دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران
دانشگاه شهید بهشتی، دانشکده مهندسی برق و کامپیوتر ،تهران، ایران، ۱ تا ۳ اسفند ۱۳۸۵

منش‌نمایی وب فارسی با استفاده از پیمایشگر وب صبا

علیرضا رضوانیان[*]، حسن بشیری[†]

چکیده

منش‌نمایی وب شاخصی است که براساس آن می‌توان به بررسی دقیق وب معنایی، وب محتوایی، ارائه طرح‌های پیشنهادی در تراکم کاری در تجارت الکترونیکی و بهینه‌سازی سرویس‌دهنده‌های وب و اینترنت پرداخت. به طور معمول از منش‌نمایی وب برای مطالعه زبان در صفحات وب و پیشرفت و نحوه استفاده آن در بین کاربران وب استفاده می‌شود. موضوع منش‌نمایی در چندین زبان خاص مورد مطالعه قرار گرفته است. اما در زبان فارسی تا کنون مطالعه‌ای انجام نشده است و صرفاً محدود به آمار گزارشات یا مصاحباتی پراکنده و غیرعلمی از وبلاگ‌های فارسی بوده است.

در این مقاله با استفاده از یک طراحی جدید از پیمایشگر وب با عنوان پیمایشگر وب صبا به بررسی منش‌نمایی وب فارسی می‌پردازیم.

واژه‌های کلیدی

بازیابی اطلاعات وب، پیمایش وب، پیمایشگر وب صبا، وب معنایی، منش‌نمایی وب فارسی

Persian Web Characterization Based on SABA Web Crawler

Alireza Rezvanian, Hassan Bashiri

Abstract

Web characterization is an indication which is used in analyzing the semantic web, web content and presenting plans in ecommerce workloads and optimizing the web and Internet services. Studying the behavior of specific language in web documents, spreading and using of the language among users are another popular usage of web characterization. Although web characterization is studied in some languages, there is no documented study of web characterization in Persian language. Documents have been limited to some informal reports and interviews of Persian web logs.

Studying and analyzing the Persian web characterization is an interesting topic we are going to discuss in details which is done with a new approach of designed web crawler entitled SABA.

Keywords

Information Retrieval, Web Crawling, SABA Crawler, Semantic Web, Persian Web Characterization

[*] دانشجوی کارشناسی مهندسی نرم‌افزار، گروه مهندسی کامپیوتر دانشگاه بوعلی‌سینا، rezvan@basu.ac.ir
[†] مدرس گروه مهندسی کامپیوتر دانشگاه بوعلی‌سینا، دانشکده فنی و مهندسی، bashiri@basu.ac.ir

۱- مقدمه

۱-۱- گسترگی اطلاعات

عوامل کلیدی در موفقیت و رشد تارنمای گسترده جهانی به حجم انبوه و بزرگی آن و عدم کنترل متمرکز بر محتوای آن دانسته می‌شود. البته از طرفی این دو موضوع از مهمترین مشکلات تشخیص اطلاعات در وب نیز محسوب می‌شود. علاوه بر آن توزیع کیفی نامتوازن از صفحات از طرفی و پراکندگی کامپیوترهای با تنوعی از ساختار و محتوا، پژوهشگران مختلف را با مشکل مواجه نموده است. پیمایش وب فرآیندی است که توسط موتورجستجو برای جمع‌آوری صفحات وب مورد استفاده قرار می‌گیرد.

تعداد صفحات وب از ۱۶،۵ میلیارد در سال ۲۰۰۳ به مقدار بیش از ۹۴ میلیارد در سال ۲۰۰۶ افزایش یافته است[۱] و آمار جمعیتی برای کاربران از ۶۰۰ میلیون نفر در سال ۲۰۰۲ به بیش از ۱،۰۹۳ میلیارد نفر در ابتدای سال ۲۰۰۷ در سراسر جهان رسیده است[۲]. نکته قابل توجه آن است که این رشد به صورت روزانه درحال افزایش است. همچنین بررسی مطالعات حاکی از دوبرابر شدن صفحات در هر ۹-۱۲ ماه و اضافه شدن روزانه ۳ میلیون صفحه است [۴،۳]. به طور متوسط موتورهای جستجو ۱۳٪ از ترافیک سایت‌های وب را تولید می‌کنند[۷]. علاوه بر این ۴۰٪ از کاربران ورودی سایت‌های وب برای اولین پیوندهای موجود در فهرست نتایج موتورجستجو را دنبال می‌کنند[۸].

"معتبرترین موتورهای جستجوی امروزی چون گوگل و التاویستا نیز تنها حجم محدودی از وب را پوشش داده و حتی حجم زیادی از اطلاعات آنها در ماه‌هایی از سال به روز نیست"[۵]. در فوریه ۲۰۰۴ گوگل از آمار ۴،۲۸ میلیارد صفحه شاخص‌گذاری شده خبر داد (که ۲۶٪ از وب به صورت وب پنهان[۱] و یا صفحات پویا مورد اهمال قرار گرفته است). پوشش صفحات مناسب و تازه به صورت شاخص‌گذاری شده به طور همزمان برای پاسخ‌دهی به نیاز کاربران تقریباً غیرممکن است. مشکل عمده موتورهای جستجو، دربخش مخازن اطلاعاتی خود نیاز به رسیدگی به اندازه و نرخ تغییرات وب است[۶].

اگرچه تخمین‌های متفاوتی برای اندازه وب زده شده است. اما همه آنها بر روی این واقعیت توافق دارند که مرتبه صفحات از چندین میلیارد گذشته است. پس در این اطلاعات هنگفت، روش‌های نوین بازیابی اطلاعات و مکاشفه‌هایی برای تسریع جستجو ضروری و غیرقابل اجتناب می‌باشد.

۱-۲- راه‌حل

با توجه به این میزان حجم بالای اطلاعات موجود بر روی وب، شاخص‌گذاری کل اطلاعات در زمان محدود و با توجه به امکانات سخت‌افزاری و شبکه‌ای موجود غیرممکن به نظر می‌رسد.

بازیابی گزینشی اطلاعات می‌تواند راه‌حل مناسبی برای بهبود در کیفیت اطلاعات جمع‌آوری و شاخص‌گذاری شده باشد. معیارهای گزینش و چگونگی استخراج آنها در این راه‌حل می‌تواند براساس میزان نرخ به‌هنگام‌سازی صفحات، تشخیص علاقه مناطق مختلف جغرافیایی به موضوع، طبقه‌بندی موضوعی، زمانی، کمی و یا برپایه تکنیک‌های پیمایشگرهای متمرکز باشد[۹].

راه‌حلی که می‌توان برای این مشکل درنظر گرفت بازیابی گزینشی اطلاعات بوده و این ایده باعث به وجود آمدن پیمایشگرهای متمرکز و موضوعی گردید. در بخش بعدی به این موضوع می‌پردازیم.

۲- پیمایشگرهای متمرکز

پیمایشگرهای وب برنامه‌هایی هستند که با پیمایش صفحه به صفحه ساختاری به شکل گراف از وب ایجاد می‌کنند. ایده پیمایشگر متمرکز اولین بار در [۱۰] مطرح گردید. این‌گونه پیمایشگرها پیوندهای موجود در صفحات بازیابی شده را براساس *موضوع* و معیارهای ارزش‌یابی تعریف شده انتخاب می‌نمایند و براساس آن اولویت‌بندی (وزن‌دهی) کرده و پیوندهای با اولویت بالاتر را زودتر پیمایش می‌کنند.

در پژوهش حاضر با استفاده از ایده پیمایشگرهای متمرکز، پیمایشگری به نام «صبا» برای پیمایش سایت‌های فارسی و منش-نمایی این مجموعه، طراحی و پیاده‌سازی کرده‌ایم که در بند ۲-۱ به معماری پیمایشگر صبا و در بند ۴ به چگونگی استفاده و آزمایش صبا در منش نمایی فارسی می‌پردازیم.

۲-۱- پیمایشگر وب صبا

نام «صبا» را برای پیمایشگر طراحی شده انتخاب کرده‌ایم، چنانچه باد **صبا** در ادبیات پارسی همواره نماد خبررسانی از یار، معشوق، دوست و مقصود بوده است. در اینجا نیز در سطح انتزاعی دیگری پیمایشگر به جمع‌آوری و خبر رسانی از اطلاعات و اسناد وب می‌پردازد.

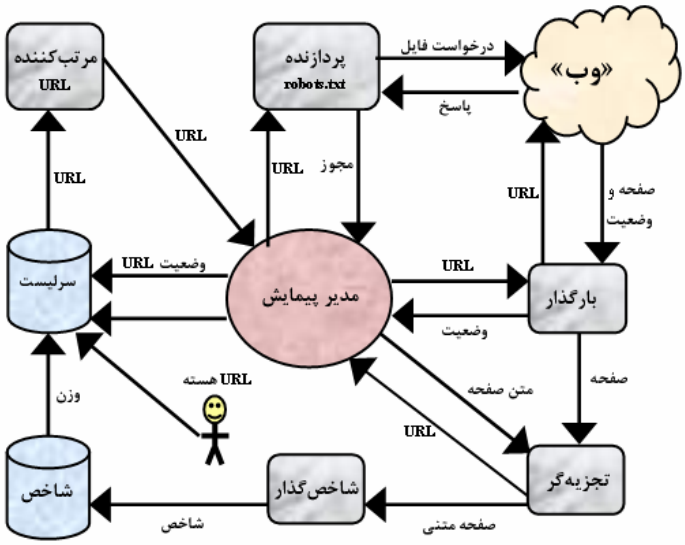
در معماری درنظرگرفته شده برای این پیمایشگر بخش‌های مختلف به شرح ذیل می‌باشد [۱۱] که معماری صبا نیز در شکل (۱) آورده شده است:

- **مدیر پیمایش**: مدیر پیمایش وظیفه هماهنگی و ارسال داده‌ای بین کلیه مولفه‌های موجود در پیمایشگر را داشته و نظارت بر افزودن URLهای جدید بازدیدنشده به داخل سرلیست را دارد.
- **مرتب‌کننده URL**: مولفه مرتب‌کننده URL براساس وزن شاخص و موضوع موردنظر به اولویت‌گذاری مابین URLها می‌پردازد.
- **پردازنده robots.txt**: براساس قانون منع ربات‌ها لازم است قبل از بارگذاری کل سایت پیمایشگر فایل

robots.txt را بارگذاری و از قسمت‌های مجاز و غیرمجاز مطلع گردد[۱۲].

- **سرلیست**: مولفه سرلیست به عنوان پایگاه داده پیوندها نگهداری URLها را به عهده دارد.
- **بارگذار**[۲]: مولفه بارگذار عمل بارگذاری صفحات و فایل‌های موردنظر را به درخواست مدیر پیمایش انجام می‌دهد.
- **تجزیه‌گر**: مولفه تجزیه‌گر وظیفه تجزیه فایل HTML و استخراج URL را داشته و URL ها را برای افزودن به پایگاه داده (سرلیست) به مدیر پیمایش فرستاده و متن خالی از برچسب‌های HTML را برای شاخص‌گذاری به مولفه شاخص‌گذار می‌فرستد.
- **شاخص**[۳]: مولفه شاخص به عنوان پایگاه داده شاخص از مشخصه واژه‌ها و عبارات نگهداری می‌کند.
- **شاخص‌گذار**[۴]: مولفه شاخص‌گذار وظیفه شاخص‌گذاری و وزن‌دهی به متن ارسالی را داشته و بردار شاخص از متن مورد نظر را به پایگاه داده شاخص ارسال می‌کند.

البته در بیشتر پیمایشگرهای طراحی شده مولفه شاخص‌گذار به عنوان مولفه‌ای از موتورجستجو در نظر گرفته می‌شود و در پیمایشگر درنظر گرفته نمی‌شود. اما از آن جهت که در این پروژه به منظور استخراج معیارهای مناسب برای منش‌نمایی وب فارسی به بردار صفحات واکشی شده نیاز است، شاخص‌گذار به عنوان مولفه‌ای در کاوشگر در نظر گرفته شده است.



شکل(۱): معماری پیمایشگر وب صبا

۳- منش‌نمایی وب

منش‌نمایی وب موضوعی است که از آن برای بررسی دقیق وب معنایی[۵]، وب محتوایی [۲۳]، طرح‌های پیشنهادی تراکم‌کاری در تجارت الکترونیکی و بهینگی سرویس‌دهنده‌ها[۲۴] در وب استفاده می‌گردد. به طور معمول از منش‌نمایی وب برای مطالعه و شناسایی زبان در صفحات وب و بررسی پیشرفت و رفتار وب در بین گرایشات کاربران استفاده می‌شود. پیمایشگر وب یکی از بهترین ابزارها برای مطالعه بر روی منش‌نمایی می‌باشد.

کشورهایی مانند برزیل[۱۳]، شیلی [۱۴]، پرتغال [۱۵]، اسپانیا [۱۶]، مجارستان [۱۷] و استرالیا [۱۸] کشورهایی هستند که زبان آنها از نظر موضوع منش‌نمایی وب مورد مطالعه قرار گرفته است. اما زبان فارسی با تقریب ۷۱ میلیون نفر گوینده در بین کشورهای ایران، افغانستان و آسیای میانه که در سطح جهان از نظر میزان جمعیتی که در رتبه ۱۵م قرار دارد[۱۹]، متاسفانه هنوز مطالعات و تحقیقاتی مدون و علمی بر روی بررسی منش‌نمایی آن صورت نگرفته است.

۳-۱- شیوه نمونه‌برداری

یکی از عمده‌ترین مشکلات پیچیده در کوشش برای منش‌نمایی وب، چگونگی بدست آوردن نمونه‌های خوب است. صفحات مهم خیلی کمی در میان انبوه صفحات غیرمهم به صورت گمشده وجود دارد که این اهمیت براساس بعضی از معیارها مانند رتبه‌بندی صفحه، شمارش ارجاع، اندازه صفحه و نمونه‌هایی از این دست مشخص می‌شود. فقط گرفتن URL به صورت تصادفی کافی نیست. در اکثر برنامه‌ها بایستی از گرفتن محتوای صفحات کم‌معنی یا بی‌معنی جلوگیری شود [۲۰].

دو راه‌کار عمده برای نمونه‌برداری صفحات وب وجود دارد:

- **نمونه‌برداری عمودی**: شامل جمع‌آوری صفحاتی که توسط نام‌های دامنه محدود می‌شوند. مانند زمانیکه صرفاً از دامنه .ir یا .ac.ir استفاده شود.
- **نمونه‌برداری افقی**: در این حالت معیار گزینش بر مبنای نام‌های دامنه نیست، بلکه یا از تراکنش‌های ثبت وقایع در پروکسی یک سازمان بزرگ یا یک سرویس‌دهنده خدمات اینترنت[۶] استفاده می‌شود یا آنکه از یک پیمایشگر وب استفاده می‌گردد. در روش اول اگرچه با استفاده از پروکسی به راحتی می‌توان صفحات مورد علاقه کاربران را یافت، اما تناوب بازدید مجدد به علت وابستگی آن به کاربر جهت کنترل غیرممکن است. درحالی‌که استفاده از پیمایشگر وب برای یافتن صفحات مورد علاقه کاربران قابل تخمین است درعین‌حال تناوب بازدید می‌تواند موازنه دقیقی داشته باشد.

۳-۲- برآورد تازگی و کهنگی

وب به طور ذاتی خیلی پویا است و پیمایش حتی قسمتی از وب زمان زیادی از جمله هفته‌ها و ماه‌ها را نیاز دارد. برطبق آمار بیش از ۵۰۳٪ از پیوندهای ارائه شده در موتورهای جستجو حاوی صفحات حذف شده است[۶]. تکنیک پیمایشگرهای متمرکز به روی بهبود تازگی و کهنگی صفحات در موتورهای جستجو متمرکز شده و عمدتاً ترکیبی از تازگی و کهنگی را برای توابع ارزشیابی استفاده می‌کند [۳۲].

تازگی صفحه p در زمان t با احتمال $u_p(t)$ از یک مدل توزیع پواسن تبعیت می‌نماید که به صورت رابطه (۱) نوشته می‌شود [۲۱]:

$$u_p(t) \propto e^{-\lambda_p t} \qquad (۱)$$

در این رابطه λ_p نرخ تغییرات صفحه p است و با استفاده از رابطه (۲) بدست می‌آید:

$$\lambda_p \approx \frac{\left(X_p - 1\right) - \frac{X_p}{N_p \log\left(1 - X_p / Np\right)}}{S_p T_p} \qquad (۲)$$

در این رابطه:

- N_p: تعداد بازدیدها از صفحه p.
- S_p: مقدار زمان از اولین بازدید صفحه p.
- X_p: تعداد دفعاتی که سرویس‌دهنده از تغییر صفحه p آگاه شده است.
- T_p: کل زمان بدون تغییر، مطابق با سرویس‌دهنده، مجموع کلیه بازدیدها از صفحه p.

حال چنانچه سرویس‌دهنده وب زمان آخرین تغییرات را در اختیار پیمایشگر نگذارد، می‌توان از بررسی مقایسه دو حالت بارگذاری شده برای بدست آوردن این معیار استفاده نمود. بنابراین در این حالت X_p تعداد زمان‌های تغییر یافته است. رابطه ۳ تخمینی از پارامتر X_p است:

$$\lambda_p = \frac{-N \log\left(1 - X_p / N_p\right)}{S_p} \qquad (۳)$$

در معادله بالا لازم است که $N_p > X_p$ باشد. بنابراین چنانچه تغییرات صفحه در زمان بازدید باشد، نمی‌توان رخداد تغییر را تخمین زد [۱۱].

۴- منش‌نمایی وب فارسی

۴-۱- مشخصات آزمایش

با استفاده از پیمایشگر صبا [۱۱] که طراحی آن در شکل (۱) توضیح داده شد، صفحاتی از ۲۳۱ سایت فارسی که پیمایش گردید. مدت زمان مورد آزمایش برای مجموعه ۲۳۱ سایت فارسی ۲۴ روز به طول انجامید. از طرف دیگر به منظور مقایسه پیمایشگر با یک سرویس‌دهنده پروکسی، آمار و ارقام سرویس‌دهنده پروکسی دانشگاه بوعلی سینا با توجه به در دسترس بودن آن استخراج شد. به منظور مقایسه تحلیلی ۴۵ روز فعالیت پروکسی مورد بررسی و تجزیه و تحلیل قرار گرفت. از سوی دیگر نیز به منظور مقایسه منش‌نمایی وب فارسی با کار مشابه انجام شده در زبان دیگر، منش‌نمایی وب شیلی [۲۱] انتخاب گردید. اگرچه حجم نمونه تهیه شده توسط پیمایشگر صبا در ظرف ۲۴ روز و پروکسی دانشگاه در مدت ۴۵ روز در مقایسه با کارهای مشابه در زبان‌های اروپایی و آمریکایی کمتر است اما نباید این نکته را فراموش کرد که میزان گستردگی زبان فارسی در وب با زبان‌های دیگر نیز اختلاف چشم‌گیری دارد.

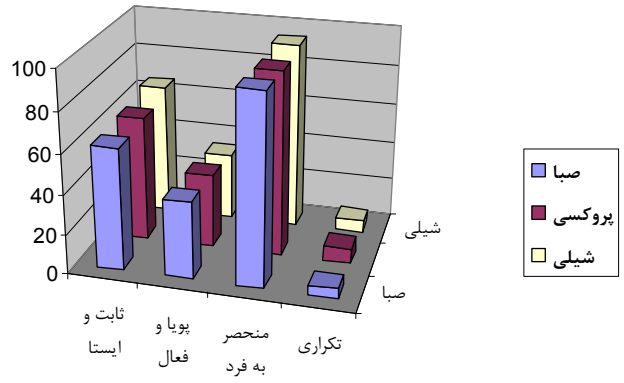
به جهت خلاصگی در ذکر عبارات سرویس‌دهنده پروکسی و پیمایشگر صبا در ادامه به ترتیب از واژه‌های کوتاه‌تر **پروکسی** و **صبا** استفاده شده است.

۴-۲- مشخصات مجموعه مورد آزمایش

حجم نمونه‌های بررسی شده در جدول (۱) و شکل (۲) نشان داده شده است.

صفحات وب بار گذاری شده	صبا		پروکسی		شیلی	
	صفحه	درصد	صفحه	درصد	صفحه	درصد
ایستا و ثابت	۱۲،۹۵۷	٪۶۱،۲۰	۱۶۶،۴۱۶	٪۶۲،۷۹	۲،۱۹۱،۵۲۲	٪۶۶،۱۵
پویا و فعال	۸،۲۱۳	٪۳۸،۸۰	۹۸،۵۷۸	٪۳۷،۲۱	۱،۱۲۱،۵۳۸	٪۳۳،۸۵
منحصربه‌فرد	۲۰،۱۳۳	٪۹۵،۱۰	۲۴۴،۶۲۰	٪۹۲،۳۸	۳،۱۱۰،۲۰۵	٪۹۳،۸۸
صفحات تکراری	۱،۰۳۷	٪۴،۹۰	۲۰،۱۷۵	٪۷،۶۲	۲۰۲،۸۵۵	٪۶،۱۲
مجموع	۲۱،۱۷۰		۲۶۴،۹۹۴		۳،۳۱۳،۰۶۰	

جدول(۱): حجم مجموعه‌های بدست آمده برحسب صفحات



شکل(۲): مقایسه مجموعه‌های بدست آمده برحسب صفحات

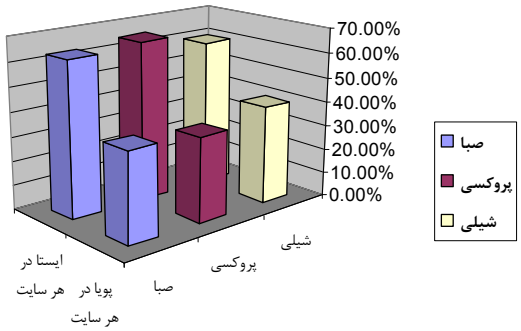
حجم نمونه‌ها و تقسیم‌بندی آنها به صفحات ایستا و پویا از آن جهت انجام گرفته است تا تغییر رفتار در هر یک از دو دسته به تفکیک قابل بررسی باشد. اگر چه مجموع صفحات بررسی شده صبا تقریبا ۱۲،۵ برابر کمتر از پروکسی و ۱۵۶ مرتبه کمتر از شیلی است اما توزیع درصدی صفحات نشان از توزیع یکنواخت صفحات ایستا (٪۶۱،۲۰ و ٪۶۶،۱۵) و پویا (٪۳۸،۸۰ و ٪۳۳،۸۵) در بین کاربران زبان فارسی و شیلی است.

از نظر تعداد صفحات ایستا و پویا در سایت‌های مختلف نیز موارد جدول (۲) و شکل (۳) قابل ملاحظه است.

همچنین اطلاعات جدول (۲) حاکی از آن است که میزان حجم سایت‌های فارسی و شیلی بر حسب تعداد صفحات از رفتار تقریبا مشابهی پیروی می‌کنند (٪۶۳،۴۶ و ٪۶۵،۴۴ و ٪۶۰،۱۸).

سایت‌های وب بار گذاری شده	صبا		پروکسی		شیلی	
	صفحه	درصد	صفحه	درصد	صفحه	درصد
متوسط صفحات ایستا در هر سایت	۵۸.۱۶	۶۳.۴۶	۵۲.۶۴	۶۵.۴۴	۴۰.۴۰	۶۰.۱۸
متوسط صفحات پویا در هر سایت	۳۳.۴۸	۳۶.۵۴	۲۷.۸۰	۳۴.۵۶	۲۶.۷۳	۳۹.۸۲
متوسط صفحات در هر سایت	۹۱.۶۴		۸۰.۴۴		۶۷.۱۳	
مجموع تعداد سایت	**۲۳۱**		**۳.۲۹۴**		**۴۹.۵۳۵**	

جدول(۲): مجموعه بدست آمده برحسب سایت



شکل(۳): مقایسه مجموعه‌های بدست آمده برحسب سایت

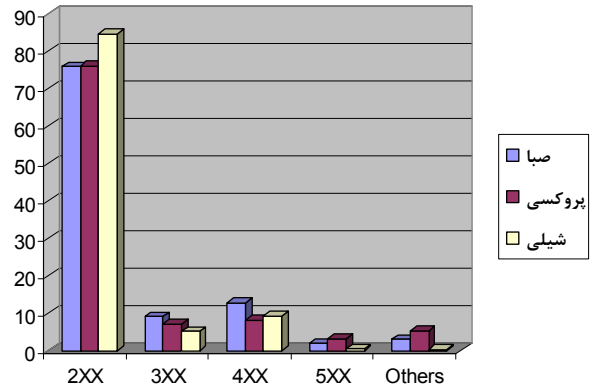
۴-۳- مقایسه براساس کد وضعیت صفحات

چندین کد وضعیت[۷] پاسخ HTTP به صورت ادغام شده درآمده که به شرح زیر آمده است:

- **حالت الف**: شامل کلیه درخواست‌هایی که منجر به انتقال و بارگذاری صفحات شده است: کد وضعیت‌های ۲۰۰ (OK) و ۲۰۶ (PARTIAL CONTENT)
- **حالت ب**: شامل کلیه هدایت‌هایی که به صفحات دیگری هدایت می‌شوند: کد وضعیت‌های ۳۰۱ (MOVED) ، ۳۰۲ (FOUND) و ۳۰۷ (TEMPORARY REDIRECT)
- **حالت ج**: شامل کلیه حالت‌هایی که از سوی سرویس‌دهنده با ممنوعیت مواجه می‌گردد: ۴۰۱ (UNAUTHORIZED) ، ۴۰۳ (FORBIDDEN) ، ۴۰۴ (NOT FOUND) و ۴۰۶ (NOT ACCEPTABLE)
- **حالت د**: حالت‌هایی که از طرف سرویس‌دهنده وب با خطا و شکست مواجه می‌گردد: ۵۰۰ (INTERNAL SERVER ERROR) ، و ۵۰۳ (UNAVAILABLE)
- **حالت ه**: حالت‌های دیگر از قبیل کد وضعیت‌نامشخص و خطاهای TCP ، DNS و به طور کلی نامشخص.

نوع حالت	صبا		پروکسی		شیلی	
	صفحه	درصد	صفحه	درصد	صفحه	درصد
حالت الف (2xx)	۱۶.۰۶۱	۷۵.۸۷	۲۰۱.۷۵۰	۷۶.۱۳	۲.۸۰۴.۵۰۶	۸۴.۶۵
حالت ب (3xx)	۱.۹۴۴	۹.۱۸	۱۸.۸۵۶	۷.۱۱	۱۷۰.۲۹۱	۵.۱۴
حالت ج (4xx)	۲.۷۱۰	۱۲.۸۰	۲۱.۶۵۲	۸.۱۸	۳۰۷.۷۸۳	۹.۲۹
حالت د (5xx)	۴۱۴	۱.۹۶	۸.۴۳۸	۳.۱۹	۱۸.۵۵۳	۰.۵۶
حالت ه	۶۴۱	۳.۰۳	۱۴.۲۹۸	۵.۴۰	۱۱.۹۲۷	۰.۳۶
مجموع	**۲۱.۱۷۰**		**۲۶۴.۹۹۴**		**۳.۳۱۳.۰۶۰**	

جدول(۳): توزیع کدوضعیت در حالت‌های مختلف



شکل(۴): مقایسه هر سه مجموعه در کدوضعیت پاسخ HTTP

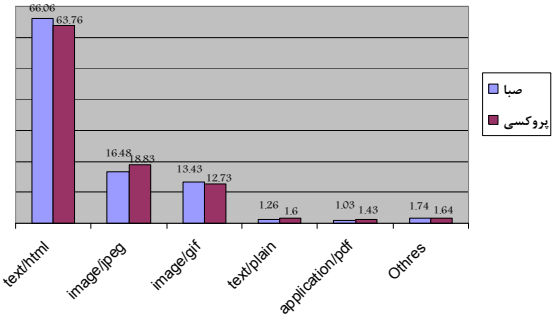
در رابطه با وضعیت یا حالت صفحات بارگذاری شده از نتایج جدول (۳) می‌توان این تحلیل را داشت که در زمان کوتاهی که از عمر وب‌سایت‌های فارسی در اینترنت می‌گذرد. در رابطه با موفقیت بارگذاری صفحات (حالت 2XX) با توجه به پایین بودن درصد صفحات بارگذاری موفق (۷۵.۸۷٪) نسبت به صفحات شیلی (۸۴.۶۵٪) هنوز نقص‌های زیادی برای صفحات می‌توان متصورشد. همچنین از نرخ بالاتر در صفحات جابجا شده (حالت 3XX) در صفحات فارسی (۹.۱۸٪) نسبت به صفحات شیلی (۵.۱۴٪) می‌توان به موقتی بودن صفحات فارسی و عدم ثبات صفحات فارسی در میزبان‌هایشان پی برد. قابل توجه است که همین توزیع به طور مشابه در صفحات غیرمجاز و ممنوعه (حالت 4XX) در صفحات فارسی (۱۲.۸۰٪) نسبت به صفحات شیلی (۹.۲۹) حاکی از ضعف برنامه نویسان و در خطاهای سرویس-دهنده (حالت 5XX) در صفحات فارسی (۱.۹۶٪) نسبت به صفحات شیلی (۰.۵۶٪) ضعف پشتیبانان را در پشتیبانی از وب‌سایت‌های فارسی نشان می‌دهد.

به طور کلی می‌توان چنین نتیجه گرفت که برنامه نویسان، پشتیبانان و مدیران وب‌سایت‌های فارسی عملکرد قابل قبولی نسبت به موارد مشابه نداشته و لازم است تدابیری از جمله آموزش فناوری‌های جدید در نظر گرفت تا از ضعف‌های موجود کاسته شود.

۴-۴- مقایسه براساس نوع فایل

نوع حالت	صبا		پروکسی	
	صفحه	درصد	صفحه	درصد
text/html	۱۳،۹۸۵	٪۶۶،۰۶	۱۶۸،۹۴۶	٪۶۳،۷۶
image/gif	۳،۴۸۸	٪۱۶،۴۸	۴۹،۹۰۵	٪۱۸،۸۳
image/jpeg	۲،۸۴۴	٪۱۳،۴۳	۳۳،۷۴۶	٪۱۲،۷۳
text/plain	۲۶۷	٪۱،۲۶	۴،۲۴۹	٪۱،۶۰
application/pdf	۲۱۸	٪۱،۰۳	۳،۷۸۵	٪۱،۴۳
Others	۳۶۸	٪۱،۷۴	۴،۳۶۳	٪۱،۶۴
مجموع	۲۱،۱۷۰		۲۶۴،۹۹۴	

جدول(۴): توزیع نوع فایل



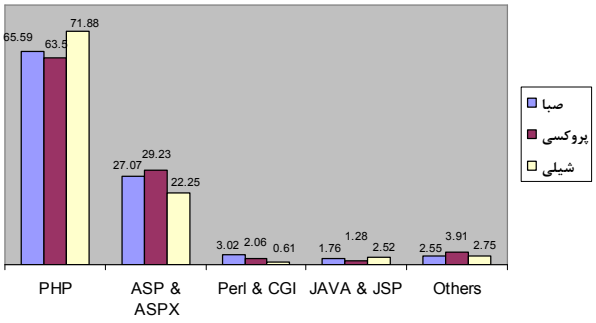
شکل(۵): مقایسه مجموعه‌ها براساس توزیع نوع فایل

۴-۵- مقایسه براساس پویایی صفحات

با توجه به این واقعیت که صفحات موجود در سایت‌ها بر پایه زبان‌های برنامه‌سازی وب و یا اسکریپ‌های مختلف تهیه شده‌اند نحوه توزیع این صفحات مورد علاقه این پژوهش بوده اسا که نتایج آزمایش آن در جدول (۵) و شکل(۶) آمده است. اطلاعات جدول (۵) نشان می‌دهد که PHP و ASP.NET محبوب‌ترین زبان‌های تولید صفحات در نمونه مورد بررسی است. همچنین میزان این محبوبیت در سایت-های شیلی به مراتب بالاتر از سایت‌های زبان فارسی است.

نوع حالت	صبا		پروکسی		شیلی	
	صفحه	درصد	صفحه	درصد	صفحه	درصد
PHP	۵،۳۸۷	۶۵،۵۹	۶۲،۵۹۷	۶۳،۵۰	۸۰۶،۰۶۱	۷۱،۸۸
ASP , ASPX	۲،۲۲۳	۲۷،۰۷	۲۸،۸۳۶	۲۹،۲۳	۲۴۹،۵۳۲	۲۲،۲۵
Perl & CGI	۲۴۸	۳،۰۲	۲،۰۳۰	۲،۰۶	۶،۸۴۱	۰،۶۱
JAVA & JSP	۱۴۵	۱،۷۶	۱،۲۶۱	۱،۲۸	۲۸،۲۶۲	۲،۵۲
مجموع	۸،۲۱۳		۹۸،۵۷۸		۱،۱۲۱،۵۳۸	

جدول(۵): توزیع صفحات پویا



شکل(۶): مقایسه مجموعه‌ها براساس توزیع صفحات پویا

همچنین میانگین حجم فایلی برای کل فایل‌های بارگذاری شده برای هر سایت به میزان 33.65KB با استفاده از پیمایشگر صبا و 27.98KB با استفاده از پروکسی بدست آمد.

۴-۶- دامنه‌های استخراجی

آمار دامنه‌های استخراجی از پیوندها به شرح زیر می‌باشد:

نام دامنه	صبا	شیلی [۲۱]	یونان [۲۲]
.com	٪۵۴،۶۳	٪۵۸،۶	٪۴۹،۲
.ir	٪۲۱،۲۰	-	-
.net	٪۱۰،۰۷	٪۶،۴	٪۸،۵
.org	٪۹،۱۵	٪۱۵،۴	٪۱۷،۹
.ac.ir	٪۳،۷۸	-	-
.edu	٪۱،۱۷	٪۱،۳	٪۲،۶

جدول(۶): توزیع دامنه‌های استخراجی در سطح وب در مقایسه با نمونه‌های شیلیایی و یونانی

۵- نتیجه و پیشنهاد ادامه پژوهش

در این مقاله، درابتدا به مسائل و مشکلات موجود در جامعه اطلاعاتی و افزایش حجم انبوه اطلاعات اشاره شد و از موتورهای جستجو و به دنبال آن پیمایشگرهای متمرکز به عنوان راه‌حلی برای حل این مشکل یاد شد. برای بررسی منش‌نمایی وب فارسی نیز از پیمایشگر صبا استفاده شد که معماری و جزئیات کلی آن ذکر گردید.

به عنوان پیشنهادات آینده می‌توان استفاده از برهه‌های زمانی مختلف همراه با ازدیاد دوره نمونه‌برداری و تمرکزسازی بر روی موضوعات ویژه را مورد آزمون قرار داد.

مراجع

[1] Internet Systems Consortium, http://www.isc.org/ , 2006.

[2] History and Growth of the Internet. http://www.internetworldstats.com/stats.htm , 2006.

[3] Hobbes' Internet Timeline v8.1. http://www.zakon.org/robert/internet/timeline/ , 204 .

[4] Steve Lawrence and C. Lee Giles, "*Accessibility of information on the web*", Nature, 400:107109, 1999.

[5] Vladislav Shkapenyuk and Torsten Suel, "*Design and implementation of a high-performance distributed web crawler*", In Proceedings of the 18th International Conference on Data Engineering (ICDE) pages 357 – 368, San Jose, California, February 2002, IEEE CS Press.

[6] Steve Lawrence and C. Lee Giles, "*Searching the World Wide Web*", Science 280(5360):98– 100, 1998.

[7] StatMarket, "*Search engine referrals nearly double worldwide*", http://websidestory.com/pressroom/pressreleases.html?id=181, 2003.

[8] Jakob Nielsen, Statistics for traffic referred by search engines and navigation directories to useit. http://www.useit.com/about/searchreferrals.html , 2003.

[9] M. Shokouhi, P. Chubak, F. Oroumchian, H. Bashiri, "*Design and Implementation of Regional Crawler as a New Sterategy for Crawling the Web*", Proceeding of 2nd IADIS international Confrence, e-society2004, Avila, Spain, July 2004.

[10] Soumen Chakrabarti, Martin van den Berg, and Byron Dom, "*Focused crawling: a new approach to topic-specific web resource discovery*", Computer Networks, 31(11–16):1623–1640, 1999.

[11] A.R Rezvanian, H. Bashiri, "*SABA Web Crawler: Design and Implementation of an Effective Web Crawler*", Thesis in Partial Fulfillment of requirement for the Degree of Bachelor of Software Engineering, Bu-Ali Sina University, Hamedan, Iran, September 2006.

[12] The Web Robots page: http://www.robotstxt.org/wc/robots.html/ , 2003.

[13] Eveline A. Veloso, Edleno de Moura, P. Golgher, A. da Silva, R. Almeida, A. Laender,B. Ribeiro-Neto, and Nivio Ziviani, "*Um retrato da web brasileira*", In Proceedings of Simposio Brasileiro de Computacao, Curitiba, Brasil, July 2000.

[14] Ricardo Baeza-Yates and B´arbara Poblete, "*Evolution of the Chilean Web structure composition*", In Proceedings of Latin American Web Conference, pages 11–13, Santiago, Chile, 2003, IEEE CS Press.

[15] Daniel Gomes and Mrio J. Silva, "*A characterization of the portuguese web*", In Proceedings of 3rd ECDL Workshop on Web Archives, Trondheim, Norway, August 2003.

[16] Ricardo Baeza-Yates, "*The Web of Spain*", UPGRADE, 3(3):82–84, 2003.

[17] Andr´as A. Bencz´ur, K´aroly Csalog´any, D´aniel Fogaras, Eszter Friedman, Tam´as Sarl´os, M´at´e Uher, and Eszter Windhager, "*Searching a small national domain – a preliminary report*", In Poster Proceedings of Conference on World Wide Web, Budapest, Hungary, May 2003.

[18] Andreas Rauber, Andreas Aschenbrenner, Oliver Witvoet, Robert M. Bruckner, and Max Kaiser. "*Uncovering information hidden in web archives*", D-Lib Magazine, 8(12), 2002.

[19] Global internet usage - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Internet_users , 2006.

[20] Monika Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork, "*On near– uniform url sampling*", In Proceedings of the Ninth Conference on World Wide Web, pages 295–308, Amsterdam, Netherlands, May 2000, Elsevier Science.

[21] Carlos Castillo, Ricardo Baeza-Yates, "*Effective Web Crawling*", Ph.D thesis, University of Chile, 2004.

[22] Stephen Dill, Ravi Kumar, Kevin S. Mccurley, Sridhar Rajagopalan, D. Sivakumar, and Andrew Tomkins, "*Self-similarity in the web*", ACM Trans, Inter. Tech., 2(3):205–223, 2002.

[23] B. Lavoie and H. F. Nielsen, "*Web Characterization Terminology & Definitions Sheet*" , http://www.w3.org/1999/05/WCA-terms/ , 1999.

[24] D. Menasce, B. Abrahao, D. Barbara, V. Almeida, F. Ribeiro, "*Fractal Characterization of Web Workloads*", In Proceedings of the 11th International World Wide Web Conference, 2002.

زیرنویس‌ها

[1] Hidden Web
[2] Loader
[3] Index
[4] Indexer
[5] Semantic Web
[6] Internet Services Provider (ISP)
[7] Status Code