تعیین Bias موتورهای جستجوگر از Robots.txt

Yang Sun, Ziming Zhuang, Isaac G. Councill, and C. Lee Giles
Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16801, USA
{ysun, zzhuang, icouncill, giles}@ist.psu.edu

مترجم : مسعود مجلسی
دانشگاه غیر انتفاعی اشراق بجنورد
Masoud.Majlesi@Gmail.com

خلاصه

موتورهای جستجوگر برای جمع آوری اطلاعات از وب بسیار به روبوت‌ها متکی هستند . این چنین فعالیت‌های خزیدن (در وب) می‌تواند در طرف سرور با به کارگیری پروتکل‌های ممانعت از روبوت‌ها در فایلی به نام Robots.txt کنترل گردد . روبوت‌های متعهد از قواعد مشخص شده در Robots.txt پیروی می‌کنند . وب سایت‌ها می‌توانند اولویت‌های دسترسی برای هر روبوت را بوسیله نام تعیین کنند . این bias ممکن است منجر به قدرت‌مندتر شدن قدرمت‌مندان گردد ، به طوری که تعداد محدودی موتورهای جستجوگر محبوب در نهایت بر وب حکمفرما شوند به دلیل اینکه آن‌ها تقدم دسترسی به منابعی را دارند که دیگران به آن دسترسی ندارند.این موضوع به ندرت مورد توجه قرار می‌گیرد ، اگرچه قرارداد Robots.txt به یک استاندارد غیر رسمی برای تنظیم روبوت‌ها تبدیل شده و موتورهای جستجوگر یک ابزار واجب برای دسترسی به اطلاعات شده است . ما یک معیار برای ارزیابی مقدار Bias برای روبوتهای ویژه‌ای که گفته شد پیشنهاد می‌کنیم . ما ۷۵۹۳ وب سایت که شامل دومین‌های آموزشی ، دولتی ، خبری و تجاری را مورد بررسی قرار داده و ۲۹۲۵ فایل Robots.txt مجزا را جمع‌آوری کرده‌ایم . نتایج و تحلیل‌های آماری از اطلاعات ثبت شده این است که روبوت‌های موتورهای جستوگر محبوب و پرتال‌های اطلاعاتی توسط اکثر وب سایت‌هایی که ما به عنوان نمونه نام برده‌ایم مورد همکاری قرار گرفته است . همچنین نتایج نشان می‌دهد که یک رابط قوی میان Bias و سهم موتورهای جستجوگر از بازار بخصوص در مورد روبوت‌های موتور جستجوگر وجود دارد.

۱. مقدمه :

شاید بدون روبوت‌ها موتورهای جستجوگری نبود . موتورهای جستجو وب ، کتابخانه‌های دیجیتال ، و بسیاری دیگر از برنامه‌های کاربردی وب همچون مرورگرهای آفلاین ، نرم‌افزارهای خرید اینترنتی و عاملین هوشمند جستجو برای بدست آوردن مستندات بسیار به روبوت‌ها وابسته‌اند . روبوت‌هایی که به نام های spider ,crawler,bot,harvesters شناخته می‌شوند ، عامل‌های خودکاری هستند برای هدایت هایپرلینک‌های وب ، کنکاش در منابع موضوعی برای مدیریت انسانی مبدإ هزینه است . به دلیل ذات خودکار روبوت‌ها ، قواعد باید طوری تنظیم گردند که چنین فعالیت‌های کنکاشی مانع از اثر نامطلوب بر تراکم کاری سرورها و نیز دسترسی به اطلاعات غیر عمومی و شخصی گردد .

پروتکل‌های ممانعت از روبوت‌ها قواعدی را برای پیروی روبوت‌ها پیشنهاد کرده است . فایلی به نام Robots.txt که شامل سیاست‌های دسترسی روبوت‌ها می باشد در فهرست ریشه یک وب سایت به کار گرفته می‌شود و برای تمام روبوت‌ها قابل دسترسی است . روبوت‌های متعهد این فایل را می‌خوانند و در طول سرکشی به وب سایت قوانین آن را اجرا می‌کنند . قرارداد Robots.txt از اواخر دهه ۹۰ در مجامع پذیرفته شده و همچنان به عنوان ابزاری فراگیر و غالب برای تنظیمات روبوت‌ها به کار می‌رود . هر چند با وجود مشخص بودن قرارداد Robots.txt برای هر دو تولید کنندگان محتوا و نیز کنکاشگران ، کارهای کمی برای بررسی فواید آن با جزئیات انجام شده است بخصوص در مقیاس وب .

وب سایت‌ها ممکن است قواعد دسترسی مختلفی را به روبوت‌ها اختصاص دهند ، این Bias می‌تواند منجر به قدرتمند شدن قدرتمندان گردد که بوسیله آن بعضی از موتورهای جستجوگر محبوب دارای امتیاز انحصاری دسترسی به منابع معینی می‌شوند که در نتیجه می‌تواند آن‌ها را باز هم محبوب تر کند . نظر به این واقعیت که کاربران اغلب موتورهای جستجوگر با پوشش اطلاعات گسترده را ترجیح می‌دهند این پدیده (قدرتمندان قدرتمندتر می‌شوند) نشانگر تاثیر زیاد انتخاب موتور جستجوگر توسط کاربران است ، که سرانجام به سهم بیشتر موتورهای جستجوگر در بازار می‌انجامد ، به عبارت دیگر (اگر چه شاید اغراق باشد) "چیزی که قابل جستجو نیست پس وجود ندارد" این پدیده دید جهت گیرانه اطلاعات روی وب را نشان می‌دهد .

۱-۱ : کارهای مشابه و همکاری‌ها :

مطالعه‌ی استفاده از Robots.txt در دانشگاه‌های انگلیس و ۱۶۳ وب سایت و ۵۳ Robots.txt دانشگاهی بررسی شده در ۱۹۹۹ . فایل‌های Robots.txt در زمینه حجم فایل و نیز پروتکل‌های ممانعت روبوت در داخل دومین‌های دانشگاهی انگلیس مورد بررسی قرار گرفتند . در سال ۲۰۰۲ Drott استفاده‌ی Robots.txt را در کمک به شاخص‌گذاری (Indexing) اطلاعات محافظت شده را مورد مطالعه قرار داد . ۶۰ نمونه از وب‌سایت‌های ۵۰۰ شرکت جهانی که به صورت دستی مورد آزمون قرار گرفت که در این آزمایش این نتیجه حاصل شد : (Robots.txt به طور زیاد در گروه مورد آزمایش استفاده نمی‌شود و در اکثر آن‌هایی که دیده می‌شود زائد است ... آن‌ها روبوت‌ها را از فهرست‌هایی محروم کرده‌اند که در هر صورت قفل شده است.) . بررسی ما نتایج متفاوتی را نشان می‌دهد که ممکن است ناشی از نمونه حجم، دومین و زمان متفاوت باشد . نمونه دیگر اشاره به جنبه حقوقی اجرای Robots.txt و نظرکلی در مورد روبوت‌های وب و استفاده از Robots.txt که ضمیمه شده . هیچ یک از تحقیقات مذکور محتویات Robots.txt را بر حسب Bias نسبت به روبوت‌های مختلف بررسی نکرده است . به علاوه نمونه مطالعات گذشته با تمایل نسبتا کمی به حجم وب انجام شده است . در این مقاله ما برای اولین بار یک مطالعه کمی در مورد این Bias و یک تحقیق جامع‌تر در زمینه استفاده از Robots.txt در وب را ارائه می‌دهیم . با به کار گرفتن کنکاشگر مخصوص Robots.txt خودمان ، مقدار قابل توجهی داده در دنیای واقعی از وب سایت‌های مختلف مجزا شامل دومین‌های آموزشی ، دولتی ، خبری و تجاری جمع‌آوری کرده‌ایم . ما این پرسش‌ها را مورد بررسی قرار داده‌ایم :

- آبا Bias روبوت وجود دارد ؟
- چگونه چنین Bias به طور کمی اندازه‌گیری می‌شود ؟
- معنی و استنباط از این Bias چیست ؟

پیشنهادات ما :

- پیشنهاد یک معیار کمی برای اندازه گیری خودکار Bias روبوت .
- با به بکارگیری معیار در نمونه‌ی بزرگی از وب‌سایت ها یافته‌های خود را در مورد روبوت‌های پرطرفدار و بالعکس ارائه می‌دهیم .

ادامه مقاله به این صورت سازمان یافته : در بخش ۲ به طور مختصر پروتکل‌های ممانعت از روبوت‌ها را معرفی می‌کنیم . در بخش ۳ جمع‌آوری اطلاعات این مطالعات را ارائه می‌دهیم . در بخش ۴ معیار Bias را ارائه می‌دهیم و شرح می‌دهیم که چگونه آن را برای اندازه‌گیری bias روبوت‌ها استفاده کرد . در بخش ۵ مشاهداتمان را روی Robots.txt و در مورد مفاهیم آن بحث می‌کنیم . در بخش ۶ نتیجه‌گیری مقاله و برای برای کارهای آینده .

۲. پروتکل ممانعت از روبوت‌ها :

پروتکل‌های ممانعت از روبوت‌ها یک قرارداد است که به مدیران وب سایت اجازه می‌دهد که به روبوت‌ها نشان دهد که کدام بخش از سایت آن‌ها نباید مورد بازدید قرار گیرد . اگر در وب سایت فایل Robots.txt نباشد روبوت‌ها برای دسترسی به تمام محتویات آزاد هستند .

قالب پروتکل‌های ممانعت از روبوت‌ها در [12] توضیح داده شده است . فایلی که Robots.txt نام دارد تحت فهرست ریشه وب سرور قرار می‌گیرد . هر خط فایل Robots.txt به صورت زیر می‌باشد :

<field>:<optionalspace>< value ><optionalspace>

سه نوع برچسب حساس به کوچک و بزرگ بودن حروف برای <field> وجود دارد که دستورات را معین می‌کند : User-Agent , Allow , Disallow . ونیز دستور غیر رسمی Crawl-Delay توسط بسیاری از وب سایت‌ها برای محدود کردن تعدد بازدیدهای روبوت استفاده می گردد .

فایل Robots.txt با یک یا چند فیلد User-Agent برای مشخص کردن اینکه کدام روبوت‌ها را شامل می‌شود آغاز می‌گردد و در ادامه تعدادی فیلدهای Allow یا Disallow که نمایان‌گر تنظیمات روبوت است می‌باشد . توضیحات در همه جای فال مجاز هستند و می‌توانند شامل فضای خالی اختیاری باشند . توضیحات با حرف توضیح '#' آغاز و Linebreak خاتمه می‌یابد . یک نمونه از Robots.txt را در زیر می‌بینید .

```
User-Agent: *
Disallow: /robots/
Disallow: /src/
Disallow: /botseer
Disallow: /uastring
Disallow: /srcseer
Disallow: /robotstxtanalysis
Disallow: /whois

User-Agent: googlebot
Disallow: /robotstxtanalysis
Disallow: /uastring

User-Agent: botseer
Disallow:
```

این فایل نشان می‌دهد که Googlebot نمی‌تواند "/robotstxtanalysis" و "/uastring" را مشاهده کند . Botseer می‌تواند تمام فهرست‌ها و فایل‌های روی سرور را بازدید کند . تمام روبوت‌های دیگر باید از قوانین * : User-agent پیروی کنند که بر طبق آن نمی‌توانند فایل‌ها و فهرست‌های زیر را بازدید کنند :

"/robots/" , "/src/" , "/botseer" , "/uastring" , "/srcseer" , "/robottxtanalysis" ,"/whois "

۳. bias روبوت :

ما $\Delta P(r)$ یک مقیاس برای میزان مساعدت روبوت‌ها در یک فایل نمونه Robots.txt را پیشنهاد می‌کنیم ، که برای اندازه گیری میزان اعتبار یک روبوت توسط مجموعه‌ای از وب سایت‌ها می‌باشد . تعریف رسمی Bias روبوت در زیر توضیح داده شده است .

۳-۱ الگوریتم بدست آوردن Bias :

تعریف ما از یک روبوت معتبر یا همکار روبوتی است که بر اساس فایل Robots.txt اجازه دسترسی به فهرست‌های بیشتری نسبت به روبوت‌های همه منظوره را دارد . روبوت‌های همه منظوره به هر روبوتی گویند که با هیچ کدام از اسامی User-Agent در فایل Robots.txt مطابقت ندارد . به عبارت دیگر نام روبوت‌های همه منظوره در فایل‌های Robots.txt ظاهر نمی‌شود .

مجموعه فایل‌های Robots.txt در مجموعه داده‌هایمان را F می‌گذاریم . یک فایل Robots.txt معین را f€F می‌گذاریم ، و مجموعه از روبوت‌های نام برده شده در Robots.txt فایل f را R می‌گذاریم . براي هر روبوت r€R الگوریتم GetBias(r,f) را که در الگوریتم ۱ آمده است را تعریف می‌کنیم . Getbias درجه اعتبار یک روبوت r را که در Robots.txt فایل f آمده را بیان می‌کند .

Algorithm 1 $GetBias(r, f)$

```
if r is * then
    return 0
end if
Construct DIR for f;
bias = 0
for all d ∈ DIR do
    if d is allowed for * then
        D_u ← d
    end if
end for
for all d ∈ DIR do
    if d is allowed for r then
        D_r ← d
    end if
end for
bias = |D_r| − |D_u|
return bias
```

مجموعه فهرست‌هایی که در Robots.txt فایل f در یک وب سایت مشخص ظاهر می‌شود را DIR می‌گذاریم . DIR برای برآورد ساختار واقعی یک فهرست به استفاده می‌شود به این دلیل که پروتکل ممانعت از روبوت‌ها به طور پیش فرض فهرست‌هایی که با Robots.txt مطابقت ندارد را به عنوان فهرست‌های مجاز در نظر می‌گیرد .

D_u € DIR مجموعه‌ای از فهرست‌هایی است یک روبوت همه منظوره برای بازدید از آن مجاز است . اگر هیچ شرطی برای User-Agent:* وجود نداشته باشد روبوت به طور پیش فرض می‌تواند به همه چیز دسترسی داشته باشد . D_r € DIR مجموعه‌ی فهرست‌هایی است که روبوت اجازه بازدید از آن را دارد . $|D_u|$ و $|D_r|$ تعداد فهرست‌ها در D_r و D_u است .

برای روبوت داده شده r الگوریتم ابتدا محاسبه می‌کند چه تعداد از فهرست‌ها در DIR در r مجاز هستند . سپس مقدار Bias را که تفاوت بین تعداد فهرست‌هایی موجود در DIR که برای روبوت r مجاز است و تعداد فهرست‌هایی که برای روبوت‌های همه منظوره مجاز هستند را بدست می‌آورد . در الگوریتم Getbias برای روبوت‌های همه منظوره مقدار مرجع صفر را برمیگرداند (GetBias Returns 0) . مقدار Bias برای روبوت‌های معتبر مقداری مثبت را برمی‌گرداند . مقدار بیشتر برای یک روبوت به معنای اعتبار بیشتر روبوت است . بالعکس مقدار bias برای روبوت‌های نامعتبر منفی می‌باشد که با ما برای Bias سازگار است . بنابراین Bias یک روبوت در Robots.txt می‌تواند توسط متغیرهای طبقه‌بندی شده نمایش داده شود که سه نوع طبقه‌بندی معتبر ، نامعتبر و بدون Bias را شامل می‌شود . به عنوان مثال Robots.txt را در http://BotSeer.ist.psu.ed*u* مورد بررسی قرار می‌دهیم .

DIR ={“/robots/”, “/src/”, “/botseer”, “/uastring”, “/srcseer”,“/robotstxtanalysis”, “/whois”}.

بر طبق الگوریتم داریم :

D_u = {null}
$D_{botseer}$ = {“/robots/”, “/src/”, “/botseer”, “/uastring”, “/srcseer”, “/robotstxtanalysis”,“/whois”}
D_{Google} = {“/robots/”, “/src/”, “/botseer”, “/srcseer”, “/whois”}.

بنابراین

$|D_u|=0$, $|D_{botseer}|=7$, $|D_{Google}|=5$

بنا بر الگوریتم ۱ :

$|D_u|-|D_u|=0$, $bias_{botseer}= |D_{botseer}|-|D_u|=7$, $bias_{Googlebet}=|D_{Google}|-|D_u|=5$

بنابراین روبوت‌هاي googlebet و botseer توسط این وب سایت مورد اعتماد هستند و در طبقه معتبر قرار مي‌گيرد و بقيه روبوت‌ها در طبقه بدون Bias.

۳-۲ اندازه گیری bias کلی :

بر اساس مقدار Bias برای هر فایل ، $\Delta P(r)$ را به منظور محاسبه برای مجموعه‌ای از فایل‌های Robots.txt ارائه می‌کنیم . $N=|F|$ تعداد کل فایل‌های Robots.txt در مجموعه داده‌ها است . $\Delta P(r)$ می‌تواند به صورت زیر تعریف گردد :

$$\Delta P(r) = P_{favor}(r) - P_{disfavor}(r)= \frac{N_{favor}(r) - N_{disfavor}(r)}{N}. \quad (۱)$$

که $N_{favor}(r)$ و $N_{disfavor}(r)$ به ترتیب برابر تعداد دفعاتی است که روبوت مورد تآیید و عدم تایید قرار می‌گیرد .

$P_{favor}(r)$ نسبتی از فایل Robots.txt r که مورد تایید قرار می‌گیرد و $P_{disfavor}(r)$ نسبتی از فایل Robots.txt r است که مورد تایید قرار نمی‌گیرد .

نسبت فایل Robots.txt که مورد تایید و یا عدم تایید قرار می‌گیرد برای بررسی آماری محاسبه ساده‌ای دارد . از آنجائیکه بیشتر از دو مقدار داریم (favor , disfavor , no bias) جداسازی دو مقدار برای محاسبه Bias کلی منتج به نتیجه دقیق نمی‌شود . به عبارت دیگر $P_{favor}(r)+P_{disfavor}(r)<1$

هر مقدار فقط یک وجه از bias را نتیجه می‌دهد .برای مثال روبوتی به نام ia_archiver 0.24% مورد تایید وب‌سایت‌های مجموعه اطلاعات ما قرار می‌گیرد و در مورد momspider 0.21% . و به ترتیب عدم تاییـد ia_archiver و momspider برابر است با 1.9% و 0% . اگر ما فقط نسبت تایید را مدنظر قرار دهیم ia_archiver نسبت به momspider درجه تایید بیشتری دارد .

$\Delta P(r)$ اختلاف بین نسبت سایت‌های که یک روبوت خاص را مورد تایید و یا عدم تایید قرار داده‌اند می‌باشد بنابراین هر دو مورد هماهنگ هستند . برای مثال بالا $\Delta p(ia_archiver)=1.66\%$ و $\Delta p(momspider)=0.21\%$ بنابراین momspider درجه تایید بیشتری از ia_archiver دارد . برای فایل r بدون bias داریم $\Delta p(r)=0$
محاسبه bias می‌تواند موارد گمراه کننده را از بین ببرد .

۳-۳ آزمودن تایید پذیری

تایید پذیری در واقع یک تابع رتبه بندی برای روبوت‌ها است . برای محاسبه دقت این تابع رتبه‌بندی ما تست کارایی رتبه‌بندی را بر اساس متد ضریب رتبه بندی kendall [11] اجرا کردیم . ضریب رتبه‌بندی به طور خلاصه در ادامه آمده است . جزئیات ارزیابی کارائی رتبه‌بندی می‌توانید در [9] ببینید . برای فایل Robots.txt به نام f ، m_a را برابر اندازه تابع bias برای تمام روبوت‌هایی که در f دیده می‌شود قرار می‌دهیم ، r_i و r_j عضو R را دو روبوت در f فرض می‌کنیم . نشان می‌دهیم $r_i<m_a$ r_j اگر r_i در رتبه بالاتری از R_j با اندازه‌گیری m_a باشد . بنابراین برای هر دو مقدار m_a و m_b kendall τ را تعریف کرده است براساس P_f که مقدار دوتایی‌هاای هماهنگ و Q_f که مقدار دوتایی‌های ناهماهنگ باشند . دو مقدار $r_i \neq r_j$ هماهنگ هستند اگر m_a و m_b با یکدیگر مطابقت داشته باشد و ناهماهنگ هستند اگر مطابق نباشند . در این مورد kendall τ را به این صورت بیان کرده است :

$$\tau_f(m_a, m_b) = \frac{P_f - Q_f}{P_f + Q_f} \qquad (۲)$$

برای هر مقداری مه m_a و m_b داشته باشند $\tau_f(m_a,m_b)$ نشان می‌دهد که دو مقدار چگونه در فایل f با هم مطابقت دارند . m_a را نشان‌گر مقدار واقعی تابع رتبه بندی روبوت می‌دانیم . در صورتیکه که ما تابع رتبه‌بندی واقعی را نمی دانیم و فقط رتبه‌بندی جزئی روبوت را برای هر Robots.txt داریم که بر اساس مقدار bias که قبلا توضیح دادیم بدست می‌آید . بنابراین $\tau_f(m_a,m_b)$ برای هر فایل Robots.txt در مجموعه داده را محاسبه می‌کنیم . اگر برای یک فایل داده شده Robots.txt $\tau_f(m_a,m_b)=1$ به این نتیجه می‌رسیم که فایل مذکور فایلی متوازن است و در غیر این صورت فایلی غیر متوازن . با محسبه فایل‌های متوازن و نامتوازن در مجموعه داده‌ها داریم : $(P+Q=N)$

$$\overline{\tau(m_a, m_b)} = \frac{P - Q}{P + Q} = 1 - \frac{2Q}{N}. \qquad (۳)$$

ما با استفاده از δP روبوت‌ها را رتبه‌بندی کردیم . لسیت رتبه‌بندی شده با رتبه‌بندی واقعی مقایسه شد . مقدار میانگین τ برابر 0.957 که معتقدیم به اندازه کافی برای اندازه‌گیری مقدار bias دقیق می‌باشد .

۴- جمع‌آوری اطلاعات :

به منظور پتانسیل bias روبوت ها در وب فعالیت مطالعاتی ما در طیف گسترده‌ای از وب سایت‌ها از دومین‌های مختلف و نیز مکان‌های مختلفی انجام شده است که با توضیحات در زیر آمده است .

۴-۱ منابع اطلاعاتی :

Open Directory Project[6] بزرگترین و گسترده‌ترین فهرست وب است که با نیروی انسانی جمع‌آوری شده است . منبع اصلی ما برای جمع‌آوری URLهای اولیه DMOZ بود به دلیل اینکه Open Directory Project تعداد گسترده‌ای از URLها را در طبقات مختلفی طبقه‌بندی کرده است . که ما را قادر ساخت تا داده‌هایمان را از دومین‌های مختلف و از مکان‌های مختلف جمع‌آوری کنیم . اطلاعات ما از این منبع سه نوع دومین آموزشی ، خبری و دولتی بود . دومین‌های دانشگاهی بعداً به دومین دانشگاهی آمریکایی ، اروپایی و آسیایی تقسیم شد .
از آنجائیکه ساختار فهرست دومین‌های تجاری در در DMOZ پیچیده بود و بسیاری از آن‌ها مشترک بودند ما از لیست ۱۰۰۰ شرکت ثروتمند در ۲۰۰۵ استفاده کردیم . قطعاً در داده‌های جمع‌آوری شده ما محدودیت‌هایی وجود دارد .اولاً بدلیل اینکه مجموعه وب‌سایت‌های DMOZ برای سایر کشورها بخصوص وب‌سایت‌های غیر انگلیسی زبان محدود بود و اکثر وب‌سایت‌ها از آمریکا بودند . دوماً بدلیل اینکه وروی‌های DMOZ توسط نیروی انسانی ویرایش و سازماندهی می‌شده احتمال وجود خطا می‌باشد . و نهایتاً ما

وب‌سایت‌های تجاری را از میان لیست ۱۰۰۰ شرکت ثروتمند جمع‌آوری کردیم که محتوی اطلاعاتی در مورد شرکت‌های بزرگ بود که این دومین‌ها نمی‌تواند نمایند شرکت های تجاری کوچک باشد و ما قصد داریم در تحقیقات آینده به این محدودیت‌ها اشاره کنیم .

۴-۲ کاوشگری برای Robots.txt

ما یک کاوشگر مخصوص برای این مطالعات پیاده‌سازی کردیم . کاوشگر با کاوش در وب سایت‌های که در DMOZ با طبقه‌بندی موضوعی ، نام وب‌سایت و سازمان‌های وابسته در مکان‌های مختلف شروع به کار کرد . سپس کاوشگر وجود Robots.txt را در دومین‌های مذکور بررسی و در صورت وجود آن‌را برای تحلیل‌های آفلاین دانلود می‌کرد. یک ماژول جدا کننده و فیلتر کننده نیز کاوشگر اغام شده بود تا تکراری‌ها را از بین ببرد و از اینکه صفحات بدست آمده در دومین‌های مقصد هستند مطمئن گردد .

علاوه بر فهرست سطح ریشه کاوشگر ما سایر مکان‌های ممکن را برای Robots.txt بررسی می‌کرد . زیر فهرست‌های وب سایت تا سطح ۳ سرکشی شد . نتایج نشان داد که موارد کمی بود که Robots.txt در فهرست ریشه نباشد که بر طبق پروتکل ممانعت از روبوت است . فایل‌های با نام غلط نیز توسط کاوشگر ما مورد بررسی قرار گرفت که دذ موارد نادری Robot.txt به جای Robots.txt استفاده شده بود .

برای مشاهده کردن مشخصات زودگذر کاوش ۵ بار برای همان مجموعه وب‌سایت از دسامبر ۲۰۰۵ تا اکتبر۲۰۰۶ . به منظور تحلیل مشخصات موقتی فایل‌های Robots.txt دانلود شده بر اساس تاریخ کاوش آرشیو شد .

۴-۳ آمار :

ما ۷۵۹۳ وب سایت مجزا را که شامل ۶۰۰ وب سایت دولتی ، ۲۰۴۷ وب سایت روزنامه ، ۱۴۸۷ وب سایت دانشگاهی آمریکایی ، ۱۴۲۰ وب سایت دانشگاهی اروپایی ، ۱۰۳۹ وب سایت دانشگاهی آسیایی و وب سایت ۱۰۰۰ شرکت را مورد کاوش و بررسی قرار دادیم . تعداد وب سایت‌هایی که Robots.txt داشتند در ۵ مرحله کاوش در جدول ۱ آمده است .

	Websites	Collected robots.txt files				
		Dec. 2005	Jan. 2006	May. 2006	Sep. 2006	Oct. 2006
Government	600	248	257	263	262	264
Newspaper	2047	859	868	876	937	942
USA Univ.	1487	615	634	650	678	683
European Univ.	1420	497	510	508	524	537
Company	1000	303	306	319	341	339
Asian Univ.	1039	140	248	149	165	160
Total	7593	2662	2823	2765	2907	2925

Table 1. Number of robots.txt found in each domain for each crawl.

توضیح بهتر استفاده از Robots.txt در وب‌سایت‌های دومین‌های مختلف در شکل ۱ که نشانگر نسبت وب‌سایت‌هایی که دارای این فایل هستند مشخص شده است .

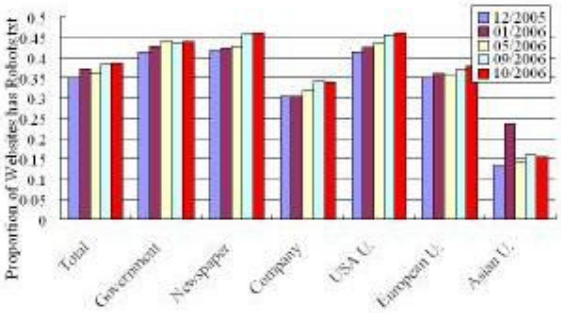
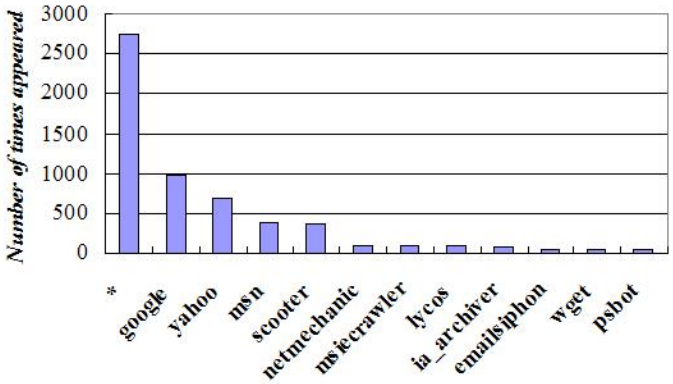


Figure 1. Probability of a website that has robots.txt in each domain.

به طور کلی به جز در مورد وب سایت‌های دانشگاه‌های آسیایی استفاده از Robots.txt افزایش یافته است . 46.02% از وب سایت‌های روزنامه‌ها اکنون Robots.txt را اجرا کرده‌اند و دومین‌های روزنامه که بیشترین پذیرش پروتکل ممانعت از روبوت را داشته‌اند . 45.93% از وب سایت‌های دانشگاهی آمریکا پروتکل ممانعت از روبوت را پذیرفته‌اند که بیشتر از سایت‌های اروپا (37.8%) و آسیا (15.4%) است . دلیل اینکه موتورهای جستجوگر و عاملین جستجوی هوشمند بسیار برای دسترسی به اطلاعات وب با اهمیت شده‌اند این نتایج مورد انتظار بود . پروتکل ممانعت از روبوت بیشتر توسط وب سایت‌های دولتی ، روزنامه‌ها و دانشگاهی در آمریکا مورد پذیرش واقع شده است که به طور گسترده‌ای از آن برای حفاظت از اطلاعاتی که نباید در اختیار عموم قرار بگیرد و توازن تراکم این وب سایت‌ها استفاده می‌شود . یک گزارش تفضیلی از استفاده از Robots.txt را می‌توانید در [6] ببینید .

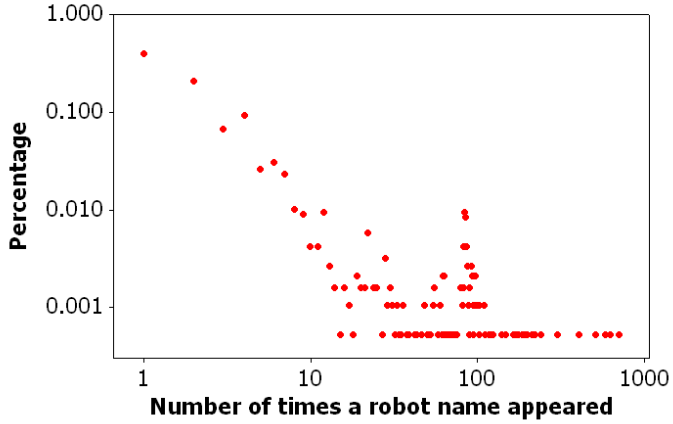
۵- نتایج :

در مجموعه داده‌های ما ۱۰۵۶ روبوت یافت می‌شود . روبوت‌های عمومی بیشترین استفاده را در فیلد User-Agent با ۲۷۴۴ بار داشته است ، که به معنی این است که 93.8% از فایل‌های Robots.txt برای روبوت‌های عمومی اجرا می‌شود . 72.4% از روبوت‌ها یک یا دوبار ظاهر شده‌اند . روبوت‌هایی که بیشتر از همه در مجموعه داده‌های ما پدیدار شده را در شکل ۲ می‌بینید .



۵-۱ تاریخچه Bias

توزیع اینکه چند بار یک روبوت مورد استفاده قرار گرفته است (شکل ۳) طی ۱۱ ماه گذشته تغییری نکرده است . بنابراین ما نتایج بدست آمده برای Bias را در آخرین کاوش که تغیری نکرده نشان می‌دهیم



از آنجائیکه اغلب روبوت‌ها یک یا دوبار در مجموعه داده‌ها ظاهر شدند ، امتیاز رده‌بندی آن‌ها در رده میانی قرار گرفته و غیر قابل تشخیص بود . با فقط روبوت‌های با امتیازات تاییدپذیری بالا و پایین را مورد بررسی قرار دادیم .۱۰ روبوت با بیشترین امتاز و ۱۰ روبوت با کمترین امتیاز را در جدول ۲ می‌بینید که N مقدار نمونه است و N_{favor} تعداد دفعاتی است که روبوت مورد تایید قرار گرفته و $N_{disfavor}$ تعداد دفعاتی است که روبوت مورد تایید نبوده است و انحراف استاندارد مطلق [3] برای $\Delta P(r)$ است . انحراف استاندارد مطلق زمانیکه

از Δp برای محاسبه درجه تایید پذیری روبوت در وب استفاده می‌شود واریانس را می‌دهد .
اندازه‌گیری Bias ما نشان می‌دهد که روبوت‌هایی که بسیار مورد تایید قرار گرفته‌اند از موتورهای جستجوگر و سازمان‌های شناخته شده مانند Google و Yahoo و MSN بوده که بیشتر از سایر روبوت‌ها مورد تایید قرار گرفته است ، دقت داشته باشید که در رده‌بندی روبوت‌های تایید نشده ΔP آن‌ها نشان دهنده این نیست که آن‌ها به ندرت در Robots.txt مورد آزمایش ظاهر شده‌اند .
به عبارت دیگر اغلب روبوت‌های تایید نشده گردآوردنده‌های ایمیل ("emailsiphon" , "CherryPicker") و مرورگرهای آفلاین ("Wget" , Webzip") بودند . از لحاظ شخصی معقول به نظر می‌رسد که مدیران وب سایت‌ها از روبوت‌هایی که هدف اصلی آن‌ها جمع‌آوری اطلاعات شخصی است ممانعت بعمل آورند . همچنین معمولاً نمی‌خواهند وب سایت آن‌ها به طور کلی توسط دیگران کپی شود . اگر چه روبوت‌هایی از شرکت‌های شناخته شده نیز هستند که مورد تایید قرار نگرفته اند مانند "MSIEcrawler" و "ia_archiver" . "MSIEcrawler" روبوتی است که در IE جاسازی شده و زمانیکه کاربران IE یک صفحه را Bookmark می‌کنند MSIEcarwler تمام صفحه و لینک‌های مرتب با آن را دانلود می‌کند (لینک ها ، تصاویر ، جاوااسکرپت و ...) .
"ia_archiver" یک کاوشگر از archive.org و Alexa.com است . یک لیست با توضیحات تقضیلی از روبوت‌های شناخته شده را در http//botseer.ist.psu.edu/namedrobots.html می‌توانید ببینید .

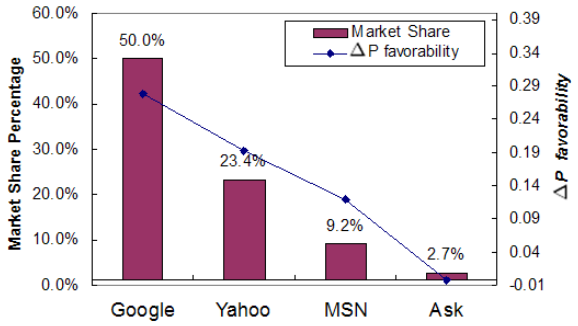
Favored Robots (Sample size $N = 2925$)				
robot name	N_{favor}	N_{disf}	$\Delta P(r)$	σ
google	877	25	0.2913	0.0084
yahoo	631	34	0.2041	0.0075
msn	349	9	0.1162	0.0059
scooter	341	15	0.1104	0.0058
lycos	91	5	0.0294	0.0031
netmechanic	84	10	0.0253	0.0029
htdig	15	3	0.0041	0.0012
teoma	13	3	0.0034	0.0011
oodlebot*	8	0	0.0027	0.0010
momspider	6	0	0.0021	0.0008
Disfavored Robots (Sample size $N = 2925$)				
robot name	N_{favor}	N_{disf}	$\Delta P(r)$	σ
msiecrawler	0	85	-0.0291	0.0031
ia_archiver	7	55	-0.0164	0.0023
cherrypicker	0	37	-0.0126	0.0021
emailsiphon	3	34	-0.0106	0.0019
roverbot	2	27	-0.0085	0.0017
psbot	0	23	-0.0079	0.0016
webzip	0	21	-0.0072	0.0016
wget	1	22	-0.0072	0.0016
linkwalker	2	20	-0.0062	0.0015
asterias	0	18	-0.0062	0.0015

Table 2. Top 10 favored and disfavored robots. σ is the standard deviation of $\Delta P(r)$.

ما همچنین به این نتیجه رسیدیم که bias روبوت‌ها در دومین‌های مختلف تغییر می‌کند . Google همیشه تایید شده ترین روبوت است . روبوت‌های بسیار تایید شده دیگر در دومین مختلف تغییر می‌کند . Yahoo (روبوت Slurt روبوت Yahoo است) و MSN نیز توسط اکثر دومین‌ها مورد تایید قرار گرفته‌اند اما آن سایر روبوت‌ها را مورد تایید قرار نداده‌اند . سایر روبوت‌های رده بالا تایید شده مربوط به کاوشگران کد باز و سازمان‌های معروف است . روبوت‌های مورد تایید قرار نگرفته برای دومین‌های مختلف تغییر می‌کند . اغلب آن‌ها گردآورندگان ایمیل و مرورگرهای آفلاین هستند . تفاوت‌ها ناشی از رفتار مختلف روبوت‌ها در دومین‌های مختلف است .

۲-۵ بازار موتورهای جستجوگر و Bias روبوت

بر اساس مطالعات اهمیت تاثیر قدرتمندتر شدن قدرتمندان ، ما برای شرکت‌های خاصی ارتباط بین Bias روبوت و سهم از بازار موتورهای جستجوگر را محاسبه نمودیم . سهم از بازار Google ، Yahoo ، MSN و Ask در ۱۱ ماه گذشته و نیز ΔP تاییدپذیری که با دو متغیر مستقل برای روبوت‌ها مطرح می‌شود . ضریب همبستگی Pearson Product-moment[15](PMCC) برای دو متغیر اندازه‌ای است برای میل دو متغیر X و Y که در یک مبحث یا ترکیب برای افزایش یا کاهش نسبت به هم اندازه‌گیری می‌شود . در مجموعه داده‌ها ما ضریب Pearson از سهم بازار ۴ شرکت و ΔP(r) روبوت آنها 0.930 با مقدار P<0.001 است . سهم از بازار موتورهای جستجوگر وBias روبوت در سپتامبر ۲۰۰۶ در شکل ۴ نشان داده شده است .



۶- نتیجه

ما یک تحقیق گسترده بر روی bias روبوت‌ها در وب بر روی اطلاعات موثق و تحلیل‌های آماری تعداد زیادی از فایل‌های Robots.txt ارائه دادیم . نتایج نشان‌دهنده این است که روبوت‌های موتورهای جستجوگر محبوب و درگاه‌های اطلاعاتی همچون Google ، Yahoo و MSN اغلب توسط وب‌سایت‌هایی که ما مورد بررسی قرار دادیم مورد تایید قرار گرفته است . این نکته اشاره به موضوع قدرتمند شدن قدرتمندان در موتورهای جستجوگر محبوب نیز دارد . ما همچنین یک تناسب مطمئن بین سهم از بازار موتورهای جستجوگر و Bias روبوت‌ها بیان کردیم . مطالعه ما نشان می‌دهد استفاده از Robots.txt در ۱۱ ماه افزایش داشته است به طوری در کاوش در مرحله اول ۲۶۶۲ و در مرحله آخر ۲۹۲۵ فایل را شامل می‌شود . ما مشاهده کردیم که 46.02% از وب‌سایت های روزنامه‌ها اکنون فایل Robots.txt را استفاده می‌کنند و دومین‌های روزنامه‌ها دومین هایی است که بیشترین پذیرش پروتکل ممانعت از روبوت را دارد . 45.93% از وب‌سایت دانشگاه‌های آمریکایی پروتکل ممانعت از وب‌سایت را پذیرفته‌اند که به طور مشخص از دانشگاه‌های اروپایی (37.8%) و آسیایی (15.4%) بیشتر است . فعالیت‌های آینده ما در آینده بیشتر تقسیم می‌شود به تحلیل مکان‌های جغرافیائی برای بررسی تایید پذیری روبوت در هر کشور .

فعالیت‌های آینده ما بیشتر در مورد تحلیل عمیق‌تر و در مقیاس بزرگتر تنظیمات و رفتارهای روبوت‌ها می‌باشد . ما سایر مقیاس‌های مربوط به bias روبوت را مورد بررسی قرار خواهیم داد . بررسی‌های تجربی ما در زمینه چگونگی تفسیر پروتکل ممانعت از روبوت‌ها در روبوت‌ها خواهد بود .

منابع و مراجع:

[1] Furtune magazine .http://money.cnn.com/magazines/fortune/-fortune500 , 2005 .

[2] M.L.Boonk , D . R . A .d .Groot , F . M . T . Brazier , and A .Oskamp . Agent exclusion on websites . In proceeding of The 4th Workshop on the Law and Eletronic Agent , 2005

[3] S.Chakrabarti , M.van den berg , and B. Dom . Focused crawling : A new approach to topic-specific web resource discovery . In Proc. of the 8th WWW conference,pages 545-562 ,1999

[4] J. Cho , H. Garcia-Molina , and L.page . efficient crawling through url ordering . In proceeding of the 7th International WWW conference , 1998

[5] T.Y Chum . World wide Web robots : An overview . online information review , 23(3): 135-142 , 1999

[6] Dmoz The open directory project . http://dmoz.org , 2005

[7] M.Drott . Indexing aids at corporate Websites: the use of robots.txt and meta tags . information processing and management , 38(2):209-219 , 2002

[8] D.Eichmann . Ethical web agent . Computer Networks and ISDN Systems , 28(1-20:127-136 ,1995

[9] t.joachims .optimizing search engine using clickthrough data . In KDD 02 : proceeding of the eight ACM SIGKDD international Conference on Knowledge discovery and data mining , pages 133-142 New York , NY , USA , 2002 . ACM Press

[10] B.Kelly and I.Peacock . Webwaching Uk web Communities : Final report for the webWatch project . British library research and innovation report , 1999

[11] M.Kendall . Rank Correlation Method .hafner , 1955

[12] M. Koster A method for web robots control . In the internet draft , The internet Engineering Task Force (IETF) , 1996

[13] R.L.Ott and M.T . Longnecker . an introduction to statistical method and data analysis . Duxbery press ; 5edition 2000

[14] G.Pant , P. Srinivasan , and F.menczer . crawling the web,chapter web dynamic . springer-verlag , 2004

[15] R.R Sokal and F.J.Rohlf . biometry . Freeman new york , 2001

[16] Y.sun , Z.Zhuang and C.L.Giles . A large Scale study for robots.txt In WWW '07 : proceeding of the 16th international conference on world wide web , page 1123-1124 New York , NY USA ,2007 .ACM Press .