on the abstracts of updated pages in order to enable monitoring of Blogs with minimum resources online. This was supported by

experimentations that showed 52% of changes to top 10 Persian Blogs were fetched in 15 minutes using limited hardware resources with a low end 1.8 Mbps network bandwidth. We also showed that this percentage can be improved if erroneous update time logging is avoided and old updates are not fetched when fetching new updates to Blogs.

References

1. Greenwood, M. “Prioritising Hyperlinks for Topic-Focused Web Crawling using Lexical and Terminological Profiling.” M.A. thesis, University of Manchester, 2009.
2. “Whole-Product-Dynamic test”, Technical report, AV-Comparatives and the University of Innsbruck's Faculty of Computer Science and Quality Engineering, 2011.
3. Henrique, W. , Ziviani, N., Cristo, M. A., Moura, E. S., Silva, A. S., Carvalho, C. “A New Approach for Verifying URL Uniqueness in Web Crawlers”, in Proceedings of the 18th International Conference on String Processing and Information Retrieval, p.237-248, October 17-21, 2011, Pisa, Italy.
4. Lee, H.T., Leonard, D., Wang, X., and Logulnov, D. “IRLbot: Scaling to 6 Billion Pages and Beyond”, in Proceedings of the 17th International WWW Conference, pp. 427-436, 2008.
5. Lewandowski, D. “A Three-Year Study on the Freshness of Web Search Engine Databases.” in Journal of Information Science, Vol. 34 No.6, pp. 817–831, 2008.
6. Agarwal, N., Kumar, S., Liu,H., Woodward, M. “BlogTrackers: A Tool for Sociologists to Track and Analyze Blogosphere.” in Proceedings of the Third International ICWSM Conference, pp. 359-360, 2009.
7. Sayyadi, H., Hurst, M., Maykov, A. “Event Detection and Tracking in Social Streams.” in Proceedings of International Conference on Blogs and Social Media (ICWSM), 2009.
8. Pathak, M., Thakre, V. “Intelligent Web Monitoring - A Hypertext Mining-Based Approach.” in Journal of the Indian Institute of Science, Vol. 86, No. 5, pp. 481-492, 2006.
9. Nanno, T., Suzuki, Y., Fujiki, T. and Okumura, M. “Automatic collection and monitoring of Japanese Blogs.” in Proceedings of International World Wide Web Conferences (WWW) 2004 Workshop on the Web logging Ecosystem: Aggregation, Analysis and Dynamics.
10. Tabansky, L. “Basic Concepts in Cyber Warfare.” in Military and Strategic Affairs Journal, Vol. 3, No. 1, 2011.
11. Swanstrom, E. “Wax Blocks, Data Banks, and File #0467839: The Archive of Memory in William Gibson’s Science Fiction.” in InterActions: UCLA Journal of Education and Information Studies, 1.2, Paper 7, 2005.
12. The U.S. Department of Defense. “Department of Defense Strategy for Operating in Cyberspace.” 2011.
13. Bakliwal, A., Arora, P., Varma, V. “Entity Centric Opinion Mining from Blogs.” in Proceedings of 24th International Conference on Computational Linguistics, pp. 53-64, 2012.
14. Shekhar, S., Oliver, D. "Computational Modeling of Spatio Temporal Social Networks: A Time-Aggregated Graph Approach.” A position paper for the Workshop on Spatio Temporal Constraints on Social Networks, Santa Barbara, 2010.
15. Hernandez-Ramos, P. "Blogs and Online Discussions as Tools to Promote Reflective Practice." in the Journal of Interactive Online Learning, Vol. 3, No. 1, 2004.
16. Mingjun, X., Hanxiang, W., Weimin, L., Zhihua, N. “A Public Opinion Classification Algorithm Based on Micro-Blog Text Sentiment Intensity: Design and Implementation.” in the Journal of Computer Network and Information Security, 3, pp. 48-54, 2011.

17. Oh, A., Lee, H., Kim, Y. “User Evaluation of a System for Classifying and Displaying Political Viewpoints of Blogs.” in Proceedings of the Third International ICWSM Conference, 282-285, 2009.
18. Nichols, J., Mahmud, J., Drews , C., “Summarizing Sporting Events Using Twitter.” in Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, New York, NY, USA, pp. 189–198, 2012.
19. Mithun, S. “Exploiting Rhetorical Relations in Blog Summarization.” PhD Thesis, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada, 2012.
20. Mathioudakis, M., Koudas, N., “TwitterMonitor: Trend Detection ver the Twitter Stream”, in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, pp. 1155–1158, 2010.
21. Shih, C., Peng, T. “Building Topic/Trend Detection System based on Slow Intelligence.” in Proceedings of the 16th International Conference on Distributed Multimedia Systems, Oak Brook, Illinois, USA. pp. 53-56, 2010.
22. Weng, J., Lee, B.S., “Event Detection in Twitter.” In Proceedings of the 5th International AAAI Conference on Blogs and Social Media, pp. 401–408, 2011.
23. Kim, D., Ki, D., Rho, S., Hwang, E. “Detecting Trend and Bursty Keywords Using Characteristics of Twitter Stream Data.” in International Journal of Smart Home, Vol. 7, No. 1, 2013.
24. Fang, F., Pervin, N., Datta, A., VanderMeer, D. “Detecting Twitter Trends in Real-Time.” in Proceedings of the 21st Workshop on Information Technologies and Systems, 2011.
25. Vakali, A., Giatsoglou, M., Antaris, S. “Social Networking Trends and Dynamics Detection via a Cloud-Based Framework Design”, in Proceedings of the 21st International Conference Companion on World Wide Web , New York, NY, USA, pp. 1213–1220, 2012.
26. Petrovic, S., Osborne, M., Lavrenko, V. “Streaming First Story Detection with Application to Twitter.” in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, PA, USA, pp. 181-189, 2010.
27. Benhardus, J. “Streaming Trend Detection in Twitter.” in National Science Foundation REU for Artificial Intelligence, Natural Language Processing and Information Retrieval, University of Colarado, 2010.
28. Wortmann, P. “Topic-based Blog Paper Search for Trend Detection.” Technical University of Kaiserslautern, Project Thesis, 2009.
29. Anderson Analytics. “Seven Social Network Segments”, Available from: http://www.andersonanalytics.com/SNStype/, 16 July, 2009, Last Accessed: 23 August, 2011.
30. Mutum, D., Wang, Q. “Consumer Generated Advertising in Blogs.” in Neal M. Burns, Terry Daugherty, Matthew S. Eastin. Handbook of Research on Digital Media and Advertising: User Generated Content Consumption, pp. 248–261, 2010.
31. Berry, R. “Blog 101: An Overview of Web log Technologies.” in STC Proceedings (Tools and Technology section), pp. 216-220, 2004.
32. Baloglu, A., Aktas. M. “BlogMiner: Web Blog Mining Application for Classification of Movie Reviews.” in Fifth International Conference on Internet and Web Applications and Services, IEEE Computer Society Transaction, pp.77-84, 2010.
33. Goncalves, M., Almeida, J., Santos, L., Laender, A., Almeida, V. “On Popularity in the Blogosphere.” in Social Computing Transaction, Published by the IEEE Computer Society, pp. 42-49, 2010.
34. Flynn, N. “Why Blog Rules?” in Blog Rules : A Business Guide to Managing Policy, Public Relations, and Legal Issues, New York, AMACOM: American Management Association, pp. 3-12, 2006.

35. Gill, A., Nowson, S., Oberlander, J. “What Are They Blogging About? Personality, Topic and Motivation in Blogs.” in Proceedings of the Third International AAAI Conference on Blogs and Social Media, pp. 18-25, 2009.
36. Invernizzi, L., Kruegel, C., Vigna, G. "Message in A Bottle: Sailing Past Censorship" in 5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs), 2012.
37. Gyongyi, Z., Garcia-Molina, H., Pedersen, J., “Web content categorization using link information”, Technical Report, Stanford University, 2006.
38. Qu, H., Pietra, A. L., Poon, S. “Automated Blog Classification: Challenges and Pitfalls.” in N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin (Eds.), Computational Approaches to Analyzing Blogs: Papers from the 2006 Spring Symposium, AAAI Press. Technical Report, pp. 184–186, 2006.
39. Ranganathan, S. R. “Library Classification on the March.” in the Sayers Memorial Volume, Library Association, London, pp. 84, 1961.
40. Yu, N., Semi-Supervised Learning for Identifying Opinions in Web Content, Ph.D. thesis, Indiana University, 2011.
41. Egele, M., Kolbitsch, C., Platzer, C. “Removing Web Spam Links from Search Engine Results.” Journal in Computer Virology, vol. 7, pp.51-62, 2011.
42. Gao, W., Tian, Y., Huang, T. “Vlogging: A Survey of Videoblogging Technology on the Web .” in ACM Computing Surveys, Vol. 42, No. 4, Paper 15, pp.1-57, 2010.
43. Hurst, M., Maykov, A. “Social Streams Blog Crawler.” in Proceedings of the 2009 IEEE International Conference on Data Engineering, pp. 1615–1618.
44. Pandey, R., Dwivedi, S. “Interoperability between Semantic Web Layers: A Communicating Agent Approach” in International Journal of Computer Applications, Volume 12, No.3, 2010.
45. Mina, Nima. “Blogs, Cyber-Literature and Virtual Culture in Iran”, George C. Marshall, European Center for Security Studies, N.15, P. 6, 2007.
46. Kargar, M., Ramli, A., Ibrahim, H., Azimzadeh, F. “Formulating Priority of Information Quality Criteria on the Blog.” in World Applied Sciences Journal Vol. 4 No. 4, pp. 586-593, 2008.
47. Alexa Co. “Statistics Summary”, Available from: www.alexa.com, Last Accessed: Dec. 21, 2011.
48. Google Co. “Google Search”, Available from: http://google.com, Last Accessed: Jan, 1, 2012.
49. Google Co. “Google Trends”, Available from: http://trends.google.com, Last Accessed: Jan, 2, 2012.
50. Google Co. “Google Search”, Available from: http://google.com, Last Accessed: Jan, 8, 2012.
51. Gedik, B., Andrade, H., Wu, K. L., Yu, P. S., Doo, M. “SPADE: The System S Declarative Stream Processing Engine.”, In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, NY, USA, pp. 1123-1134. 2008.
52. Suzumura, T., Oiki, T. “StreamWeb: Real-Time Web Monitoring with Stream Computing.”, In Proceedings of the 2011 IEEE International Conference on Web Services, IEEE Computer Society, Washington, DC, USA, pp. 620-627, 2011.
53. Foo, K. C., Jiang, Z. M., Hassan, A. E., Zou, Y., Martin, K., Flora, P. “Modelling the performance of Ultra-Large-Scale Systems Using Layered Simulations.”, 2011 IEEE International Workshop on the Maintenance and Evolution of Service-Oriented and Cloud-Based Systems (MESOCA 2011), Williams-burg, VA, USA, 2011.
54. Richards, R. “Content Syndication: RSS and Atom” in Pro PHP XML and Web Services Book, Apress, pp. 521-566, 2006.
55. Roden, T. “Realtime Syndication” in Building the Realtime User Experience Book, O’Reilly Media, Inc., pp. 9-36, 2010.