مجله علمی-پژوهشی رایانش نرم و فن‌آوری اطلاعات

جلد ۵، شماره ۳، پاییز سال ۱۳۹۵ شاپا: ۱۰۰۶-۲۳۸۳

ارائه روشی بر مبنای پیوند جهت بهبود تشخیص صفحات فریب‌آمیز در گراف وب فارسی

مرضیه پارویی*[۱]، علی‌محمد زارع بیدکی[۲]

*۱- نویسنده مسئول: دانشجوی کارشناسی‌ارشد، دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد،یزد، ایران،parooie@stu.yazd.ac.ir

۲- دانشیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، یزد، ایران،alizareh@yazd.ac.ir

چکیده- امروزه با افزایش چشمگیر استفاده از اینترنت و همچنین رشد بسیار فزاینده صفحات وب، استفاده از موتورهای جستجو اهمیت بیشتری پیدا کرده است. در نتیجه بعضی از افراد برای بدست آوردن مخاطب بیشتر و افزایش سود ناشی از آن تلاش می‌کنند موتورهای جستجو را گمراه کنند و رتبه صفحات مورد نظر خود را با روش‌های نامشروع بالا ببرند. شناسایی این صفحات می‌تواند نقشی اساسی در بهبود عملکرد موتورهای جستجو و بالا بردن اطمینان کاربران به آنها گردد. نظر به اهمیت کشف صفحات وب فریب‌آمیز، در این مقاله روشی جدید بر مبنای اطلاعات پیوندها جهت شناسایی صفحات فریب‌آمیز در گراف وب فارسی ارائه می‌گردد. در این روش با بهره‌گیری از اطلاعات پیوندها، ابتدا توده‌های فریب‌آمیز شناسایی می‌شود و سپس امتیاز منفی آنها در کل گراف انتشار می‌یابد. برای بررسی صحت عملکرد الگوریتم ارائه شده، این روش بر روی داده‌های موتور جستجوی فارسی پارسی‌جو پیاده‌سازی شده است و نتایج ارزیابی‌های صورت گرفته بهبودی برابر با ۲۱,۲٪ را در فاکتور دقت نشان می‌دهد.

واژه‌های کلیدی: موتور جستجو، وب‌فریبکارانه، رتبه‌بندی.

۱- مقدمه

مالکین سایت‌های اینترنتی همواره علاقمندند تا در نتایج تولیدشده از موتورهای جستجوی[۱] معروف در رده‌های بالا قرارگیرند. علت این موضوع را می‌توان در بازخورد غالب کاربران اینترنت جستجو نمود که در ۸۵٪ موارد اگر نتیجه مورد انتظار خود را در ۱۰ پیوند اول نتایج یا حداکثر در پیوندهای صفحه اول نتایج نمایش داده شده نیابند به صفحات و نتایج بعدی مراجعه ننموده و عبارت پرس‌وجوی[۲] خود را عوض می‌کنند. به همین دلیل و در راستای افزایش کارآمدی و درآمد، همواره موتورهای جستجو در تلاش‌اند تا سایت‌های مرتبط با پرس‌وجوی کاربر را در رتبه‌های بالاتر قرار دهند. همین تلاش موتورهای جستجو می‌تواند توسط سودجویان مورد سوء استفاده قرار گرفته و پدیده‌ای تحت عنوان صفحه وب فریب‌آمیز را ایجاد نماید.

واژه‌ی وب فریب‌آمیز[۳] به ترفندهایی گفته می‌شود که برای فریب دادن الگوریتم‌های رتبه‌بندی موتورهای جستجو و تغییر نتیجه جستجو به سمت صفحات وب هدف استفاده می‌شوند[۲].

با وجودی که استفاده از صفحات فریب‌آمیز همیشه با مقاصد تجاری و سوء استفاده نیست اما از آنجایی که در این قبیل صفحات اطلاعاتی متفاوت با اطلاعات مد نظر کاربر ارائه می‌گردد و هدف از ایجاد آنها تنها اعمال نفوذ در الگوریتم‌های رتبه‌بندی[۴] موتورهای جستجو است، در نتیجه همه آنها را می‌توان تحت عنوان صفحات فریب‌آمیز دسته‌بندی نمود. هدف اصلی از صفحات فریب‌آمیز افزایش رتبه کاذب صفحات اینترنتی از طریق فریب الگوریتم‌های موتورهای جستجو می‌باشد. در این راستا آنچه اهمیت بیشتری دارد، دستیابی هرچه سریعتر به جایگاه بالاتر در میان ده سایت اول نتایج جستجو با پرس‌وجوهای متفاوت می باشد.

امروزه با گسترده‌تر شدن واژگان موجود در فضای مجازی و نیاز همگان به جستجو برای دستیابی به انبوه اطلاعات در وب، نقش موتورهای جستجو و الگوریتم‌های رتبه‌بندی مورد استفاده در آنها بسیار اهمیت یافته است. به همین منظور نگاه افراد سودجو یا منفعت طلب نیز به این مقوله بیشتر گردیده است. همانند نامه‌های فریب‌آمیز که در حوزه نامه‌های الکترونیکی[۵] موجب اتلاف فراوان

[1] Search engines
[2] Query
[3] Spam
[4] Ranking
[5] Email

www.SID.ir

Archive of SID

وقت و انرژی می‌گردند، صفحات وب فریب‌آمیز نیز موجب کاهش کیفیت نتایج جستجو و اتلاف وقت کاربران می‌شوند. با افزایش تعداد این صفحات فریب‌آمیز بالاجبار تعداد صفحاتی که می‌بایست توسط خزشگرها[1] مورد بررسی قرار گرفته و توسط شاخص‌گذارها مرتب گردند به تناسب افزایش می‌یابد. این موضوع علاوه بر اتلاف منابع موتورهای جستجو موجب افزایش زمان جستجو در پاسخ به پرس‌وجوی کاربر نیز می‌گردد.

تاکنون روش‌های بسیاری برای مقابله با پدیده وب فریب‌آمیز ایجاد شده‌اند. با این حال جذابیت و سود اقتصادی این پدیده موجب شده است تا همواره از یک طرف محققین روش‌های جدیدی برای مقابله ارائه نموده و از طرف دیگر فریبگرها روش‌هایی برای عبور از محدودیت‌ها ارائه نمایند. در نتیجه تاکنون هیچ روشی که بتواند به صورت نظری، غیرقابل نفوذ بودن خود را تضمین نماید ارائه نشده است.

موتورهای جستجو معمولاً صفحات وب را براساس دو فاکتور زیر رتبه‌بندی می‌کنند [۲]:

۱. ارتباط پرس‌وجو و صفحه وب: این ارتباط معمولاً از طریق اندازه‌گیری شباهت درخواست ورودی و متن صفحات وب بدست می‌آید.

۲. محبوبیت صفحه وب: منظور از محبوبیت صفحه در واقع میزان اهمیت آن صفحه مستقل از پرس‌وجو است. این اهمیت مثلاً ممکن است با توجه به تعداد ارجاعاتی که از صفحات دیگر به این صفحه شده است اندازه‌گیری شود.

فریب‌گرها نیز سعی می‌کنند با استفاده از ترفندهای مختلف این دو فاکتور را به صورت غیرواقعی افزایش دهند.

ایجاد پیوندهای زیاد به منظور بالا بردن امتیاز صفحه هدف، یکی از روش‌های ایجاد صفحات وب فریب‌آمیز است. در این روش فریبگر در گراف وب اختلال ایجاد کرده و سعی دارد با استفاده از تعداد صفحات زیادی که پیوندهای بسیاری به یکدیگر دارند، الگوریتم رتبه‌بندی موتور جستجو را فریب داده و رتبه صفحات هدف را بالا ببرد [۲].

مشاهدات این پژوهش نشان داده، در زبان فارسی و بخصوص در کشور ما ایران، تعداد وبلاگ‌ها بسیار زیاد است. دلیل این مهم را می‌توان هزینه پایین تولید محتوا و ایجاد وبلاگ در مقابل وب‌سایت دانست. این مسئله به راحتی زمینه کاری فریبگرها را فراهم می‌کند تا بتوانند با هزینه کم اقدام به تولید صفحات فریب‌آمیز کرده و به اهداف تجاری، سیاسی و غیره برسند.

رتبه‌بندی موتور جستجو باید ماهیت صفحات (وبلاگ یا وب‌سایت) را مدنظر قرار داده و حتی در صورتی که صفحه وبلاگ از دید الگوریتم رتبه‌بندی مورد استفاده از امتیاز بالایی برخوردار باشد، رتبه‌بندی را به گونه‌ای انجام دهد که صفحات ابتدایی نتایج، مالامال از صفحات وبلاگ نشود. بدین منظور لازم است در رتبه بندی، وزن‌های مختلفی برای صفحات سایت‌ها و وبلاگ‌ها در نظر گرفته شود.

روشی که برای شناسایی صفحات وب فریب‌آمیز در این مقاله ارائه شده است مبتنی بر اطلاعات پیوند است. براساس بررسی‌های انجام شده در این پژوهش مشخص شد که هر صفحه فریب‌آمیز متعلق به یک میزبان فریب‌آمیز است. با توجه به این مسئله شناسایی یک میزبان فریب‌آمیز (به جهت اینکه میزبان مجموعه‌ای از چندین صفحه است) بسیار کارآمدتر از یک صفحه وب فریب‌آمیز است. بنابراین شناسایی بر روی میزبان‌ها صورت گرفته است.

این روش شامل پنج فاز است در انتهای این روش برای هر میزبان یک ضریب محاسبه می‌شود که میزان فریب‌آمیز بودن آن میزبان را براساس پیوندهای آن میزبان نشان می‌دهد.

در ادامه این بخش به بیان یک‌سری از الگوریتم‌های ارائه شده در این زمینه می‌پردازیم. در بخش دوم الگوریتم پیشنهادی را بیان می‌کنیم. سپس در بخش سوم به ارزیابی الگوریتم پیشنهادی و بررسی نتایج می‌پردازیم. درنهایت در بخش چهارم نتیجه‌گیری و کارهای آینده را بیان می‌کنیم.

۱-۱- الگوریتم HostRank

در نگاه کلی الگوریتم‌های مبتنی بر پیوند، به دو دسته وابسته به پرس‌وجو و مستقل از پرس‌وجو تقسیم می‌شوند. در روش‌های مستقل از پرس‌وجو مانند HostRank[۳] و PageRank [۴] رتبه بندی به صورت برون‌خط[2] و با استفاده از کل گراف وب انجام می‌شود. در نتیجه به ازای هر پرس وجو رتبه هر صفحه ثابت است. اما در روش وابسته به پرس وجو یا حساس به موضوع، رتبه‌بندی در گرافِ شامل مجموعه صفحات مرتبط با پرس وجوی کاربر انجام می‌شود.

در الگوریتم HostRank برای محاسبه اهمیت یک صفحه از ساختار پیوندی وب و هم‌چنین از ساختار سلسله‌مراتبی آن استفاده می‌شود. در این روش از یک ساختار پیوندی جدید وب استفاده می‌

[1] Crawler

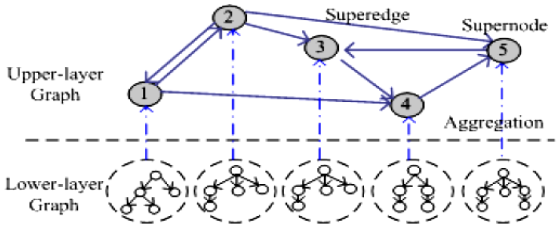
[2] offline

www.SID.ir

شود که شامل دو لایه است. الگوریتم HostRank برخلاف الگوریتم PageRank که وب را به صورت یکنواخت نگاه می‌کند، وب را به صورت لایه لایه می‌بیند.

در این روش یک مدل قدم زدن تصادفی جدید ارائه شده که بر مبنای آن فرض می‌شود کاربری که به دنبال اطلاعات است با شروع از لایه بالا و یا پرش تصادفی به یکی دیگر از گره‌های لایه بالاتر و یا به زیرپیوندهای سلسله مراتبی لایه پایین‌تر می‌رود.

شکل ۱ ساختار سلسله مراتبی در گراف وب را نمایش می‌دهد. بر اساس این مدل راه رفتن تصادفی، یک الگوریتم تجزیه و تحلیل پیوند جدید به نام رتبه سلسله‌مراتبی برای محاسبه اهمیت صفحات وب ارائه شده است. این الگوریتم اجازه می‌دهد امتیاز یک گره برتر[۱] که در لایه بالایی گراف قرار دارد به گره‌های لایه پایین منتشر شود. الگوریتم HostRank نشان می‌دهد که تراکم توزیع پیوندها در لایه بالایی گراف بیشتر از بقیه گراف است. این مسئله باعث می‌شود که مشکل "غنی‌تر شدن اغنیا"[۲] حل شود.



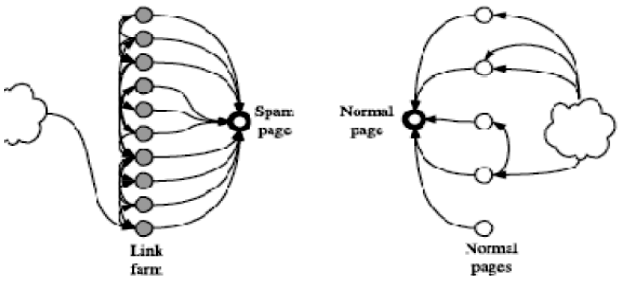
شکل ۱: ساختار سلسله مراتبی در گراف وب[۳]

الگوریتم‌های مبتنی بر پیوند، عمل رتبه بندی را بر مبنای گراف مسطح[۳] انجام می‌دهند و ساختار سلسله مراتبی[۴] وب را نادیده می‌گیرند. آنها از دو مشکل عمده رنج می‌برند: پراکندگی زیاد صفحات وب و سوگیری نسبت به صفحات قدیمی که باعث می‌شود صفحات جدید با درجه ورودی کم از رتبه پایینی برخوردار شوند. الگوریتم HostRank که برای فائق آمدن بر دو مشکل فوق ارائه شده است. هر دوی ساختار سلسله مراتبی و اتصالی وب را در رتبه‌بندی لحاظ می‌کند. در این روش صفحات ابتدا در یک ساختار سلسله مراتبیِ دایرکتوری، میزبان و یا دامنه[۵] که گره برتر نامیده می‌شود قرار داده می‌شوند و عمل آنالیز اتصال بر روی گراف به‌دست آمده، انجام می‌شود. سپس درجه‌ی اهمیت (ارزش) محاسبه شده‌ی هر گره بین صفحاتی که آن گره شامل می‌شود توسط ساختار سلسله مراتبی توزیع می‌شود.

[1] Super node
[2] Rich-get-richer
[3] Flat
[4] Hierarchical
[5] Domain

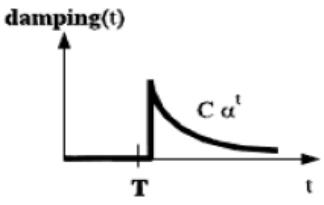
۲-۱- الگوریتم TruncatedPageRank

اساساً این الگوریتم برای شناسایی مزرعه فریب‌آمیز[۶] به کار می‌رود و این اصطلاح برای سایت‌هایی به کار می‌رود که به صورت توده‌ای به هم پیوند می‌دهند. نمونه ای از مزرعه فریب‌آمیز در شکل ۲ نمایش داده شده است.



شکل۲: مزرعه فریب‌آمیز[۵]

در الگوریتم TruncatedPageRank[۶] به منظور کاهش رتبه صفحاتی که رتبه آنها به واسطه پیوندهای مستقیم افزایش یافته است، تابع damping به صورت زیر تعریف می‌گردد. این رابطه در شکل ۳ نمایش داده شده است.



$$\text{damping}(t) = \begin{cases} 0 & t \le T \\ C\alpha^t & t > T \end{cases}$$

شکل ۳: damping factor در الگوریتم TruncatedPageRank[۶]

در این رابطه C یک ثابت نرمال‌سازی و α فاکتور میرایی نام دارد. ثابت نرمال‌سازی باید طوری انتخاب شود که $\sum_{t=0}^{\infty} \text{damping}(t) = 1$ این رابطه برقرار باشد. بنابراین $C = \frac{1-\alpha}{\alpha^{T+1}}$ خواهد بود.

این تابع صفحاتی که رتبه‌شان را از پیوندهای خیلی کوتاه و از

[6] Link farm

www.SID.ir

طریق صفحات نزدیک به خود به دست آورده‌اند، جریمه می‌کند. مکانیزم این الگوریتم به صورتی است که اگر یک گره اکثر امتیاز خود را از گره‌های همسایه خود دریافت کند و در حقیقت امتیاز دریافتی یک گره به صورت توزیع شده از کل وب نباشد، این گره مشکوک به فریب‌آمیز است و امتیاز بالایی را به خود اختصاص می‌دهد. این روش بسیار مشابه PageRank است با این تفاوت که پشتیبان‌های خیلی نزدیک، سهم خود را در افزایش رتبه صفحه مورد نظر از دست می‌دهند.

شبه کد این الگوریتم در شکل ۴ نمایش داده شده است. فاکتور T در این شبه کد فاصله همسایگی را نمایش می‌دهد. لازم به ذکر است که مفهوم همسایگی در اینجا تعداد کلیک برای رسیدن از مبدا به مقصد است.

باید توجه داشت در این الگوریتم در صورتی که $T = -1$ باشد، در واقع همان PageRank معمولی محاسبه می‌شود. مقدار PageRank و TruncatedPageRank برای یک پیکره آموزشی در شکل ۵ نمایش داده شده است. همانطور که ملاحظه می‌شود این دو رتبه‌بندی همبستگی نزدیکی با هم دارند. اما این همبستگی با افزایش متغیر T کاهش می‌یابد.

```
Require: N: number of nodes, 0 < α < 1: damping factor,
    T≥ −1: distance for truncation
 1: for i : 1 ... N do {Initialization}
 2:    R[i] ← (1 − α)/((α^(T+1))N)
 3:    if T≥ 0 then
 4:       Score[i] ← 0
 5:    else {Calculate normal PageRank}
 6:       Score[i] ← R[i]
 7:    end if
 8: end for
 9: distance = 1
10: while not converged do
11:    Aux ← 0
12:    for src : 1 ... N do {Follow links in the graph}
13:       for all link from src to dest do
14:          Aux[dest] ← Aux[dest] + R[src]/outdegree(src)
15:       end for
16:    end for
17:    for i : 1 ... N do {Apply damping factor α}
18:       R[i] ← Aux[i] ×α
19:       if distance > T then {Add to ranking value}
20:          Score[i] ← Score[i] + R[i]
21:       end if
22:    end for
23:    distance = distance +1
24: end while
25: return Score
```

شکل ۴:شبه کد الگوریتم TruncatedPageRank

۳-۱- روش AntiTrustRank

این روش دقیقاً عکس الگوریتم TrustRank [۶] است و امتیاز منفی را خلاف جهت الگوریتم TrustRank در گراف انتشار می‌دهد. در این روش ابتدا باید یک مجموعه اولیه از صفحات فریب‌آمیز انتخاب شود. این مجموعه باید به گونه‌ای انتخاب شود که خصوصیات زیر را داشته باشد:

اولاً این مجموعه باید شامل صفحاتی باشد که انتشار AntiTrustRank[۷] با هزینه کمتری انجام شود.

دوما این مجموعه باید شامل آن دسته از صفحات فریب‌آمیز باشد که PageRank آنها مقدار بالاتری دارد. زیرا در موتورهای جستجو تشخیص صفحات فریب‌آمیز که در رتبه‌های بالاتری قرار می‌گیرند از اهمیت بسزایی برخوردار است.

انتخاب صفحات فریب‌آمیز با رتبه بالا منجر به انتخاب صفحاتی می‌شود که تعداد پیوندهای ورودی زیادی دارند. بنابراین انتشار AntiTrustRank هم با هزینه کمتری انجام خواهد پذیرفت. بنابراین نکته اول هم خود به خود تحقق می‌یابد. مراحل الگوریتم AntiTrustRank بدین شرح است:

۱. ابتدا مجموعه اولیه صفحات فریب‌آمیز با رتبه بالا انتخاب می‌شود.

۲. ترانهاده ماتریس گراف وب محاسبه می‌شود.

۳. الگوریتم PageRank به ازاء مجموعه تولید شده در مرحله ۱ اجرا می‌شود.

۴. تمام صفحات را براساس رتبه‌های بدست آمده در مرحله قبل، به صورت نزولی مرتب کرده یا اینکه تمام صفحاتی که رتبه آنها از یک مقدار آستانه‌ای بیشتر است به عنوان فریب‌آمیز معرفی می‌شوند.

۲- الگوریتم پیشنهادی

این الگوریتم از ۵ فاز تشکیل شده است. نمودار فازهای این الگوریتم درشکل ۶ نمایش داده شده است.

۲-۱- پیاده‌سازی الگوریتم HostRank

در فاز اول ابتدا الگوریتم HostRank پیاده‌سازی شد. جزئیات این الگوریتم در مقدمه بیان شده است. با پیاده‌سازی این الگوریتم در فاز اول برای هر میزبان یک امتیاز محاسبه شد.

Archive of SID

www.SID.ir

۲-۲- پیاده‌سازی الگوریتم TruncatedHostRank

با بررسی گراف وب فارسی و مشاهده مکانیزم فریبگرها برای بالا بردن امتیاز یک صفحه به صورت غیرقانونی، این نتیجه حاصل شد که اکثر فریبگرها از مکانیزم مشابهی برای بالا بردن امتیاز صفحات خود استفاده می‌نمایند. به این صورت که برای بالابردن امتیاز یک صفحه اقدام به تولید یک توده یا مزرعه فریب‌آمیز می‌کنند. این مسئله به این معناست که تعداد زیادی صفحه مشابه ایجاد کرده و به هم پیوند می‌دهند و به این صورت باعث اختلال در الگوریتم رتبه‌بندی موتور جستجو شده و امتیاز صفحه مورد نظر را بالا می-آورند. از این رو در مرحله اول اقدام به شناسایی این توده‌های فریب‌آمیز شد.

با توجه به اینکه در میان صفحات وب فریب‌آمیز، پیوندهای کوتاه زیادی وجود دارد. برای شناسایی این توده‌ها، الگوریتم TruncatedPageRank [۴] مد نظر قرار گرفت اما با توجه به این نکته که یک میزبان مجموعه‌ای از چندین صفحه است و شناسایی یک میزبان به عنوان فریب‌آمیز بسیار کارآمدتر از شناسایی یک صفحه به عنوان فریب‌آمیز است، این الگوریتم بر روی میزبان‌ها و در حقیقت به صورت TruncatedHostRank پیاده‌سازی شد تا میزبان‌هایی که به صورت توده‌ای به هم پیوند می‌دهند شناسایی شود. اساساً این الگوریتم برای شناسایی مزرعه فریب‌آمیز[1] به کار می‌رود و این اصطلاح برای سایت‌هایی به کار می‌رود که به صورت توده‌ای به هم پیوند می‌دهند.

مکانیزم این الگوریتم به صورتی است که اگر یک گره اکثر امتیاز خود را از گره‌های همسایه خود دریافت کند و در حقیقت امتیاز دریافتی یک گره به صورت توزیع شده از کل وب نباشد، این گره مشکوک به فریب‌آمیز است و امتیاز بالایی را به خود اختصاص می‌دهد. این روش بسیار مشابه PageRank است با این تفاوت که پشتیبان‌های خیلی نزدیک، سهم خود را در افزایش رتبه صفحه مورد نظر از دست می‌دهند.

در نهایت پس از انتشار امتیازها، گره‌هایی که نسبت TruncatedHostRank به HostRank آنها از یک حد آستانه‌ای بالاتر باشد مشکوک هستند. در حقیقت این گره‌ها اکثر امتیاز خود را از گره‌های همسایه خود دریافت کرده‌اند که این نسبت در آنها بالا رفته است. همچنین گره‌هایی که این نسبت در آنها با هم برابر است رفتار مشابهی داشته‌اند و یک توده را تشکیل می‌دهند. این الگوریتم با $T = 1$ تا $T = 4$ پیاده‌سازی شد و برای این گراف بهترین جواب در $T = 2$ حاصل شد.

در اینجا تنها از فاکتور نسبت TruncatedHostRank به HostRank به عنوان معیاری برای شناسایی درجه فریب‌آمیز بودن صفحه استفاده شده است.

با پیاده‌سازی این الگوریتم تعداد زیادی از توده‌های گراف شناسایی شد ولی با توجه به وسعت گراف وب و ترفندهای زیاد فریبکاران هنوز گره‌های بسیاری برچسب نخورده باقی ماند. زیرا تمام صفحات فریب‌آمیز به صورت توده‌ای عمل نمی‌کنند و در نتیجه نمی‌توان تمام صفحات وب فریب‌آمیز را با استفاده از این الگوریتم شناسایی کرد. در نتیجه برای گسترش کار و شناسایی دیگر صفحات فریب‌آمیز از الگوریتم AntiTrustRank استفاده شد.

۲-۳- فاز سوم: پیاده‌سازی الگوریتم AntiTrustRank

برای گسترش کار از این توده‌های شناسایی شده به عنوان نقطه شروع[2] استفاده شد تا بتوان از این طریق، امتیاز منفی آنها را در کل گراف انتشار داد.

یکی از ایرادات اساسی که به الگوریتم AntiTrustRank وارد است انتخاب نقاط شروع است که کاملا به صورت دستی و تجربی صورت می‌پذیرد. اما در این مقاله همانطور که در قسمت پیش به طور کامل شرح داده شد انتخاب نقاط شروع به صورت اتوماتیک انجام می‌شود. توده‌های به دست آمده در مرحله قبل به عنوان نقاط شروع این مرحله استفاده می‌شود. بدین ترتیب این گره‌ها به عنوان نقطه شروع به الگوریتم AntiTrustRank داده شد.

مراحل الگوریتم AntiTrustRank به این صورت است:

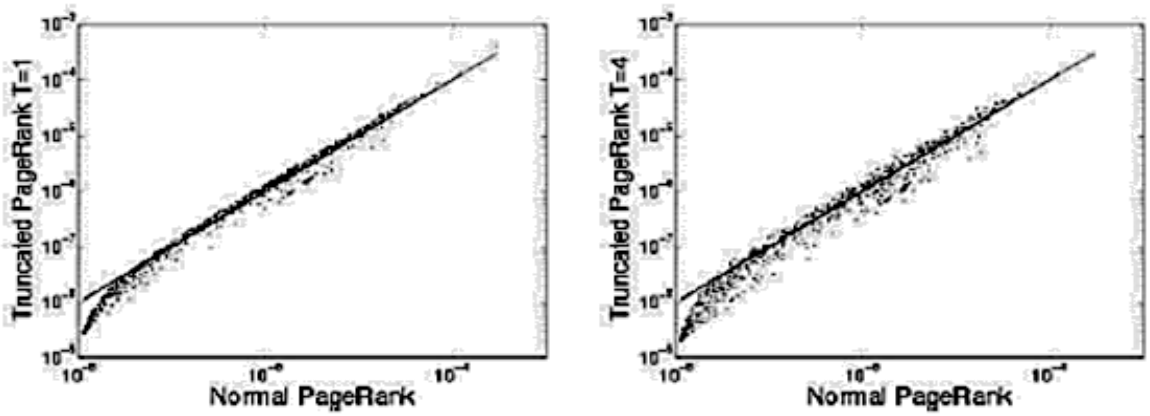
در ابتدا یک مجموعه اولیه از صفحات فریب‌آمیز با رتبه بالا ایجاد کرده و به عنوان نقطه شروع از آن استفاده می‌شود. سپس ترانهاده ماتریس گراف وب محاسبه می‌شود و الگوریتم PageRank به ازای نقاط شروع اجرا می‌شود. سپس تمام صفحات براساس رتبه‌های بدست آمده به صورت نزولی مرتب می‌شوند. یا تمام صفحاتی که رتبه آن‌ها از یک مقدار آستانه‌ای بیشتر است به عنوان فریب‌آمیز معرفی می‌شوند.

[1] Link farm

[2] seed

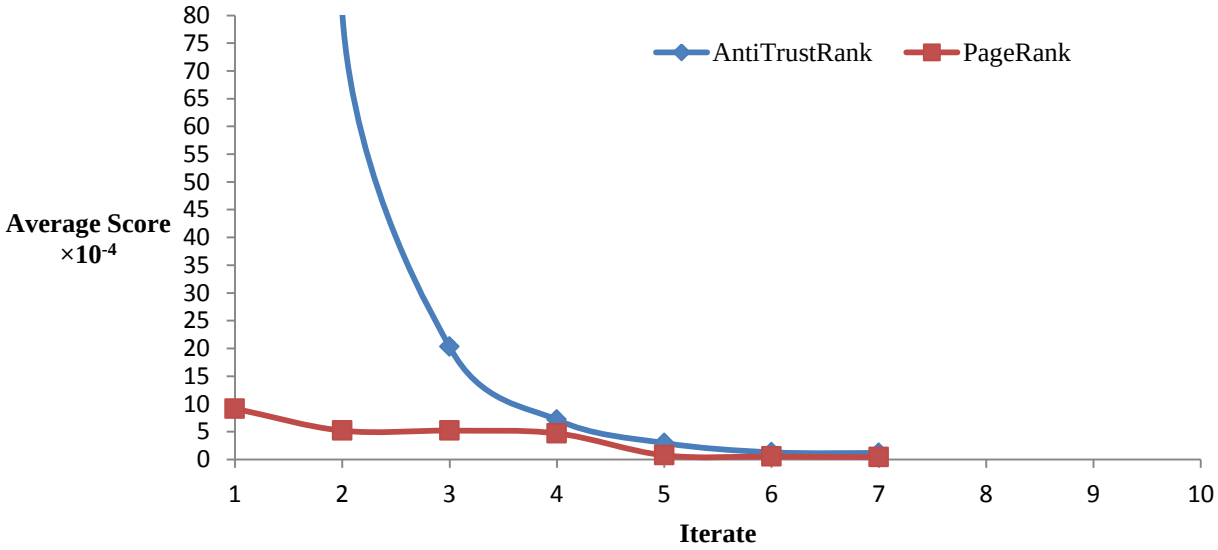
www.SID.ir

Archive of SID



شکل ۵: نمایش همبستگی PageRank و TruncatedPageRank[۶]

شکل ۶: مراحل الگوریتم پیشنهادی



شکل ۷: مقایسه انتشار در PageRank و AntiTrustRank

www.SID.ir

روند این الگوریتم بدین صورت است که نقاط شروع امتیاز منفی می‌گیرند و در طی یک عملیات انتشار هر گره امتیاز خود را به گره‌های والد خود می‌دهند. در حقیقت عملیات انتشار در این الگوریتم دقیقاً عکس عملیات انتشار در الگوریتم PageRank است. منطق این الگوریتم بیان می‌کند، صفحه‌ای که به یک صفحه وب فریب‌آمیز پیوند می‌دهد فریب‌آمیز است و تا حدی باید امتیاز منفی بگیرد. اما میزان این امتیاز منفی را تعداد پیوندهایی که به صفحات وب فریب‌آمیز داده مشخص می‌کند.

پس از پیاده‌سازی و اعمال این الگوریتم امتیازات بدست آمده تقریباً برای تمام گره‌ها یکسان بود و اطلاعی در مورد فریب‌آمیز بودن یا نبودن یک گره نمی‌داد. در حقیقت امتیاز منفی به درستی انتشار داده نشده بود و این مسئله به این علت ایجاد می‌شد که گراف از پیوستگی کامل برخوردار نبود و در نتیجه عملیات انتشار به درستی صورت نمی‌گرفت و باعث می‌شد که هر چقدر هم امتیاز اولیه نقاط شروع بالا بود، باز هم در طی انتشار و در همان مرحله اول این امتیاز شکسته شود و بقیه مراحل انتشار درست مشابه قبل ادامه یابد و امتیاز منفی انتشار داده نشود.

در نتیجه با توجه به این اصل (که در حقیقت همان منطق AntiTrustRank است) که یک صفحه که به صفحه وب فریب‌آمیز پیوند دهد، فریب‌آمیز است و باید امتیازی بیشتر از فرزند خود دریافت کند، روند این الگوریتم به این صورت تغییر داده شد که در طی عملیات انتشار، امتیازی که یک گره منتشر می‌کند کاهش نیابد. بلکه این امتیاز مقداری افزایش یابد و در حقیقت امتیازی که به گره والد منتقل می‌شود بیشتر از گره فرزند باشد در هر مرحله انتشار، امتیاز اولیه گره دوباره به آن اضافه می‌شود این مسئله باعث می‌شود امتیاز منفی صفحات فریب‌آمیز در طی عملیات انتشار از بین نرود و در هر مرحله انتشار این امتیاز دوباره در گراف تزریق شود.

شکل ۷ انتشار در الگوریتم PageRank و AntiTrustRank را مقایسه می‌کند. این شکل میانگین امتیازات یک میلیون گره را در مراحل مختلف انتشار نمایش می‌دهد.

همانطور که در شکل مشاهده می‌کنید امتیاز اولیه AntiTrustRank، فقط در دو مرحله ابتدایی انتشار منتقل می‌شود و در مراحل بعدی میانگین امتیازات به شدت افت می‌کند تا اینکه در نهایت پس از چند مرحله انتشار دقیقاً مانند PageRank ادامه می‌یابد در نتیجه نمی‌توان با این الگوریتم امتیاز منفی را به درستی انتشار داد اما با استفاده از روش ارائه شده امتیاز کاملاً در این گراف انتشار می‌یابد.

در این مرحله مشکلی ایجاد شد که در وب، گره‌های خوبی بودند که در عین حال که به گره‌های خوب زیادی پیوند داده بودند در این میان به یک یا چند صفحه وب فریب‌آمیز هم پیوند داده بودند و این پیوندها باعث شده بود که امتیاز منفی بالایی دریافت کنند و این صحیح نبود.

برای حل این مشکل، امتیازی که از طریق انتشار به یک گره داده می‌شد بر درجه خروجی همان گره تقسیم شد و این مسئله باعث شد که گره‌ای که به تعداد زیادی از گره‌ها پیوند داده است و تعداد کمی از این پیوندها به صفحات وب فریب‌آمیز بوده است، به همان میزان امتیاز منفی کمتری دریافت کند و بدین صورت روند امتیازدهی تصحیح شد. در واقع با این روش برای هر گره میزان تأثیر آن گره در فریبکار جلوه دادن گره‌های همسایه‌اش در نظر گرفته می‌شود.

این روند باعث شد میزبان‌هایی با تعداد صفحات زیاد که به یک و یا تعداد کمی صفحه فریب‌آمیز پیوند داده‌اند، به عنوان فریب‌آمیز شناخته نشوند.

سپس امتیازی که هر یک از گره‌ها از طریق انتشار دریافت کردند به صورت عددی بین صفر و یک نرمال گردید و این عدد به عنوان یک ضریب در امتیاز کلی، که برای یک صفحه محاسبه شده، ضرب می‌شود.

در شکل ۸ شبه کد این الگوریتم آورده شده است. همانطور که در این شکل مشاهده می‌شود در ابتدای برنامه ابتدا برای هر میزبان یک امتیاز HostRank محاسبه شده و این امتیاز را ذخیره می‌کنیم. سپس برای هر میزبان یک امتیاز TruncatedHostRank محاسبه کرده و نسبت این دو را می‌یابیم و در صورتی که از یک آستانه‌ای بالاتر بود جزو صفحات فریب‌آمیز محسوب می‌شود. سپس وارد عملیات انتشار می‌شویم صفحاتی که در مرحله قبل به عنوان فریب‌آمیز شناسایی شده‌اند را به عنوان نقطه شروع به الگوریتم داده و امتیاز منفی آن‌ها را انتشارمی‌دهیم.

۳- ارزیابی

برای سنجش نتیجه اجرای این الگوریتم از داده‌های پارسی‌جو [۹] که شامل یک مجموعه صفحات خزش شده و ایندکس شده بود استفاده شد. این مجموعه از ۱۱ میلیون میزبان تشکیل شده است. برای ارزیابی این الگوریتم از ۱۵۰ پرس‌وجو با موضوعات مختلف استفاده شده که به صورت تصادفی از بین ۱۵۰۰ پرس‌وجو انتخاب شدند.

Archive of SID

www.SID.ir

برای سنجش نتایج از فاکتور p@n استفاده شد. فاکتور p@n به این صورت تعریف می‌شود: نسبت تعداد سندهای مرتبط در n نتیجه بالا به کل سندهای بازیابی شده.

نتایج در دو حالت، بدون الگوریتم شناسایی صفحات وب فریب‌آمیز و با وجود الگوریتم شناسایی صفحات وب فریب‌آمیز محاسبه شد که در نمودار شکل ۹ نمایش داده شده است. نمودار شکل ۹ نتایج این الگوریتم را نمایش می‌دهد. روند کلی نمودار p@n در حالت استاندارد نزولی است. اما در صورتی که فرایند بازیابی تحت تأثیر صفحات فریب‌آمیز باشد، ممکن است در نقاطی از این نمودار افت‌های محسوسی دیده شود. مطابق نمودار شماره یک (به رنگ آبی) که روند تغیرات p@n موتور جستجو را به صورت خالص نشان می‌دهد. در ردیف‌های دو و سه و هفت که بیشترین تعداد صفحات فریب‌آمیز ظاهر می‌شده‌اند دقت سیستم دارای افت‌های چشمگیر بوده است ولی در نمودار شماره دو (به رنگ قرمز) که p@n موتور جستجو را پس از کشف صفحات فریب‌آمیز نمایش می‌دهد، این افت‌های ناگهانی برطرف شده و نمودار نشانی از تاثیرات چشمگیر صفحات فریب‌آمیز را در کاهش p@n نشان نمی‌دهد. اما در ردیف‌های هفت تا ده نمودار دچار افت شدیدی شده است و این به این دلیل است که این ردیف‌ها حاوی صفحات فریب‌آمیز نبوده‌اند و در نتیجه اعمال الگوریتم روی آنها تاثیر چندانی نداشته و مقادیر p@n تقریبا با مقدیر p@n قبل از اعمال الگوریتم برابر است.

۴- نتیجه‌گیری

رتبه‌بندی صفحات وب یکی از وظایف مهم و اصلی هر موتور جستجو است. به علت فرصت‌های شغلی کلانی که صفحات وب با رتبه بالا ایجاد می‌کنند افراد تلاش می‌کنند تا با روش‌های مختلف رتبه صفحات خود را بالا ببرند. از آنجایی که اکثر کاربران فراتر از دو صفحه نخست نتایج جستجو نمی‌روند حضور یک صفحه در نتایج بالای موتورهای جستجو به معنای بازدید کننده بیشتر و هم‌چنین به معنای درآمد بیشتر است.

نظر به اهمیت کشف صفحات وب فریب‌آمیز در این مقاله یک روش جدید بر مبنای اطلاعات پیوند یک صفحه ارائه شده است. که با بهره‌گیری از ویژگی‌های پیوندهای یک میزبان، میزبان‌های فریب‌آمیز را شناسایی می‌کند.

این روش شامل ۵ فاز است و بر روی داده‌های موتور جستجوی فارسی پارسی‌جو پیاده‌سازی شده است و نتایج ارزیابی‌های صورت گرفته، بهبودی برابر با ۲۱,۲٪ را در فاکتور p@n نشان داده است.

در این مقاله فقط از اطلاعات پیوند یک میزبان استفاده شده است. در حقیقت از اطلاعات متن یک صفحه و بازخورد کاربر هیچ استفاده‌ای نشده است.

استفاده از ویژگی‌هایی مانند پرس‌وجو و کلیک کاربر به عنوان کارهای آینده باقی خواهد ماند. علاوه بر این ارائه روشی جهت شناسایی دامنه‌های[1] فریب‌آمیز و شناسایی الگوی خودکار تولید محتوا برای مثال وجود برچسب‌های متنوع و کلمات کلیدی با آمار جستجوی بالا به عنوان کارهای آینده باقی میماند.

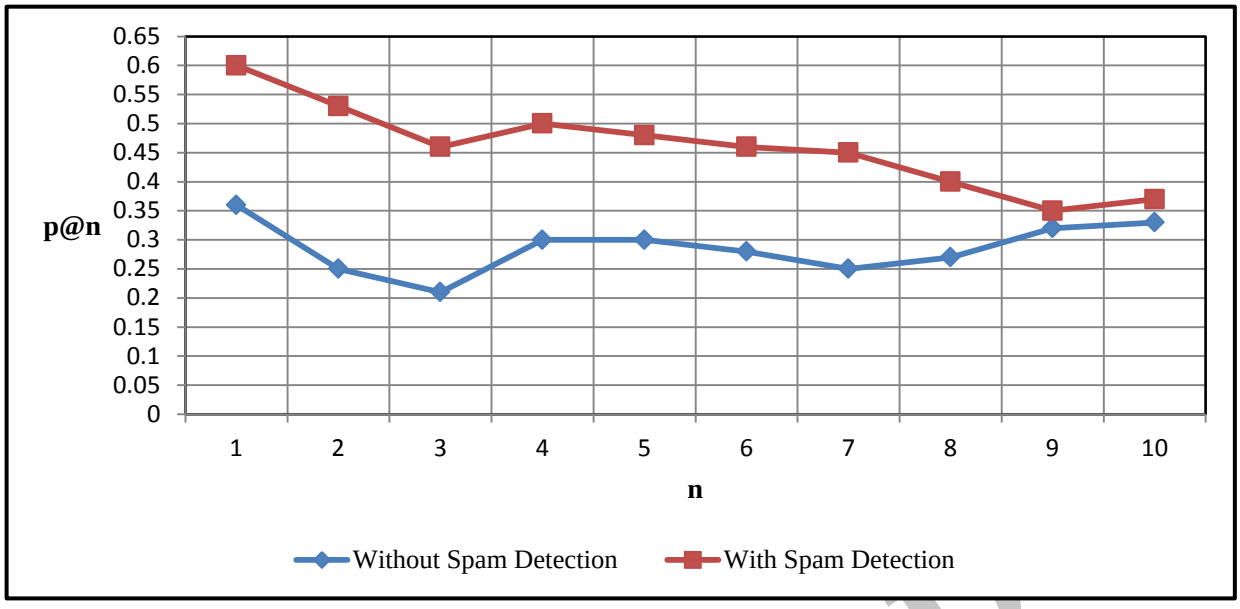
```
Input:
        Number of Host nodes :N
        distance for truncation: DST
        The initial value for Spam pages: p
        Convergence Factor: Threshold
        Incremental coefficient: k
        Time: T
Variable:
        HostRank: HR
        TruncatedHostRank: THR
        RatioOfTruncatedHostRankToHostRank: RatioTHRToHR

Output:
        Score for each Host
Function  CalculateLinkScore
For i ← 1 to N
        HostRank[i] ← HostRank(i) // Calculate Normal HostRank for each Host
End
For i ← 1 to N
        THR[i] ← THR(i) // Calculate TruncatedHostRank
End
For i ← 1 to N
        RatioTHRToHR[i] ← THR[i] / HR[i]
        if RatioTHRToHR[i] > DST
        SpamHost[i] ← p
        else SpamHost[i] ← 1/N
End
T ← 1
While(| Score^t - Score^t-1 | < Threshold)
        For i ← 1 to N
            For all link from src to dest
                Score[src] ← PageRank[i]/Outdegree[i]
                T++
                Score[src] ← (Score[src] + k * T * SpamHost[src])/Outdegree[src]
            End
        End
End
```

شکل ۸:شبه کد الگوریتم ارائه شده

[1] Domain

www.SID.ir



شکل ۹: مقایسه نتایج قبل از اعمال الگوریتم شناسایی صفحات فریب‌آمیز و پس از اعمال الگوریتم شناسایی صفحات فریب‌آمیز

مراجع

[1] M. Luckner, M. Gad and P. Sobkowiak, "*Stable web spam detection using features based on lexical items*", Computers & Security, vol. 46, pp. 79–93, 2014.

[2] A.M. ZarehBidoki, M.A. Golshani, and E. Mousakazemi-Mohammadi ", *Design and Implementation of Persian document crawling/ranking system and Implementation of a Persian Search Engine*", Itre,Tehran, Iran, 2012.(in persian)

[3] G.-R. Xue, Q. Yang, H.-J. Zeng, Y. Yu, and Z. Chen, "*Exploiting the hierarchical structure for link analysis*", Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "*The pagerank citation algorithm: bringing order to the web*," Technical Report, Standford Univ.,1998.

[5] B. Wu and B. D. Davison, "*Identifying link farm spam pages*", Special interest tracks and posters of the 14th international conference on World Wide Web,pp. 820-829, 2005.

[6] L. Becchetti, C. Castillo, D. Donato, S. Leonardi and R. Baeza-Yate,", *Link-Based Characterization and Detection of Web Spam", Proceeding of the 6th International Workshop*

on Adversarial Information Retrival on the Web (AIRWEB), 2006.

[7] Z. Gyongyi, H. Garcia-Molina and J. Peddersen, "*Combating web spam with trustrank*", Proceedings of the Thirtieth international conference on Very large data bases volume 30.VLDB Endowment, Torento, Canada, pp. 576-587, 2004.

[8] V. Krishnan, R. Raj, "*Web spam detection with anti-TrustRank*" , Proceeding of the 2nd International Workshop

on Adversarial Information Retrival on the Web (AIRWEB),pp. 37-40, 2006.

[9] www.parsijoo.ir

Archive of SID

www.SID.ir

An improved link-based method for spam detection in Persian web graph

Marzie Parooie[1*], AliMohammad Zareh Bidoki[2]
1*- Corresponding Author: Department of Electrical and Computer, Yazd University, Yazd, Iran.
2- Department of Electrical and Computer, Yazd University, Yazd, Iran.
[1*]parooie@stu.yazd.ac.ir, [2]Alizareh@yazd.ac.ir

Abstract- Today using the internet has spread wildly, and increasing number of web pages leads to importance of using search engines, therefore some people try to misguide search engines to have more customers and benefit. They increase the rank of their pages by some illegal ways. search engines to. Identify of this kind of web pages can improve search engines and attract confidence to user.
According to importance of finding spam pages, the research is presented a new linke-based way to detect spam pages in Persian web graph. This way, first link farms detectes. Finally, the negative scores of spam pages propagate in whole of web graph.
This way was implemented on data of Parsijoo search engine and the result of data analyses indicates 21.2% improvement in p@n factor.

Keywords- Search Engine, Spam, Ranking.

Archive of SID

www.SID.ir