مجله علمی پژوهشی «علوم و فناوری‌های پدافند غیرعامل»
سال چهارم، شماره۲، تابستان ۱۳۹۲؛ ص ۱۳۶- ۱۲۵

کشف برخط تغییرات وب‌نوشت‌ها با ردیابی چکیده آنها

مهدی نقوی[1*]، محسن شریفی[2]

۱- دانشجوی دکتری و ۲- استاد دانشگاه علم و صنعت ایران

(دریافت: ۱۳۹۱/۰۹/۱۳، پذیرش: ۱۳۹۲/۰۳/۲۲)

چکیده

رشد نمایی اطلاعات در فضای سایبر در کنار پیشرفت سریع فناوری‌های مرتبط با آن، شکل جدیدی از رقابت بین کشورها و فرهنگ‌های مختلف برای به‌دست آوردن برتری اطلاعاتی در این فضای بسیار مهم و ارزشمند را ایجاد کرده است. برای همه ذینفعان بسیار مهم است که نقش برتری را در تولید اطلاعات و پایش اطلاعات بار شده در این فضا بازی کنند. نقش برتر یک ذینفع در فضای سایبر، فرصت‌های جدیدی را در اختیار وی قرار داده و ظرفیت‌های پدافند غیرعامل وی را نیز در زمینه‌های سیاسی، اخلاقی، علمی، آموزشی و تفریحی افزایش می‌دهد. لیکن به‌دست آوردن برتری در پایش اطلاعات انبوه در این فضای سایبری نیازمند بهره‌گیری از فناوری‌های پایش بی‌درنگ با استفاده از رویکردها و فناوری‌های جدید کاملاً متفاوت از فناوری‌های سنتی پایش است. در این مقاله، با تمرکز بر وب‌نوشت‌ها به‌عنوان بخش مهمی از رسانه‌های اجتماعی فضای سایبر، فضای هدف را محدود ساخته و یک روش جدید کشف برخط تغییرات ایجاد شده در وب‌نوشت‌ها معرفی می‌شود. همچنین نشان داده شده که این روش، بهتر از سایر روش‌ها و فناوری‌هایی عمل می‌کند که نیاز به همکاری و هماهنگی با ارائه دهندگان اطلاعات دارند. دستیابی به این مطلوب با ارائه یک معماری جدید و محدود کردن فضای جستجوی تغییرات از حجم عظیمی از وب‌نوشت‌ها به چکیده وب‌نوشت‌ها، میسر شده است.

کلید واژه‌ها: فضای سایبر، پدافند غیرعامل، رسانه‌های اجتماعی، وب‌نوشت، چکیده وب‌نوشت، کشف برخط تغییرات.

Online Detection of Changes in Blogs by RSS Tracking

M. Naghavi[*], M. Sharifi
Iran University of Science and Technology

(Received: 03/12/2012; Accepted: 12/06/2013)

Abstract

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 dominance in this critical and invaluable space. It has thus become quite critical to all stakeholders to play a leading and dominant role in generation of information and monitoring of voluminous information uploaded to this space. The dominant role of a stakeholder in cyberspace provides him with new opportunities and builds up his capacity for passive defense in the political, ethical, scientific, educational and recreational fields. However, to gain dominance in the monitoring of vast amount of information in cyberspace requires new techniques and approaches for online monitoring quite different from traditional ones. Concerned with the latter case, we limit our focus in this paper on blogs as an important part of the cyberspace social media and present a new technique for online detection of changes in blogs. We show that our technique works favorably compared to techniques that require cooperation and synchronization between information providers. This is achieved by providing a new architecture and restricting the search for changes in high volumes of blogs only to changes in the RSS (Rich Site Summary) of each blog.

Keywords: Cyberspace, Passive Defence, Social Media, Blog, RSS Blog, Online Change Detection.

[*]Corresponding Author E-Mail: msharifi@iust.ac.ir

۱. مقدمه

تا پایان سه ماهه دوم ۲۰۱۰ میلادی، حدود ۲۴۰ میلیون دامنه وب ثبت شده [۱] که حدود ۱۳۰ میلیون آن فعال است [۲]. هر چند اطلاعات دقیقی از تعداد صفحات موجود در اینترنت به‌طور رسمی موجود نیست، ولی تعداد صفحه‌های نمایه‌گذاری شده توسط جویشگر[1] گوگل در ژانویه ۲۰۰۸ حدود ۳۰ میلیارد و توسط جویشگر یاهو ۳۷ میلیارد صفحه برآورد شده است [۳]. این در حالی است که گزارش‌های ارائه شده حاکی از دست یافتن جویشگر گوگل به یک تریلیون آدرس منحصر بفرد می‌باشد [۴] که به دلایل مختلف از جمله نداشتن ظرفیت‌های لازم نتوانسته است همه آنها را واکشی و نمایه‌گذاری نماید. بررسی‌ها نشان می‌دهد که روند به‌روز بودن اطلاعات موتورهای جستجو رو به کاهش است. یعنی هر چند موتورهای جستجو تجهیزات و منابع خود را ارتقا می‌دهند، ولی همواره نسبت به روند رشد اطلاعات در وب عقب هستند. به‌طور مثال میزان صفحاتی که در جستجوگر گوگل در سال ۲۰۰۵ به‌روز بوده‌اند حدود ۸۳ درصد بوده و این میزان در سال ۲۰۰۷ به ۲۴ درصد کاهش پیدا کرده است [۵].

در این مقاله، روشی ارائه خواهد شد که از تغییرات به‌وجود آمده در وب‌نوشت‌های[2] کشوری مانند ایران به‌صورت برخط آگاهی پیدا می‌کنیم، بدون آن که به منابع زیاد پردازش و یا پهنای باند زیاد نیاز باشد. تحقیق‌های دیگر مانند ردیابی و تحلیل وب‌نوشت‌ها [۶]، تشخیص و کشف رویدادها در شبکه‌های اجتماعی [۷] و دیده‌بانی تغییرات به‌وجود آمده در وبگاه‌ها [۸] در این زمینه انجام شده که برون‌خط بوده و یا از تکنولوژی‌هایی نظیر Ping [۹] استفاده کرده‌اند که بسیاری از سرویس دهنده‌های وب‌نوشت به‌خصوص در ایران از آن پشتیبانی نمی‌کنند.

آنچه که در این مقاله بر آن تمرکز شده، به‌دست آوردن این اطلاعات به‌صورت برخط و به‌موقع، با توجه به محدودیت‌های منابع مانند قدرت پردازش و پهنای باند است. تمرکز این مقاله بر روی وب‌نوشت‌ها و طراحی سیستمی که به‌روزرسانی‌های به‌وجود آمده در آنها را به‌صورت برخط اعلام نماید، ایجاد ظرفیت جدیدی برای پدافند غیر عامل خواهد بود. بسیاری از فعالیت‌های گروهی رسمی و یا غیر رسمی در زمینه‌های مختلف سیاسی، علمی، آموزشی، تفریحی و اخلاقی در فضای سایبر رخ می‌دهد. اشراف اطلاعاتی بر این فضا و آگاهی از تغییراتی که در صحنه آن رخ می‌دهد، نوعی پدافند غیر عامل در برابر مخاطرات این‌گونه فعالیت‌ها است. سیستم طراحی شده که در این مقاله ارائه می‌شود، پیش‌نیاز رصد برخط فضای وب خواهد بود. این سیستم هم‌اکنون بر پایه همین معماری پیاده‌سازی شده و پدیده‌های مهم[3] فضای وب‌نوشت‌های کشور را کشف و ارائه می‌کند.

در بخش ۲ این مقاله، فضای سایبر، رسانه‌های اجتماعی و وب‌نوشت‌ها تشریح خواهند شد. در بخش ۳ کارهای مرتبط انجام شده معرفی می‌شوند. چالش‌های بررسی وب‌نوشت‌ها در بخش ۴ و اهدافی که از بررسی وب‌نوشت‌ها و آگاهی به‌موقع از تغییرات آنها حاصل می‌شود در بخش ۵ ارائه می‌شوند. در بخش ۶ سامانه برخط کشف تغییرات وب‌نوشت‌ها با توجه به وضعیت موجود وب‌نوشت‌ها در کشور و محدویت‌های پیش رو بررسی شده و در بخش ۷ معماری سامانه ODCB[4] به‌همراه نتایج پیاده‌سازی و آزمایش‌های انجام شده ارائه می‌شود. در این بخش به چگونگی آگاهی برخط و به‌موقع از تغییرات به‌وجود آمده در وب‌نوشت‌های یک کشور نمونه مانند ایران پرداخته شده و در ادامه نتایج مورد ارزیابی و تحلیل قرار می‌گیرند. در آخرین بخش این مقاله جمع‌بندی و نتیجه‌گیری صورت گرفته است.

۲. فضای سایبر

فضای سایبر[5] ترکیبی از کلمه فضا و لغت سایبر می‌باشد که از سایبرنتیک مشتق شده است. سایبرنتیک نظریه‌ای است که مناسبات انسان و ماشین و مناسبات ماشین‌ها با یکدیگر را تبیین می‌کند. این نظریه را "نوربرت وینر" در سال ۱۹۴۸ مطرح کرده است [۱۰]. ترکیب "فضای سایبر" اولین بار در سال ۱۹۸۴ در یک رمان علمی تخیلی آمده [۱۱] و در حال حاضر از جمله عبارت‌هایی است که در فضای وب و شبکه اینترنت بسیار استفاده می‌شود. این عبارت معادل فضای مجازی به‌کار برده شده و ترکیبات بسیاری نظیر شهروند سایبر، پول سایبر، فرهنگ سایبر، تجارت سایبر و عبارت‌های مختلفی از این نوع را شامل می‌شود.

اهمیت فضای سایبر و یا فضای مجازی روزبه‌روز مشهودتر می‌شود، به‌طوری که وزارت دفاع آمریکا مرکز فرماندهی سایبری امریکا را تأسیس کرده و با انتشار یک سند راهبردی درباره "عملیات در فضای سایبر"، اعلام کرده است که در این فضا دست به عملیات نظامی خواهد زد [۱۲]. پنتاگون در این راهبرد اعلام کرده است فضای سایبر نیز مانند زمین، هوا و دریا عرصه جنگی است و آن را به‌عنوان یک میدان جدید عملیاتی معرفی کرده است.

معادل قرار دادن فضای سایبر با سایر عرصه‌های جنگی، نشانگر اهمیت عرصه فضای سایبر برای ایالات متحده آمریکا است. در چنین فضایی شناخت هر چه بیشتر این عرصه بسیار مهم بوده و می‌تواند در تبدیل تهدیدها به فرصت ها بسیار کمک نماید. رسانه‌های اجتماعی در فضای سایبر یکی از بهترین و کم‌هزینه‌ترین بسترها برای آگاهی از افکار عمومی و کنترل آنها هستند. در ادامه این رسانه‌ها و وب‌نوشت‌ها به‌طور مختصر معرفی خواهند شد.

[1] Search Engine
[2] Blog/Weblog
[3] Hot Trends
[4] Online Detection of Changes in Blogs
[5] Syberspace

۲-۱. رسانه‌های اجتماعی در فضای سایبر

رسانه‌های اجتماعی در فضای سایبر یکی از سریع‌ترین پدیده‌های رو به‌رشد فضای وب است که امکان تعاملات مختلف را بین کاربران پدید آورده‌اند. این رسانه‌ها انواع مختلفی از خدمات مانند رابطه‌های دوستانه تعاملی، رابطه‌های یک‌سویه، باخبر شدن از وضعیت دوستان به‌صورت برخط، اعلام نظرات، پیام های خصوصی و عمومی، آلبوم های چند رسانه‌ای، رویدادها و انواع چت های متنی، گفتاری و ویدیوئی را برای کاربران خود فراهم می‌کنند. بر اساس تعاریف و مشخصات این رسانه‌ها [۱۶-۱۳] می‌توان آنها را در قالب جدول (۱) دسته‌بندی و ارائه کرد.

اطلاعات به‌روز و بی‌شمار رسانه‌های اجتماعی، دانشمندان را ترغیب کرده تا با استفاده از این اطلاعات ارزشمند، کاربردهای بسیاری در زمینه‌های مختلف از جمله استخراج نظرهای مردم از وب‌نوشت‌ها [۱۳ و ۱۷]، خلاصه‌سازی خودکار وب‌نوشت‌ها و رویدادها [۱۹-۱۸]، استخراج پدیده‌ها و رویدادهای مهم از وب‌نوشت‌ها، میکروبلاگ‌ها و شبکه‌های اجتماعی [۱۳و ۲۸-۲۰] را تعریف و مورد استفاده قرار دهند.

جدول ۱. انواع رسانه‌های اجتماعی وب

رسانه‌ها	شرح مختصر	مثال
وب‌نوشت‌ها	فضای باز شبکه سازی کاربران با هم و ثبت نظرات و اطلاعات برخط	Blogger Wordpress Blogfa PersianBlog
شبکه‌های اجتماعی	فضای مشترک اختصاصی‌تر از وب‌نوشت‌ها-ژورنال‌های شخصی برخط	MySpace Facebook LinkedIn
ویکی‌ها	فضای مشترک کاربران برای ذخیره، ویرایش و دیدن محتوا	Wikipedia Wikia Wikinews
فروم‌ها	ویژه بحث و گفتگو- از قدیمی‌ترین نوع رسانه‌های اجتماعی برخط	StonedLizard DiscussionLounge
میکرو بلاگ‌ها	تلفیقی از شبکه‌های اجتماعی و وبلاگ‌نویسی با انواع پیام‌های فوری	Twitter Pownce Jaiku

۲-۲. وب‌نوشت‌ها

وب‌نوشت‌ها به‌عنوان یکی از متداول‌ترین رسانه‌های اجتماعی، تحولی اساسی در ارتباطات اجتماعی فضای وب به‌وجود آورده اند. مهم‌ترین خصیصه‌ای که وب‌نوشت‌ها را از سایر سایت‌های معمولی متمایز می‌سازد، خاصیت همه‌گیر بودن و به‌روز شدن لحظه‌ای آن است. با توجه به این ویژگی‌ها، بررسی وب‌نوشت‌ها می‌تواند ما را به اطلاعات با خواص مشخصی رهنمون سازد.

وب‌نوشت نوعی وبگاه است که معمولاً توسط یک فرد با نوشتن یادداشت‌ها، توضیحات، اخبار غیر رسمی مربوط به موضوعات خاص، شرح وقایع و گذاردن مطالب صوتی، تصویری و ویدئویی و پیوندهای آنها ایجاد می‌شود. اکثر وب‌نوشت‌ها، اجازه می‌دهند بازدید کننده‌ها نظرات و پیام‌های خود را از طریق ابزارهای وب‌نوشت‌ها به یکدیگر ارسال نمایند. این نوع اعمال، پویایی سایت ها را بیشتر کرده و یکی از مواردی که باعث تمایز وب‌نوشت‌ها از سایر سایت‌های ایستا می‌شود، همین پویایی آنها است [۲۹].

وب‌نوشت‌ها در سال ۱۹۹۷ مطرح و تعریف شده‌اند [۳۰]. تا سال ۲۰۰۹ میلادی بیش از ۱۳۳میلیون وب‌نوشت توسط موتور جستجوی تکنوراتی، که موتور جستجویی برای وب‌نوشت‌ها می‌باشد، نمایه شده [۳۱] و بیش از ۹۰۰هزار پست وب‌نوشت جدید در هر ۲۴ ساعت ثبت شده است [۳۲]. از آنجا که وب‌نوشت در سال ۱۹۹۷ ابداع شده [۳۳]، این آمار بیانگر رشد بسیار سریع و غیر قابل پیش‌بینی وب‌نوشت‌ها است. ایجاد یک وب‌نوشت در هر ثانیه در سال ۲۰۰۶ [۳۴] و ارتقاء آن به ۱۱ وب نوشت در هر ثانیه در سال ۲۰۱۰ میلادی، نشانگر رشد شتابان وب‌نوشت‌ها و اهمیت آنها است.

۳. پیشینه تحقیق

اگروال و همکارانش [۶] سامانه‌ای را برای کمک به جامعه‌شناسان جهت ردیابی و تحلیل وب‌نوشت‌ها در سال ۲۰۰۹ ارائه کرده‌اند. در این سامانه پس از معرفی وب‌نوشت‌های مورد نظر، آنها را واکشی کرده و پس از نمایه‌گذاری، آماده برای ارائه تحلیل و آمار مورد نظر می‌سازد. یکی از این ابزارها، تحلیل‌گر تکرار واژه است. این ابزار می‌تواند تعداد تکرار واژه‌های کلیدی را که در یک بازه زمانی در وب‌نوشت‌ها استفاده شده‌اند را مشخص نماید. همچنین می‌تواند نویسنده‌های وب‌نوشت‌ها را به فعال با نفوذ، غیر فعال با نفوذ، فعال بدون نفوذ، و غیر فعال بدون نفوذ، طبقه‌بندی نماید. این نرم‌افزار یک سامانه آگاه ساز به کاربر ارائه می‌دهد. با این سامانه کاربر می‌تواند واژه‌های مشخصی را در فهرست دیده‌بانی تعیین نموده و در صورتی که این واژه در پست جدیدی بیاید، از طریق ایمیل به کاربر خبر داده شود. از مزیت‌های این سامانه این است که هر دو عمل جمع‌آوری اطلاعات و تحلیل اطلاعات را خودش انجام می‌دهد. از نقاط ضعف آن این است که به‌صورت برخط کار نکرده و به‌جای اینکه فقط تغییرات را واکشی نماید کل صفحات وب‌نوشت‌ها را واکشی می‌کند.

خزش گراف شبکه‌های اجتماعی برخط، مورد بررسی یه و همکاران [۳۵] در سال ۲۰۱۰ میلادی قرار گرفته است. چالش‌های مورد بررسی در این کار، مشابه چالش‌هایی بوده که ما در کار خود با آن روبرو بوده‌ایم. از جمله مواردی که در این تحقیق بررسی شده است، خزش سریع‌تر گره‌های شبکه‌های اجتماعی برخط، تأثیر شبکه های اجتماعی و کاربران حفاظت شده آنها بر روی خزشگرها، چگونگی تعریف مشخصات گراف کامل و تفاوت آن با زیرگراف های خزش شده می‌باشد. برای انتخاب گره‌ها از الگوریتم‌های BFS [۳۶]، شانسی، حریصانه، و حریصانه فرضی [۳۵و ۳۸- ۳۷] استفاده شده است. برخی از چالش‌های خزش شبکه‌های اجتماعی مانند حجم زیاد اطلاعات، تعداد زیاد صفحات پویا، خزش مؤثر، مشکل دریافت

اطلاعات از سرویس‌دهندگان شبکه‌های اجتماعی و به‌روزرسانی بی‌درنگ این شبکه‌ها، در این تحقیق بررسی شده است. سیاه‌چاله‌ها که مسیر بن‌بست برای خزشگرها محسوب می‌شوند، یکی از مشکلات بزرگ خزش شبکه های اجتماعی برخط بیان شده است. این سیاه‌چاله‌ها از آنجا به‌وجود می‌آید که کاربران اطلاعات شخصی خود را به‌صورت محرمانه حفظ کرده و از دستیابی دیگران به آنها جلوگیری می‌کنند. ادامه گراف و ساخت آن از طریق این‌گونه کاربران که بسیار هم هستند، بدون همکاری سرویس‌دهندگان شبکه‌های اجتماعی تقریباً غیرممکن است.

تشخیص و کشف رویدادها در شبکه های اجتماعی، کار دیگری است که در سال ۲۰۰۹، مورد بررسی صیادی و همکارانش [۷] در شرکت مایکروسافت قرار گرفته است. در این تحقیق گرافی از کلمه‌های کلیدی جهت مستندات ایجاد شده و برای هر رویداد، خوشه‌ای از این کلمه‌های کلیدی منظور شده است. شبکه کلمه‌های کلیدی بر اساس همکاری در ایجاد یک سند مشترک، ساخته می‌شود. در این گراف، کلمه های کلیدی کم تکرار فیلتر شده و کلمه‌های باقیمانده، گره‌ها را تشکیل داده و یال‌ها بر اساس همکاری آنها در تشکیل سند ایجاد می‌شوند. اگر گره‌های دارای کلمه‌های i و j در ایجاد سند همکاری داشته باشند یال $e_{i,j}$ ایجاد می‌شود. در صورتی که اشتراک کلمه های کلیدی همکار، کمتر از آستانه تعیین شده باشد، یال مربوطه حذف می‌شود. اگر احتمال دیدن کلمه iام در سند، به شرط وجود کلمه jام در سند از یک حد آستانه نیز کمتر باشد، یال حذف می‌شود. در این آزمایش برای دستیابی به اطلاعات مورد نیاز از ۱۸۰۰۰ پست وب نوشت در یک بازه دو ماهه استفاده شده است.

اُه و همکارانش [۱۷] به بررسی یک سامانه طبقه‌بندی نقطه نظرهای سیاسی پست‌های وب‌نوشت‌ها و ارزیابی کاربران وب از نتایج آن پرداخته اند. این سامانه مبتنی بر یک مدل یادگیر طبقه‌بندی کننده با استفاده از الگوریتم یادگیری هدایت شده است. داده‌های ورودی آخرین پست‌های وب‌نوشت‌ها بوده که به‌عنوان نظرات لیبرال و یا محافظه‌کار دسته‌بندی شده و سپس با ارائه یک مدل اولیه برای بازیابی و طبقه‌بندی وب‌نوشت‌های سیاسی، نتایج طبقه بندی شده به کاربر نشان داده شده و مورد ارزیابی او قرار گرفته و در بهبود یادگیری سامانه نیز مورد استفاده قرار می‌گیرد. در این کار از آدرس‌های از پیش آماده شده استفاده شده و بر اساس آن اطلاعات وب‌نوشت‌ها را واکشی و دسته‌بندی کرده است. در پیاده‌سازی سیستم فوق از مفاهیم داده‌کاوی، خوشه‌بندی و گراف استفاده شده است.

جیل و همکارانش [۳۹] در سال ۲۰۰۹، شخصیت و انگیزه نویسنده و موضوع وب‌نوشت‌ها را مورد بررسی قرار داده‌اند. در این کار، با استناد و ارجاع به تئوری خصیصه، شخصیت را به تعدادی از عوامل قابل اندازه گیری یا صفات تجزیه کرده و بر اساس یک مدل پنج عاملی شخصیت را مورد بررسی قرار داده است. همچنین بررسی شده است که چگونه موضوعات، مورد بحث و گفتگوی وب نوشت ها قرار می‌گیرند. این مقاله در درک بهتر موضوعات وب‌نوشت‌ها، نویسندگان و علاقه‌مندی‌های آنها کمک می‌کند.

پاتاک و همکارش [۸] روش هوشمندی برای دیده‌بانی تغییرات به‌وجود آمده در سایت ها، ارائه کرده‌اند. در این روش علاقه‌مندی کاربران منظور شده و میزان تغییرات به‌وجود آمده رتبه بندی شده است. یک سامانه نمونه به‌نام WebMon برای آزمایش روش ارائه شده نیز عرضه شده است. برای منظور کردن علاقه های کاربران، از کلید واژه‌های وزن‌گذاری شده توسط کاربر استفاده شده است. از الگوریتم مدل فضای برداری (VSM) [۴۰] برای محاسبه میزان تغییرات به‌وجود آمده، استفاده شده است. کار با آدرس سایت هایی که کاربر مشخص کرده است شروع شده و نتیجه آن با ارائه فهرستی از تغییرات به‌وجود آمده به‌صورت یک عدد اعشاری بین صفر و یک اعلام می‌شود.

۴. چالش‌های بررسی وب‌نوشت‌ها

با توجه به حجم عظیم وب‌نوشت‌ها، بررسی و مطالعه آنها با چالش‌هایی روبرو است. در ادامه این چالش‌ها بیان می‌شود.

۴-۱. به‌روز شدن سریع وب‌نوشت‌ها

به لحاظ زمانی ماهیت وب‌نوشت‌ها بسیار زودگذر است. وب‌نویس‌ها علاقه‌مند هستند که مطالب جدید در پست‌ها قرار داده و خوانندگان نیز علاقه‌مند هستند که بر وب‌نوشت‌ها توضیح بنویسند. بنابراین بسیار مناسب خواهد بود که به‌محض تغییرات در وب‌نوشت‌ها و گذاشتن پست جدید توسط کاربران، به نحوی باخبر شده تا بتوانیم از اطلاعات جدید بهره مند گردیم. فناوری Ping برای پشتیبانی از این چنین نیازمندی‌هایی ارائه شده است [۴۱]. فناوری Ping که برای اولین بار در سال ۲۰۰۱ معرفی شد [۴۲]، یک روش فراخوانی پردازه از راه دور مبتنی بر XML است که وب‌نوشت‌های یک ماشین خدمت‌گذار که خدمت‌گذار Ping نام دارد را از تغییرات به‌وجود آمده در پست‌های خود باخبر می‌سازد [۹].

۴-۲. زودگذر بودن ماهیت وب‌نوشت‌ها

اطلاعات وب‌نوشت‌ها نسبت به گذر زمان حساس هستند. در اغلب موارد به‌دست آوردن اطلاعات پست‌های وب‌نوشت‌ها و پردازش آنها در زمان‌های کوتاه بسیار مهم است. وب‌نوشت‌ها دارای ویژگی‌هایی هستند که اطلاعات آنها پس از گذشت زمان خاصی بی‌اثر می‌شوند. به‌همین دلیل بررسی وب‌نوشت‌ها باید در زمان معین انجام پذیرد.

۴-۳. مشکل دسته‌بندی وب‌نوشت‌ها

دسته‌بندی وب‌نوشت‌ها یکی از مسائل اساسی بازیابی اطلاعات وب است [۴۳]. دسته‌بندی وب‌نوشت‌ها توسط انسان و حتی برای ماشین بسیار مشکل است. به این دلیل برای انسان مشکل است چون حجم عظیمی از وب‌نوشت‌ها در برابر او قرار داشته و نیروی

انسانی زیادی لازم دارد تا بتواند با بررسی و عقل و دانش انسانی آنها را دسته‌بندی نماید. برای ماشین مشکل است چون از ساختار مشخص و معینی پیروی نکرده و به علاقه‌ها و سلیقه‌های افراد بستگی داشته و بسیاری از اوقات به‌صورت نامنظم، سرگردان و غیر معقول نوشته شده‌اند. برای دسته‌بندی وب‌نوشت‌ها می‌توان از ویژگی‌های زبان‌شناسی عنوان پست‌ها در وب‌نوشت‌ها و متن پیوندها استفاده کرد. آزمایش‌ها نشان داده است که این‌گونه دسته‌بندی‌ها خیلی موفق نبوده‌اند، چرا که وب‌نوشت‌ها را نمی‌توان تحت عناوین مشخص و منظم دسته‌بندی کرد [۴۴] . برای یک وب‌نوشت نمی‌توان به‌طور قطع یک دسته‌بندی مشخص تعیین نمود و ممکن است در چندین دسته‌بندی قرار گیرد. وجود آثار با طبقه بندی چندوجهی مطابق نظر رانگاناتان [۴۵]، ارائه دهنده طبقه‌بندی وجهی، کاملاً طبیعی بوده و این موضوع برای وب‌نوشت‌ها مشهودتر است [۴۴].

۴-۴. وب‌نوشت‌های هرز و پیوند به آنها

صفحات مزاحم وب‌نوشت‌ها که وب‌نوشت هرز نام گرفته‌اند، یکی از مشکلات اساسی بررسی وب‌نوشت‌ها بیان می‌شوند. طبق گزارش تکنوراتی در هر روز بین ۳۰۰۰ تا ۷۰۰۰ وب‌نوشت هرز جدید ایجاد شده و گاهی نیز به ۱۱۰۰۰ صفحه در روز می‌رسد[۴۶]. پیوندهای هرز از توضیحات و پاسخ هایی که به‌صورت پویا پشتیبانی می‌شوند، به‌وجود می‌آید. صفحات هرز ناشی از توضیحات، به آسانی در وب‌نوشت‌ها قابل ایجاد شدن هستند. یک هرزنویس می‌تواند یک نرم‌افزار تولید کند که به‌صورت تصادفی به وب‌نوشت‌های مختلف دسترسی پیدا کرده و در پست‌های توضیحات آنها، پیوندهای بازگشت به صفحه هرز را بگذارد [۳۳]. صفحات و پیوندهای مزاحم یک چالش اساسی محسوب می‌شوند. این پیوندها ارتباطات کاذبی را برقرار می‌کنند که از بار معنایی برخوردار نیستند.

۴-۵. استفاده از زبان غیر رسمی در وب‌نوشت‌ها

استفاده از زبان‌های گفتاری و غیررسمی یا محاوره‌ای در وب‌نوشت‌ها باعث تنوع بسیار زیادی در واژه‌ها و مفاهیم آنها می‌شود. زبان وب‌نوشت‌ها حاوی واژه‌های متعدد و جدید مربوط به اصطلاحات وب‌نوشتی (مانند کلیک، کامنت، پست، آپدیت، آف گذاشتن و ...) و همچنین واژه‌های بیگانه با رسم‌الخط بیگانه است. همچنین این نوشته‌ها حاوی غلط‌های تایپی و نگارشی به نسبت زیادی هستند. این موارد درک مفهوم واژه‌ها و بررسی هدفمند وب‌نوشت‌ها را مشکل می‌سازد. همچنین معادل‌سازی عبارت‌ها را مشکل‌تر ساخته و برای ایجاد لغت‌نامه، نقش ماشین را کمتر کرده و نیاز به حضور نیروی انسانی را بیشتر می‌نماید.

۴-۶. وجود میلیون‌ها وب‌نوشت خالی بدون پست

با توسعه سرویس دهندگان وب‌نوشت، حجم زیادی از وب‌نوشت‌ها توسط کاربران غیر فعال در این زمینه، ایجاد شده که بی‌استفاده رها شده‌اند. در این صورت بررسی میلیون‌ها صفحه خالی باعث هدر رفتن بخش عظیمی از منابع مانند پهنای باند و پردازنده می‌شود. یکی از راه‌های جلوگیری از این امر، تشکیل بانک صفحات خالی است. هر چند شناخت صفحات خالی نیز، عاری از مشکل نخواهد بود، زیرا بسیاری از صفحات وب‌نوشت‌ها در ابتدای تشکیل خالی بوده و به‌تدریج تکمیل می‌شوند، بنابراین باید مکانیزم مناسبی جهت شناخت صفحات خالی پیاده‌سازی شود.

۴-۷. از هم گسستگی مستندات وب‌نوشت‌ها

در وب‌نوشت‌ها، مستندات معمولاً از هم گسسته هستند. در حالی که در صفحات معمول وب از طریق زیرصفحه‌ها به یکدیگر متصل هستند [۴۱]. این باعث می‌شود هنگام خزش و واکشی صفحات وب‌نوشت‌ها، روند پیوسته و مرتبطی را نتوان طی کرد. این مشکل در وب‌نوشت‌هایی که توسط سرویس دهندگان آنها امکان درج پیوندها حذف شده است بیشتر خودش را نشان خواهد داد، زیرا از پیوندهای به صفحات دیگر کمتر استفاده می‌شود.

۵. اهداف استفاده از وب‌نوشت‌ها برای آگاهی برخط

بر طبق آنچه در بخش های قبلی اشاره شد، ویژگی‌های خاص وب‌نوشت‌ها آنها را به منابع ارزشمندی تبدیل کرده است که بررسی هدفمند آنها می‌تواند ما را به اطلاعات ارزشمند و گرانبهایی برساند. اهدافی که از بررسی وب‌نوشت‌ها حاصل خواهد شد در ادامه بیان می‌شود.

۵-۱. یافتن علاقه‌مندی‌های جامعه وب‌نوشت

با پیشرفت فناوری‌های مربوط به اینترنت، رشد قابل توجهی در تعداد کاربران اینترنت به‌وجود آمده است. تا ژوئن ۲۰۱۰ میلادی نزدیک به دو میلیارد (۱,۹۶۶,۵۱۴,۸۱۶) کاربر اینترنت در جهان برآورد شده است [۴۷]. بخشی از این تعداد عظیم، کاربران وب‌نوشت‌ها بوده که یا نویسنده وب‌نوشت هستند و یا خواننده آن می‌باشند. یافتن علاقه‌مندی‌های جامعه وب‌نوشت، می‌تواند فرصت‌های فراوانی را پیش رو قرار دهد. از جمله این فرصت‌ها، حضور فعال در تجارت با شناخت نیازمندی‌های کاربران، فرهنگ‌سازی بر بستر علاقه‌های فردی و گروهی افراد، انتشار هدفمند نقطه‌نظرهای سیاسی با شناخت گرایش‌های کاربران و آگاهی از روند گسترش جرم‌های خاص بر بستر اینترنت است.

۵-۲. اطلاع از روند روز موضوعات مختلف

در جوامع امروزی وقوع اتفاقات و وقایع مختلف روند سریعی دارد. به‌ویژه با گسترش فناوری‌های جدید رسانه‌ای، انتشار آنها نیز به روش‌های مختلف و با سرعت بالایی صورت می‌پذیرد. بخشی از این اتفاقات در دنیای وب منتشر می‌شود. اطلاع از روند روزانه اتفاقات در موضوعات مورد علاقه و آگاهی از اخبار و پیشامدهای مهم که در دنیای وب منتشر می‌شود، از دیگر کاربردهای بررسی برخط وب‌نوشت‌ها است.

۵-۳. به‌دست آوردن نبض جامعه کاربران وب

به‌دست آوردن نبض و تب یک جامعه از طریق متون وب‌نوشت‌ها کاربردی دیگر از بررسی وب‌نوشت‌ها است. آگاهی از حساسیت‌های جامعه و محورهایی که افکار جامعه حول آن متمرکز می‌شود، مورد علاقه جامعه‌شناسان و سیاستمداران است. طی ۱۰ سال گذشته تعداد کاربران اینترنت نسبت به سال ۲۰۰۰ میلادی، ۴۴۴ برابر شده است [۴۸]. با گسترش روزافزون کاربران وب از سطوح مختلف جامعه، صحنه وب به‌تدریج به عصاره یک جامعه نزدیک می‌شود. در این صورت به‌دست آوردن نبض این جامعه، نمایانگر نبض جامعه بزرگتر خواهد بود.

۵-۴. آگاهی از تأثیر حوادث بر روی افکار عمومی

شاید بتوان جامعه کاربران وب که بخشی از اجتماع بشری هستند را به‌عنوان نماینده بخشی از افکار عمومی یک جامعه در نظر گرفت. در این صورت با زیر نظر داشتن نقطه‌نظرهای ایشان، می‌توان به نقطه نظرهای جامعه مورد نظر نیز پی برد. حوادث و پیشامدهای طبیعی مانند زلزله، سیل، طوفان، رخدادهای غیر طبیعی مانند جنگ و یا رخدادهای اجتماعی مانند انتخابات، اعمال قوانین سراسری که گریبان‌گیر آحاد جامعه است مانند افزایش و یا کاهش حقوق و مالیات، جامعه را به‌شدت متأثر می‌سازد. آگاهی سریع از تأثیرات این اتفاقات بر روی افکار عمومی جامعه، می‌تواند باعث اندیشیدن تدابیر لازم جهت جلوگیری از بحران‌های مختلف و پیش بینی‌های لازم شود. بررسی وب می‌تواند کمک شایانی جهت آگاهی از جهت گیری افکار عمومی در این‌گونه موارد بنماید.

۶. سامانه برخط کسب خلاصه تغییرات وب‌نوشت‌ها

اطلاع سریع و به‌موقع از تغییرات وب‌نوشت‌ها ما را در رسیدن به اهدافی که در بخش قبل بیان شد، نزدیک می‌نماید. صحنه وب‌نوشت‌ها نمودی از اتفاقاتی است که در جامعه رخ می‌دهد. آگاهی از تغییرات به‌وجود آمده در وب‌نوشت‌ها ما را به سمت اتفاقات جدید در جامعه رهنمون می‌سازد. سامانه آگاه‌ساز تغییرات وب‌نوشت‌ها بر همین اساس طراحی شده است. در این بخش، این سامانه را تشریح کرده و معماری آن ارائه می‌شود. در ارائه این معماری، محدودیت‌هایی وجود دارد که شناخت آنها ما را در ارائه معماری مناسب یاری می‌نماید. همچنین برای ارائه نتایج بهتر، مناسب است که به‌جای بررسی کلیه وب‌نوشت‌ها، بر روی بخش مشخصی از وب‌نوشت‌ها تمرکز شود. به‌طور مثال می‌توان تمرکز بر روی وب‌نوشت‌های مرسوم در یک کشور و یا یک زبان خاص داشت. در ادامه، لزوم تمرکز روی بخش خاصی از وب‌نوشت‌ها را بیان کرده و محدودیت‌های موجود برای بررسی وب‌نوشت‌ها ذکر خواهد شد. سپس با توجه به موارد فوق، معماری سامانه آگاه‌ساز برخط تغییرات وب‌نوشت‌ها ارائه شده است.

۶-۱. تمرکز روی حوزه‌ی خاصی از وب‌نوشت‌ها

همان‌گونه که در بخش یک بیان شد، تعداد صفحات وب که تا ژانویه ۲۰۰۸ میلادی توسط یکی از موتورهای جستجو نمایه گذاری شده‌اند، حدود ۳۰ میلیارد صفحه بوده است [۳]. بررسی این حجم از اطلاعات، مستلزم تجهیزات و هزینه‌های بسیار گران‌قیمتی است. اگر از کل فضای وب، فقط روی وب‌نوشت‌ها تمرکز شود، حوزه عمل از فضای ۳۴ میلیارد صفحه‌ای به فضای ۱۳۳ میلیون صفحه‌ای [۳۲] محدود خواهد شد. لازم است با انتخاب حوزه خاصی از وب‌نوشت‌ها مانند کشور و یا زبان، دامنه هدف را به چند میلیون وب‌نوشت کاهش داد تا بتوان با تجهیزات قابل قبول و معماری مناسبی به هدف مورد نظر دست پیدا کرد.

۶-۲. محدودیت‌های بررسی وب‌نوشت‌ها

با توجه به ویژگی‌های خاص وب‌نوشت‌ها، محدودیت‌هایی نظیر زمان دسترسی به نتایج، ظرفیت پردازش و پهنای باند از جمله مواردی هستند که معماری ارائه شده را تحت تأثیر خواهند داشت. در ادامه به بررسی مختصر این محدودیت‌ها پرداخته می‌شود.

بررسی فضای هدف در زمان معین و قابل قبول از الزامات بررسی برخط وب است. اگر نتایج کار پس از زمان مورد انتظار آماده شود، اطلاعات به‌دست آمده ارزش و کارایی خود را از دست داده و مفید واقع نخواهد شد.

به‌دلیل محدودیت در تهیه سخت افزارهای انبوه، ظرفیت پردازش محدود است. معماری ارائه شده باید بتواند محدودیت سخت‌افزار و تجهیزات را مد نظر داشته و به‌گونه‌ای ارائه شود که با ظرفیت‌های پردازش محدود قابل پیاده‌سازی باشد. ظرفیت پهنای باند نیز از دیگر مواردی است که به‌دلیل تحمیل هزینه، نوعی محدودیت محسوب می‌شود. معماری پیشنهادی باید بتواند ظرفیت‌های کم پهنای باند را نیز پشتیبانی کند.

۶-۳. محدود ساختن حوزه وب‌نوشت‌ها

طبق گزارش اعلام شده توسط سایت تکنوراتی، زبان فارسی یکی از ۱۰ زبان اول وب‌نوشت‌ها در سال ۲۰۰۷ به‌شمار آمده [۴۹] و در سال ۲۰۰۸ تعداد وب نوشت ها در ایران حدود ۲ میلیون برآورد شده است [۵۰]. این آمار نشانگر این مسئله مهم است که وب نوشت نویسی در ایران طرفداران بسیار زیادی دارد و می‌توان با تمرکز روی آن اطلاعات ذی‌قیمتی را از وب استخراج نمود. برای بررسی وب‌نوشت‌های فارسی، تعداد ۸۱ مورد از سرویس‌دهندگان وب‌نوشت‌های فارسی شناسایی شده و مورد بررسی قرار گرفتند. شناسایی این سرویس‌دهندگان با اجرای برنامه نگاشت و کاهش[1] [۵۱] روی نتایج خزش ۳۲ میلیون صفحه فارسی که بر روی پایگاه داده توزیع شده HBase در آزمایشگاه موتور جستجوی ملی صورت گرفت و تعداد ۲۱۳۰۰ آدرس وب‌نوشت استخراج گردید. با پالایش

[1] Map-Reduce

این آدرس‌ها، استخراج سرویس‌دهندگان اصلی با نظر نهایی نیروی انسانی متخصص صورت پذیرفت.

برای بررسی این سرویس‌دهندگان و تهیه آمار مورد نظر، از سرویس‌های ارائه شده توسط شرکت الکسا [۵۲] استفاده شده است. همچنین علاوه بر استفاده از سرویس‌های الکسا، برای استخراج دیگر آمارهای مورد نیاز، از عملگر site: موتور جستجوی گوگل بهره برده و در نهایت جدولی تهیه شده که به‌ازاء بررسی ۸۱ سرویس‌دهنده مختلف وب‌نوشت فارسی، نتایج ۶ فاکتور را نشان می‌دهد. این فاکتورها مربوط به رتبه جهانی، رتبه در ایران، اعتبار، برد، میزان صفحه به کاربر و تعداد صفحات موجود در هر سرویس‌دهنده را نشان می‌دهد. رتبه جهانی، رتبه وب‌نوشت در کل سایت‌های جهان را نشان می‌دهد. رتبه در ایران، نشانگر رتبه ترافیک وب‌نوشت نسبت به کل سایت‌های ایران است. فاکتور اعتبار، میزان ارجاع سایت‌های دیگر به وب‌نوشت‌ها را مشخص می‌کند. همچنین فاکتور برد، نشانگر درصد کاربران وب‌نوشت نسبت به کل کاربران جهان است و فاکتور صفحه به کاربر، متوسط تعداد صفحه بازدید شده توسط هر کاربر را نشان می‌دهد. برای به‌دست آوردن فاکتور ششم که تعداد تقریبی صفحات موجود در هر سرویس‌دهنده را نشان می‌دهد از عملگر site: جستجوگر گوگل استفاده شده است [۵۳].

در جدول (۲) فقط اطلاعات مربوط به ده سرویس دهنده اول وب‌نوشت‌های فارسی آورده شده است. این ده سرویس‌دهنده از نظر همه فاکتورها به‌جزء صفحه بر کاربر، دارای رتبه بالا هستند.

جدول ۲. مشخصات ده میزبان اول وب‌نوشت‌های فارسی

سرویس دهنده	رتبه جهانی	رتبه در ایران	اعتبار	برد	صفحه بر کاربر	تعداد صفحات (میلیون)
Blogfa.com	۱۸۷	۳	۱۷۸۵۰۸	۰/۵۲۹۸	۴/۷۱	۵۰
Mihanblog.com	۴۲۷	۵	۴۷۳۵۴	۰/۲۷۱۶	۳/۲۵	۱۸/۷
Persianblog.ir	۵۷۲	۶	۵۷۰۸۷	۰/۲۲۱۴	۲/۶۴	۹/۳
Blogsky.com	۱۰۴۵	۱۳	۳۵۱۹۶	۰/۱۳۵۲	۲/۵۶	۳/۸
Persiangig.com	۲۴۷۳	۳۸	۲۳۹۴۱	۰/۰۵۹۹	۲/۷۲	۰/۷۱۵
Parsiblog.com	۲۵۲۵	۳۴	۱۶۵۴۹	۰/۰۶۵۶	۱/۹۳	۸/۲۹
Iranblog.com	۳۲۱۳	۵۵	۹۶۷۴	۰/۰۴۸۱	۲/۵۸	۱/۱
Rozblog.com	۴۴۶۱	۷۱	۸۳۰۷	۰/۰۳۴	۳/۰۶	۵
Parsfa.com	۵۵۸۲	۷۸	۱۷۰۵	۰/۰۳۳۷	۱/۷	۰/۸۱۷
Loxblog.com	۷۹۳۹	۱۲۳	۵۵۳۰	۰/۰۲۰۷	۲/۹	۴/۵۳

جدول (۳) که از جمع‌بندی اطلاعات مربوط به ۸۱ سرویس‌دهنده شناسایی شده به‌دست آمده است، مقایسه معدل ارقام ده سرویس‌دهنده اول با معدل ارقام کل وب‌نوشت‌های فارسی را نشان می‌دهد. با بررسی این آمار و ارقام می‌توان نتیجه گرفت که برای جمع‌آوری اطلاعات از وب‌نوشت‌ها، بررسی ۱۰ سرویس‌دهنده اول کافی است زیرا از لحاظ حجم، ۸۶ درصد تعداد صفحات موجود را پوشش داده و سایر نسبت‌ها نیز نشان می‌دهد که سایر سرویس‌دهنده‌ها تأثیر جدی در نتایج به‌دست آمده نخواهند داشت.

شکل (۱) نمودار مقایسه تعداد بازدیدها از ۵ سرویس دهنده اول وب‌نوشت‌های فارسی از ابتدای سال ۲۰۰۹ میلادی تا انتهای سال ۲۰۱۱ که با استفاده از سرویس‌های سایت گوگل‌ترندز [۵۴] استخراج شده است را نشان می‌دهد. این نتایج نیز آمار و ارقام جمع‌آوری شده قبلی در مورد سرویس‌دهندگان اول را تأیید می‌کند.

جدول ۳. مقایسه ده سرویس‌دهنده اول و کل

مقایسه ده سرویس دهنده اول با کل	رتبه جهانی	رتبه در ایران	اعتبار	برد	صفحه بر کاربر	تعداد صفحات (میلیون)
کلیه سرویس‌دهندگان	۱۰۷۶۲۳۲	۸۴۸۱	۵۱۴۴	۰/۰۲	۲/۹	۱۱۸
ده سرویس دهنده اول	۲۸۴۲	۴۲/۶	۳۸۳۸۵	۰/۱۴	۲/۸	۱۰۲

شکل ۱. مقایسه تعداد بازدیدها از ۵ سرویس‌دهنده اول

۶-۴. به‌روزرسانی وب‌نوشت‌ها

برای برآورد تجهیزات مورد نیاز معماری مورد نظر جهت خزش صفحات به‌روز شده، لازم است تعداد این صفحات مشخص شده تا بتوان بازه به‌روزرسانی وب‌نوشت‌ها برای واکشی صفحات در زمان مربوطه را تخمین زد. به‌همین منظور با استفاده از عملگر site: و سرویس بازه زمانی گوگل، آمار ۲۰ سرویس‌دهنده اول در تاریخ ۸ ژانویه ۲۰۱۲ استخراج شد [۵۵]. نتایج به‌دست آمده در بازه‌های یک ساعت، یک روز، یک هفته و یک ماه ثبت شده است. جدول (۴) نتایج این آزمون برای ۱۰ سرویس‌دهنده اول را نشان می‌دهد.

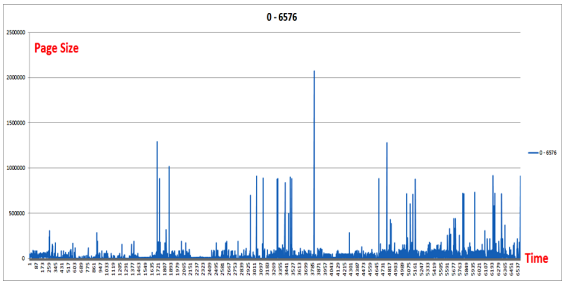
بر طبق جدول (۴)، مجموع میانگین صفحات به‌روز شده در یک روز برای ۱۰ وب‌نوشت اول ۴۷۴۵۸۶ صفحه است. اگر بخواهیم به صورت برخط از تغییرات به‌وجود آمده خبردار شویم، باید بتوانیم این حجم صفحه را به‌صورت روزانه واکشی نماییم. یعنی باید قادر باشیم در هر ثانیه فقط ۵/۵ صفحه را واکشی کنیم.

برای به‌دست آوردن پهنای باند مورد نیاز، لازم است که از حجم متوسط هر صفحه آگاهی حاصل شود. شکل (۲) نتایج آزمایش بررسی اندازه صفحات وب‌نوشت‌های فارسی را نشان می‌دهد. در این آزمایش فقط صفحات به‌روز شده ۴ سرویس‌دهنده اول واکشی شده است. با توجه به اینکه مجموع تعداد کل صفحات این

سرویس‌دهنده‌ها ۸۱/۴ میلیون صفحه بوده که ۶۹ درصد کل وب‌نوشت‌ها را شامل می‌شود، نتایج به‌دست آمده از آزمایش‌ها قابل قبول خواهد بود. این نمودار نشان می‌دهد که حجم صفحات دارای پراکندگی زیادی نبوده و حجم صفحاتِ خیلی زیاد و یا خیلی کم، کمتر اتفاق می‌افتد.

جدول ۴. میزان به‌روزرسانی وب‌نوشت‌های فارسی

سرویس دهنده	تعداد صفحات به‌روز شده در				
	یک ساعت	یک روز	یک هفته	یک ماه	میانگین (روزانه)
Blogfa.com	۱۳۳۰۰	۳۸۴۰۰۰	۲۵۰۰۰۰۰	۳۴۲۰۰۰۰	۲۹۳۵۸۶
Mihanblog.com	۷۶	۱۳۶۰۰۰	۸۷۱۰۰۰	۲۲۵۰۰۰۰	۸۴۳۱۳
Persianblog.ir	۳۰۷	۹۸۰۰۰	۴۷۴۰۰۰	۱۱۲۰۰۰۰	۵۲۶۰۴
Blogsky.com	۱	۱۳۵۰۰	۱۴۷۰۰۰	۴۰۷۰۰۰	۱۲۰۲۳
Persiangig.com	۰	۶۱	۹۸۳۰	۴۱۷۰۰	۷۱۴
Parsiblog.com	۳۰	۴۴	۱۲۳۰۰	۶۱۵۰۰	۱۱۴۳
Iranblog.com	۰	۴۱	۳۸۵	۱۷۰۰۰	۱۶۶
Rozblog.com	۲۶	۴۱۰۰۰	۲۹۶۴۷۰	۵۶۲۶۰۰	۲۵۶۸۳
Parsfa.com	۰	۱۶۸	۱۳۵۰۰	۱۱۴۰۰۰	۱۴۷۴
Loxblog.com	۷	۲۷۱	۳۴۹۰۰	۱۸۳۰۰۰	۲۸۸۱
مجموع	۱۳۷۴۷	۶۷۳۰۸۵	۴۳۵۹۳۸۵	۸۱۷۶۸۰۰	۴۷۴۵۷۶



شکل ۲. بررسی اندازه صفحات وب‌نوشت‌های فارسی

جدول (۵) نتایج کلی این آزمایش را نشان می‌دهد. با توجه به نتایج این جدول، می‌توان پهنای باند و همچنین فضای ذخیره‌سازی مورد نیاز را محاسبه کرد. طبق جدول (۵) حجم متوسط یک صفحه وب‌نوشت فارسی ۴۰۴۶۱ بایت محاسبه شده است. بنابراین برای واکشی ۵/۵ صفحه در ثانیه، به پهنای باند ۱/۸ Mbp s نیاز است. با توجه به جداول (۳ و ۴) فضای مورد نیاز روزانه برای ذخیره‌سازی نیز ۱۹/۲ گیگابایت برآورد می‌شود.

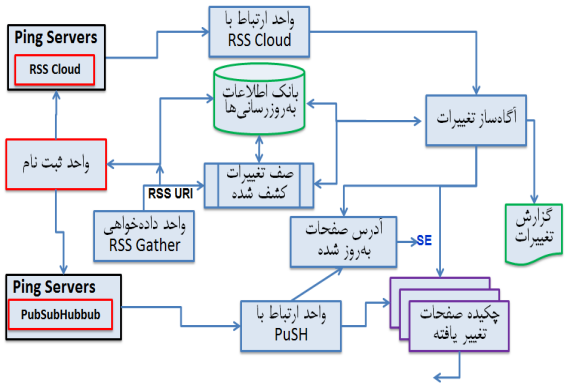
۷. معماری و پیاده‌سازی آگاه‌ساز، ارزیابی و تحلیل نتایج

بر اساس مطالعات و آزمایش‌های مختلفی که نتایج آن را در بخش‌های قبل ارائه شده، معماری سامانه ODCB مطابق شکل (۳) ارائه می‌شود.

جدول ۵. مشخصات وب‌نوشت‌های واکشی شده

بازه واکشی صفحات: از ساعت ۱۹:۱۹ روز ۲۰۱۲/۱/۲۶ تا ساعت ۷:۵۷ روز ۲۰۱۲/۱/۲۷	
تعداد صفحات واکشی شده	۶۵۷۵ صفحه
تعداد صفحات صفر بایتی	۳۲۰۷ صفحه
حجم صفحات واکشی شده	۱۳۶۲۷۳۵۶۶ بایت
میانگین حجم هر صفحه	۴۰۴۶۱ بایت
اندازه بیشینه صفحه	۲۰۷۱۴۷۹ بایت
اندازه کمینه صفحه	۷۰۴ بایت

در معماری ارائه شده ترکیبی از فناوری‌های مبتنی بر Ping و داده‌خواهی[1] در نظر گرفته شده است. واحدهای RSS Cloud و PuSH که مبتنی بر Ping و واحد RSS Gather که مبتنی بر داده‌خواهی هستند، وظیفه جمع‌آوری اطلاعات را دارا هستند. این اطلاعات در یک صف آرایش داده می‌شود که با استفاده از یک پایگاه داده ایجاد می‌شود. اطلاعات نگهداری شده در این پایگاه داده شامل آدرس‌ها، زمان بازدید از میزبان، بازه به‌روزرسانی هر میزبان و سایر اطلاعات مورد نیاز است. واحد ارتباطی RSS Cloud بر اساس فناوری RSS Cloud کار می‌کند. این فناوری گونه‌ای توسعه یافته بر پایه پروتکل RSS بوده که آگاه‌سازی بی‌درنگ را فراهم می‌آورد. همچنین این فناوری یک سرویس وب است که بر اساس HTTP- POST، XML-RPC یا SOAP [۵۶-۵۷] پیاده‌سازی می‌شود. این فناوری برای پردازه‌ها این امکان را فراهم می‌کند که در یک خوشه ابری ثبت نام کرده تا آنها را از به‌روز رسانی‌ها آگاه نمایند [۵۸].



شکل ۳. معماری سامانه ODCB

واحد PuSH که از فناوری PubSubHubbub پشتیبانی می‌کند، مسؤل خبرگیری از سرویس‌دهنده‌هایی است که از این فناوری استفاده می‌کنند. فناوری PuSH مشابه Cloud است با این تفاوت که محتوای تغییر یافته را نیز ارسال می‌کند. این پروتکل یک استاندارد کاملاً باز بوده که به‌صورت غیر متمرکز عمل می‌کند. هنگامی که مطلب جدیدی پست می‌شود، ناشر به یک مرکز اطلاع داده و این مرکز تغییرات را برای کلیه مشترکین ارسال می‌کند.

[1] Polling

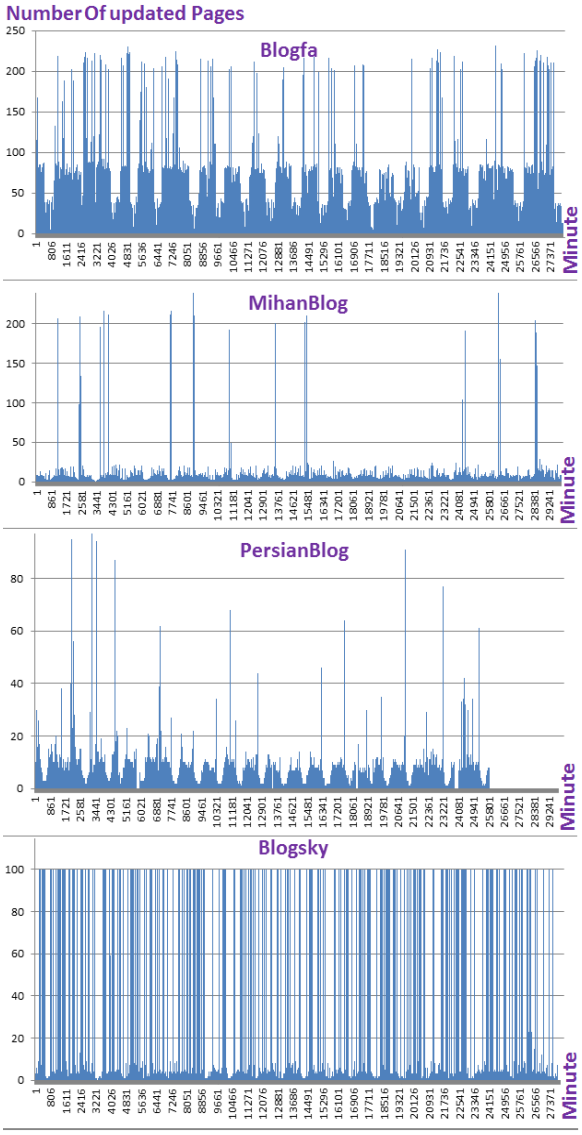
به‌همین جهت لازم نیست مشتریان جهت گرفتن اطلاعات جدید، مرکز را Ping کنند [۵۹]. مطابق معماری سامانه، پس از تعامل با ماشین‌های خدمتگذار Ping و ثبت نام در واحد ثبت نام واحدهای RSS Cloud و PuSH آماده استفاده از آگاه‌سازی‌ها می‌شوند.

چالش اصلی در مورد فناوری‌های یاد شده این است که این فناوری‌ها در صورتی کارا هستند که سرویس‌دهندگان وب‌نوشت‌ها از آنها پشتیبانی نمایند. در صورت عدم پشتیبانی وب‌نوشت‌ها از این فناوری‌ها، امکان استفاده از آنها میسر نخواهد بود. بسیاری از وب‌نوشت‌ها از این فناوری‌ها پشتیبانی نمی‌کنند. طبق بررسی‌هایی که از ۱۰ سرویس دهنده اول به‌عمل آمد، مشخص شد که تمامی ۱۰ وب‌نوشت اول فارسی از این فناوری‌ها پشتیبانی نمی‌کنند. بنابراین باید برای آن دسته از سرویس‌دهندگان وب‌نوشت که از این فناوری‌ها پشتیبانی نمی‌کنند، راه حل دیگری ارائه شود. واحدهای RSS Gather و RSS Notifier که بر اساس روش داده‌خواهی عمل می‌کنند، برای همین منظور در نظر گرفته شده‌اند.

برای پیاده‌سازی روش داده‌خواهی باید وضعیت به‌روزرسانی وب‌نوشت‌ها مورد بررسی قرار بگیرد. طبق نتایج آزمایش‌های انجام شده در بخش قبل، زمان لازم برای واکشی صفحات به‌روز شده وب‌نوشت‌های فارسی با پهنای باند ۱/۸ Mbps، یک روز می‌باشد که برای آگاه‌سازی برخط نامناسب بوده و قابل قبول نیست.

برای مقابله با این مشکل، در نگاه اولیه می‌توان افزایش پهنای باند را به‌عنوان یک راه حل مد نظر قرار داد. در این صورت اگر بخواهیم زمان آگاه‌سازی را از یک روز به ۱۰ دقیقه برسانیم، به پهنای باندی در حدود 260 Mbps نیاز است. فراهم کردن این مقدار پهنای باند فقط به‌دلیل آگاهی از تغییرات وب‌نوشت‌ها به‌صرفه نبوده و به‌لحاظ منابع مالی به‌سختی امکان‌پذیر است. یک راه حل مناسب این است که حجم مورد نیاز برای دانلود کاهش داده شود. خزش RSSها راه‌حلی است که در تحقیق حاضر ارائه شده است. با توجه به حجم کم RSSها، زمان واکشی آنها می‌تواند بسیار کم‌تر از واکشی کل اطلاعات باشد، به‌طوری که آگاهی برخط نسبت به تغییرات ممکن شود. برای به‌دست آوردن زمان و پهنای باند مورد نیاز معماری ارائه شده، آزمون دیگری طراحی و پیاده‌سازی شده است. در این آزمون با استفاده از معماری پیشنهادی در یک بازه ۲۷ روزه از تاریخ ۱۷ می تا ۱۳ ژوئن سال ۲۰۱۲ تغییرات وب‌نوشت‌های مربوط به ۴ سرویس‌دهنده اول وب‌نوشت‌ها در جدول (۲) که ۷۰ درصد حجم وب‌نوشت‌ها را شامل می‌شوند، واکشی شده است. در این آزمایش از یک ماشین دو پردازنده چهار هسته‌ای با فرکانس ۲/۵ GHz که دارای ۱۶ نخ و پهنای باند ۲ Mbps استفاده شده است. نمودارهای این آزمایش در شکل (۴) و نتایج آن در جدول (۶) نشان داده شده است.

برنامه نوشته شده برای واکشی تغییرات به‌صورت موازی بر روی هر ۴ سرویس دهنده عمل می‌کند. شیوه اجرای برنامه به این صورت است که برنامه به سرویس دهنده مربوطه متصل شده و اطلاعات مربوط به تغییرات را از طریق RSS واکشی کرده و پس از یک دقیقه توقف، به‌طور مجدد به همان سرویس‌دهنده متصل شده و عمل واکشی را تکرار می‌کند.



شکل ۴. نمودار به‌روزشدگی ۴ وب‌نوشت اول در ۲۷ روز

جدول ۶. اطلاعات سرکشی به ۴ سرویس‌دهنده اول

	blogfa	persianblog	mihanblog	blogsky	جمع
تعداد بازدیدها	۲۸۱۵۷	۲۵۸۲۳	۳۰۰۸۲	۲۷۷۹۵	۱۱۱۸۵۷
بیشینه صفحات تغییر یافته هر بازدید	۲۳۱	۱۱۰	۲۴۷	۱۰۰	۲۴۷
کل صفحات تغییر یافته	۸۱۴۴۵۷	۷۷۶۵۰	۷۶۰۸۸	۵۱۸۰	۹۷۳۳۷۵
تعداد صفحات تغییر یافته در هر بازدید	۲۹	۳	۳	۳	۳۸

مطابق جدول (۶)، در مدت ۲۷ روز به‌طور میانگین، به تعداد ۲۷۹۶۴ بار به هر سرویس دهنده سرکشی کرده و اطلاعات تغییریافته آنها دانلود شده است. یعنی برای هر واکشی به‌طور متوسط ۸۳ ثانیه وقت صرف شده است. با توجه به اینکه پس از هر واکشی یک دقیقه توقف کرده و دوباره به آن سایت متصل شده، می‌توان از این عدد ۶۰ ثانیه کم کرد و به عدد ۲۳ ثانیه رسید. به‌عبارت دیگر می‌توان در مدت زمان ۲۳ ثانیه تغییرات این وب‌نوشت‌ها را (که ۷۰ درصد وب‌نوشت‌های فارسی را پوشش می‌دهند)، و در مدت ۳۳ ثانیه تغییرات همه وب‌نوشت‌های فارسی را واکشی نمود. این یک رویکرد برای بررسی برخط وب‌نوشت‌ها و مانیتور کردن آنها است. در ادامه عملکرد این معماری مورد بررسی بیشتر قرار می‌گیرد.

مطابق معماری ارائه شده در این جدول، پس از واکشی تعدادی RSS توسط واحد RssGather، واحد RssNotifier هر RSS واکشی شده را بررسی و آدرس‌های وب‌نوشت‌هایی را که تغییرات داشته‌اند استخراج و در یک جدول ذخیره می‌نماید. برخی از سرویس‌دهندگان، صفحه‌ای دارند که در آن آخرین وب نوشت‌هایی به‌روز شده درج می‌شوند. با پردازش این صفحه، می‌توان آدرس این وب‌نوشت‌ها را به‌دست آورد. بنابراین با بررسی دوره‌ای این صفحه در بازه‌های زمانی معین می‌توان از تغییرات به‌عمل آمده در وب‌نوشت‌ها، در زمانی کوتاه آگاهی یافت. بازه زمانی مذکور نسبت به سرویس‌دهندگان مختلف، متفاوت خواهد بود. به‌عنوان مثال، بازه بررسی صفحه مذکور در سرویس‌دهنده blogfa از رابطه (۱) محاسبه شده است:

$$\text{StepVisit} = \begin{cases} \text{StepVisit}/2, & UpdateFlag = 1 \\ \text{StepVisit} + 120, & UpdateFlag = 0 \end{cases} \quad (۱)$$

در رابطه فوق StepVisit بازه بررسی بر حسب ثانیه و UpdateFlag به‌روز بودن صفحه هنگام مراجعه را نشان می‌دهد.

۸. نتیجه‌گیری

استقبال کاربران از وب‌نوشت‌ها به عنوان یکی از پرکاربردترین رسانه‌های غیر رسمی بسیار بیشتر و متنوع‌تر از رسانه‌های رسمی است. آمار و ارقام نشان می‌دهد که محبوبیت این نوع رسانه در ایران بسیار زیاد است. به همین سبب این رسانه‌ها در فضای سایبر هم نوعی فرصت و هم نوعی تهدید محسوب می‌شوند. اگر آمادگی بهره‌گیری از چنین فرصت‌هایی را داشته باشیم، می‌توانیم به خوبی از این منظر نفع برده و از آن به عنوان نقاط قوت بهره‌جویی نماییم. در این مقاله وب‌نوشت‌های فارسی به لحاظ تعدد/ نرخ بازدید، تعداد صفحات، اندازه صفحات، رتبه و اعتبار، میزان به‌روزرسانی آنها مورد آزمایش‌های مختلف قرار گرفته و نتایج آن بررسی و مقایسه شده‌اند. همچنین معماری پیشنهادی سامانه جمع‌آوری تغییرات وب‌نوشت‌ها که به صورت برخط عمل می‌کند، ارائه شده و پیاده‌سازی و بهینه‌سازی آن نیز ارائه شده است. مهم‌ترین قسمت‌های این سامانه، واحدهای آگاه‌ساز تغییرات و داده‌خواهی می‌باشد. یک سازوکار مشخص برای آگاهی از تغییرات وب‌نوشت‌هایی که از یک الگوی ثابت استفاده می‌کنند، طراحی شده و برای سرویس‌دهندگانی که دارای الگوی خاص خود هستند، سازوکار خاص آن سرویس‌دهنده را طراحی و پیاده‌سازی کرده‌ایم. در صورت تغییر در الگوی به‌روزرسانی، یک مکانیزم اطلاع‌رسانی برای آگاهی از تغییرات به وجود آمده در سرویس دهنده منظور شده است. در این مقاله ما نشان دادیم که با توجه به معماری ارائه شده با منابع محدود در حد یک سرور با دو پردازنده چهار هسته‌ای و پهنای باند ۲ Mbps می‌توان به صورت برخط از تغییرات به وجود آمده در وب‌نوشت‌های فارسی مطلع شد.

۹. مراجع

[1] Greenwood, M. "Prioritising Hyperlinks for Topic-Focused Web Crawling using Lexical and Terminological Profiling"; M.A. Thesis, Univ. of Manchester, 2009.

[2] "Whole-Product-Dynamic Test"; Technical Report, AV Comparatives and the Univ. of Innsbruck's Faculty of Computer Science and Quality., 2011.

[3] Lee, H. T.; Leonard, D.; Wang, X.; Logulnov, D. "IRLbot: Scaling to 6 Billion Pages and Beyond"; In Proc. of the 17th Int. WWW Conf. Beijing 2008, 427-436.

[4] 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 Proc. of the 18th Int. Conf. on String Processing and Information Retrieval, Pisa, Italy, 2011, 237-248.

[5] Lewandowski, D. "A Three-Year Study on the Freshness of Web Search Engine Databases"; J. Inform. Sci. 2008, 34, 817–831.

[6] Agarwal, N.; Kumar, S.; Liu, H.; Woodward, M. "BlogTrackers: a Tool for Sociologists to Track and Analyze Blogosphere"; In Proc. of the Third Int. ICWSM Conf. 2009, 359-360.

[7] Sayyadi, H.; Hurst, M.; Maykov, A. "Event Detection and Tracking in Social Streams"; In Proc. of Int. Conf. on Weblogs and Social Media (ICWSM) 2009.

[8] Pathak, M.; Thakre, V. "Intelligent Web Monitoring- A Hypertext Mining-Based Approach"; J. Indian Institute Sci. 2006, 86, 481-492.

[9] Nanno, T.; Suzuki, Y.; Fujiki, T.; Okumura, M. "Automatic Collection and Monitoring of Japanese Weblogs"; In Proc. of Int. World Wide Web Conf. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics 2004.

[10] Tabansky, L. "Basic Concepts in Cyber Warfare"; J. of Military and Strategic Affairs 2011, 3, 75-92.

[11] Swanstrom, E. "Wax Blocks, Data Banks, and File #0467839: the Archive of Memory in William Gibson's Science Fiction"; J. Educ. Inform. Stud. 2005, 1, 1-24.

[12] The U.S. Dep't of Defense. "Department of Defense Strategy for Operating in Cyberspace"; July 2011.

[13] Bakliwal, A.; Arora, P.; Varma, V. "Entity Centric Opinion Mining from Blogs"; In Proc. of 24th Int. Conf. on Computational Linguistics, 2012, 53-64.

[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. "Web Logs and Online Discussions as Tools to Promote Reflective Practice"; J. Interac. Online Learn. 2004, 3, 1-16.

[16] Mingjun, X.; Hanxiang, W.; Weimin, L.; Zhihua, N. "A Public Opinion Classification Algorithm Based on Micro-Blog Text Sentiment Intensity : Design and Implementation"; J. Comput. Network Inform. Secur. 2011, 3, 48-54.

[17] Oh, A.; Lee, H.; Kim, Y. "User Evaluation of a System for Classifying and Displaying Political Viewpoints of Weblogs"; in Proc. of the Third Int. ICWSM Conf. 2009, 282-285.

[18] Nichols, J.; Mahmud, J.; Drews, C. "Summarizing Sporting Events Using Twitter"; In Proc. of the ACM Int. Conf. on Intelligent User Interfaces (IUI), New York 2012, 189–198.

[19] Mithun, S. "Exploiting Rhetorical Relations in Blog Summarization"; Ph.D. thesis, Dep't of Computer Science and Software Eng. Concordia . Montreal, Canada, 2012.

[20] Wortmann, P. "Topic-Based Blog Article Search for Trend Detection"; Project Thesis, Technical Univ. of Kaiserslautern, 2009.

[21] Petrovic, S.; Osborne, M.; Lavrenko, V. "Streaming First Story Detection With Application to Twitter"; In Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10), Stroudsburg 2010, 181-189.

[22] Benhardus, J. "Streaming Trend Detection in Twitter"; in National Science Foundation REU for Artificial Intelligence, Natural Language Proc. and Information Retrieval, Univ. of Colarado, 2010.

[23] Shih, C.; Peng, T. "Building Topic/Trend Detection System Based on Slow Intelligence"; in Proc. of the 16th Int. Conf. on Distributed Multimedia Systems, DMS, Illinois, USA 2010, 53-56.

[24] Weng, J.; Lee, B. S., "Event Detection in Twitter"; In Proc. of the 5th Int. AAAI Conf. on Weblogs and Social Media 2011, 401–408.

[25] Mathioudakis, M.; Koudas, N. "Twittermonitor: Trend Detection Over the Twitter Stream"; in Proc. of the ACM SIGMOD Int. Conf. on Management of Data, New York, 2010, 1155–1158.

[26] Kim, D.; Ki, D.; Rho, S.; Hwang, E. "Detecting Trend and Bursty Keywords Using Characteristics of Twitter Stream Data"; Int. J. of Smart Home 2013, 7, 209-219.

[27] Fang, F.; Pervin, N.; Datta, A.; VanderMeer, D. "Detecting Twitter Trends in Real-Time"; In Proc. of the 21st Workshop on Information Tech. and Systems (WITS) 2011.

[28] Vakali, A.; Giatsoglou, M.; Antaris, S. "Social Networking Trends and Dynamics Detection Via a Cloud-Based Framework Design"; in Proc. of the 21st Int. Conf. Companion on World Wide Web, New York 2012, 1213–1220.

[29] 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, 2010, 248–261.

[30] Berry, R. "Blog 101: an Overview of Weblog Technologies"; in Proc. of the Tools and Tech. Section (STC) 2004, 216-220.

[31] Baloglu, A.; Aktas. M. "Blog Miner: Web Blog Mining Application for Classification of Movie Reviews"; in Proc. of Fifth Int. Conf. on Internet and Web Applications and Services, IEEE Computer Society Transaction 2010, 77-84.

[32] Goncalves, M.; Almeida, J.; Santos, L.; Laender, A.; Almeida, V. "On Popularity in the Blogosphere"; In Social Computing Transcation, Published by the IEEE Computer Society, 2010, 42-49.

[33] Gao, W.; Tian, Y.; Huang, T. "Vlogging: A Survey of Videoblogging Technology on the Web", J. ACM Comput. Surv. 2010, 42, 1-57.

[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, 2006, 3-12.

[35] Ye, S.; Lang, J.; Wu, F. "Crawling Online Social Graphs"; in Proc. of the 2010 Asia Pacific Web Conf. 2010, 236-242.

[36] Najork, M.; Wiener, J. L. "Breadth-First Search Crawling Yields High-Quality Pages"; In Proc. of the World Wide Web Conf. (WWW'01) 2001, 114–118.

[37] Lang, J. "Encouraging User Engagement with Online Social Networks"; Ph.D. Thesis, the Univ.of California, 2011.

[38] Ye, S. "Online Social Network Measurements and Search Privacy Protection"; Ph.D. Thesis, The Univ. of California, 2010.

[39] Gill, A.; Nowson, S.; Oberlander, J. "What Are They Blogging About? Personality, Topic and Motivation in Blogs"; In Proc. of the Third Int. AAAI Conf. on Weblogs and Social Media 2009, 18-25.

[40] Buckley, C. "Implementation of the SMART information Retrieval System"; TR85-686, Computer Science Dep't. Cornell Univ. 1985.

[41] Hurst, M.; Maykov, A. "Social Streams Blog Crawler"; In Proc. of the 2009 IEEE Int. Conf. on Data Engi. 2009, 1615–1618.

[42] 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.

[43] Gyongyi, Z.; Garcia-Molina, H.; Pedersen, J. "Web Content Categorization Using Link Iinformation"; Technical Report, Stanford University, 2006.

[44] 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 Weblogs: Papers from the 2006 Spring Symposium, AAAI Press. Technical Report, 2006, 184–186.

[45] Ranganathan, S. R., "Library Classification on the March"; In the Sayers Memorial Volume, Library Association, London, 1961, 84.

[46] Yu, N. "Semi-Supervised Learning for Identifying Opinions in Web Content"; Ph.D. Thesis, Indiana Univ. 2011.

[47] Pandey, R.; Dwivedi, S. "Interoperability between Semantic Web Layers: A Communicating Agent Approach"; Int. J. Comput. Appl. 2010, 12, 28-32.

[48] Internet World Stats, "The Internet Big Picture, World Internet Users and Population Stats";http://internetworldstats.com/stats.htm, June 30, 2010; Last Accessed: Feb. 28, 2011.

[49] Mina, N. "Blogs, Cyber-Literature and Virtual Culture in Iran"; George C. Marshall, European Center for Security Studies, 2007, 15, 6.

[50] Kargar, M.; Ramli, A.; Ibrahim, H.; Azimzadeh, F. "Formulating Priority of Information Quality Criteria on the Blog"; World Appl. Sci. J. 2008, 4, 586-593.

[51] Zhou, P.; Lei, J.; Ye, W. "Large-Scale Data Sets Clustering Based on Map Reduce and Hadoop"; J. Comput. Inform. Sys. 2011, 7, 5956-5963.

[52] Alexa Co. "Statistics Summary"; http://.alexa.com; Last Accessed: Dec. 21, 2011.

[53] Google Co. "Google Search"; http://google.com; Last Accessed: Jan, 1, 2012.

[54] Google Co. "Google Trends."; http://trends.google.com; Last Accessed: Jan, 2, 2012.

[55] Google Co. "Google Search"; http://google.com; Last Accessed: Jan, 8, 2012.

[56] Allan, R. J.; Ashworth, M. "A Survey of Distributed Computing, Computational Grid, Meta-Computing and Network Information Tools"; Computational Science and Eng. Dep't, CCLRC Daresbury Laboratory, Daresbury, Warrington WA4 4AD, UK, 2001, 38-42.

[57] Ploscar, A. "XML-RPC vs. SOAP vs. REST Web Services in Java – Uniform Using Wswrapper"; Int. J. Comput. 2012, 4, 215-223.

[58] Richards, R. "Content Syndication: RSS and Atom"; Pro PHP XML and Web Services Book, Apress, 2006, 521-566.

[59] Roden,T. "Realtime Syndication"; In Building the Realtime User Experience Book, O'Reilly Media Inc., 2010, 9-36.