مروری بر الگوریتم‌های رتبه بندی صفحات وب

عطیه زراعتکار برفوئیه[1]، حمید میروزیری[2] و مصطفی قاضی زاده احسانی[3]

[1] بخش مهندسی کامپیوتر دانشگاه شهید باهنرکرمان،کرمان،atiyezeraatkar@eng.uk.ac.ir

[2] بخش مهندسی کامپیوتر دانشگاه شهید باهنرکرمان،کرمانhmirvaziri@gmail.com

[3] بخش مهندسی کامپیوتر دانشگاه شهید باهنرکرمان،کرمان ،mghazizadeh@uk.ac.ir

چکیده:

حجم صفحات وب امروزه در حال افزایش است. میلیاردها صفحه وب وجود دارد که کاربران با آنها در ارتباط هستند. نکته‌ی مهم و قابل توجه چگونگی استفاده و مفید بودن صفحات است. در مرحله‌ی اول که یک کاربر پرس‌وجویی را وارد می‌کند نیاز دارد تا بهترین پاسخ‌ها را موتور جست‌وجو برای او بیاورد. بنابراین داشتن موتور جست‌وجو کارا و دقیق و با سرعت بالا لازمه‌ی کار است. اساس کار موتورهای جست‌وجو الگوریتم‌های رتبه‌بندی است.

الگوریتم‌های رتبه‌بندی از روش‌های وب کاوی استفاده می‌کنند. وب کاوی به ساختار کاوی، محتوا کاوی و کاربرد کاوی وب تقسیم می‌شود. الگوریتم‌های مختلفی از سال‌ها پیش ارائه شده‌اند. بهترین الگوریتم‌ها از ترکیب سه روش وب کاوی به دست می-آیند. در ابتدا الگوریتم‌های ساختاری مانند Page Rank **مورد بررسی قرار گرفته‌اند. سپس به آن جنبه محتوا کاوی اضافه شده است و الگوریتم‌هایی مانند** WPCR **مورد بحث قرار گرفته‌اند و در نهایت نظر کاربر به عبارتی دیگر جنبه‌ی کاربرد کاوی اضافه شده است. این الگوریتم‌ها مرور و مقایسه شده اند.**

کلمات کلیدی : page rank، in link ، out link، weighted، visit time ،content ،of link

۱- مقدمه:

وب شامل داده‌هایی از قبیل متن ، تصویر، ویدئو و داده‌های چندرسانه‌ای است. برای پردازش و به دست آوردن نتایج و الگوهای خوب از این داده‌ها نمی‌توان از روش‌های داده کاوی و متن کاوی به تنهایی استفاده کرد. بنابراین می‌توان روش‌های وب کاوی که ترکیبی از داده کاوی است استفاده کرد.

وب کاوی به سه روش ساختار کاوی وب[1] ، محتوا کاوی وب[2] و کاربرد کاوی وب[3] تقسیم می‌شود. ساختار کاوی با گراف وب کار می کند و محتوا کاوی با استفاده از روش‌های متن کاوی و داده کاوی محتوا وب را بررسی می‌کند. کاربرد کاوی بیشتر بر روی نظر کاربر کار می‌کند و از log file ها استفاده می کند.

پایه و اساس الگوریتم‌هایی که در این مقاله بررسی می‌شوند PageRank است. که از گراف وب به عنوان ورودی استفاده می‌کند. سپس با اضافه کردن پرس‌وجوی کاربر به عنوان ورودی جنبه‌های محتوایی اضافه شده است. برای افزودن جنبه کاربرد کاوی تعداد رؤیت لینک به عنوان ورودی الگوریتم اضافه می‌شود. با بررسی الگوریتم‌های ارائه شده، ترکیبی از سه روش وب کاوی بهترین جواب را می‌دهد.

۲- وب کاوی:

وب کاوی به معنای بکار گیری تکنیک‌های داده کاوی و آماری و هوش مصنوعی برای کشف و استخراج خودکار اطلاعات مفید و ساختار یافته از اسناد و سرویس‌های وب است.

وب کاوی روی داده‌هایی مانند محتوای صفحات وب، اطلاعات دسترسی کاربر و هایپرلینک میان صفحات کار می کند.

وب علاوه بر اینکه دارای مجموعه عظیمی از اطلاعات است، حاوی مجموعه‌ای پویا از پیوندها برای دسترسی به صفحات وب و استفاده از اطلاعات نیز است که یک مجموعه غنی برای داده کاوی ایجاد می کند.
وب کاوی شامل مراحل زیر است:

- پیدا کردن منبع: این مرحله شامل بازیابی اسناد وب مورد نظر است.
- انتخاب اطلاعات و پیش پردازش: به صورت خودکار اطلاعات خاصی از اسناد بازیابی شده، انتخاب و پیش پردازش می‌شوند.
- تعمیم: به صورت خودکار الگوهای عام در یک یا چندین سایت وب کشف می‌شوند.
- تحلیل: الگوهای به دست آمده در مرحله قبل اعتبارسنجی و تفسیر می‌شوند.

مشکلات موتورهای جست و جو فعلی:

- جواب‌های زیادی بر می‌گرداند.
- جواب‌ها کیفیت پایینی دارند.
- هنوز هم بسیاری از داده‌های پنهان را جستجو نمی‌کنند.

برای رفع مشکلات موتور جست وجو نیاز به یک الگوریتم رتبه بندی خوب داریم. مقاله به این موضوع پرداخته است.

۲-۱- کاوش محتوای وب:

کاوش محتوای وب اکتشاف اطلاعات مفید از محتوی، داده‌ها و اسناد وب است. وب شامل اطلاعات زیادی در قالب HTML، اخبار، انواع مجلات، کتاب های الکترونیکی، متن، تصویر، صدا و ویدیو است. کار با این اسناد و استخراج اطلاعات از آنها وظیفه کاوش محتوا وب است.

کاوش محتوا در وب را میتوان از دو دید بررسی کرد: از دید بازیابی اطلاعات و از دید پایگاه داده ها. در بازیابی اطلاعات به تسهیل یا بهبود فرایند جست و جوی اطلاعات یا فیلتر کردن اطلاعات برای کاربران پرداخته می‌شود. اما از دید پایگاه داده‌ها هدف ارائه مدلی از داده‌های وب و یکپارچه سازی آن‌ها است به طوری که پرس‌وجوهای پیچیده‌تر از پرس‌وجوهای مبتنی بر کلمات کلیدی قابل پردازش باشند. [1]

یکی از مشکلات روش های محتوایی پهنش رتبه است. پهنش به دوصورت تغییر دادن محتوا و پیوند صفحات انجام می شود. Spammer ها با تغییر دادن محتوای اسنادشان و اضافه کردن کلمات کلیدی به داخل و مکان‌های پنهان صفحه سعی در بالا بردن شباهت صفحه‌ی خود با پرس‌وجوهای آن حوزه دارند. بدین ترتیب با زیاد اشاره کردن از داخل سایت های دیگر به سایت خود رتبه‌ی آن را افزایش می دهند.[2]

۲-۲- کاوش ساختار وب:

وب دارای یک ساختار ناهمگن بزرگ است که اسناد به یک دیگر متصل هستند و یک گراف بزرگ را تشکیل می دهند. صفحه های وب به عنوان گره‌ها و ابرلینک‌ها به عنوان لبه‌های اتصال دهنده بین دو صفحه مرتبط است.

پیوندهای وب دارای اطلاعات ارزشمندی هستند به همین دلیل الگوریتم‌های جدید رتبه بندی بر اساس پیوند ارائه شدند. این روش می‌تواند برای دسته بندی صفحات وب استفاده شود و یا از آن برای تولید اطلاعاتی از قبیل میزان تشابه وب سایت ها استفاده کرد.

این روش مشکل پهنش رتبه در روش های محتوایی را برطرف می کند و کمک می کند تا بتوانیم جواب های دقیق تری را بدست بیاوریم. اگر پرس-وجو مربوط به یک تصویر، صدا و فایل ویدیویی باشد کاوش ساختاری بسیار کمک می کند زیرا در اتصالات یک منبع یا صفحه وب توصیف دقیق تری از محتوای آن وجود دارد که از کاوش محتوی صفحه مفید تر است.

الگوریتم‌های بر مبنای اتصال به دو دسته وابسته به پرس‌وجو و مستقل از پرس‌وجو تقسیم می شوند. در روش‌های مستقل از پرس‌وجو رتبه بندی به صورت برون خط و با استفاده از کل گراف وب انجام می‌شود، در نتیجه به ازای هر پرس‌وجو رتبه هر صفحه ثابت است. اما روش وابسته به پرس‌وجو رتبه بندی در گراف شامل مجموعه صفحات مرتبط با پرس‌وجوی کاربر انجام می‌شود.

یکی از مشکلات این روش غنی تر شدن اغنیا است. قرار گرفتن همیشه صفحات محبوب[۴] در صدر لیست ارائه شده به کاربر، باعث می‌شود تا کاربر فقط صفحات خاصی را ببیند و در نتیجه صفحات تازه متولد شده ی با کیفیت بالا که کسی به آنها اشاره نمی کند نتوانند در دید کاربران قرار گیرند. این مشکل باعث می‌شود صفحات محبوب مرتباً محبوب تر شده و تعداد پیوند به آنها افزایش یابد. [2]

۲-۳- کاوش در دسترسی و استفاده از وب:

هر روز مراجعه کنندگان به یک سایت وب اطلاعات ارزشمندی از خود بر جای می گذارند. چگونه این اطلاعات استخراج و تبدیل به اطلاعات تجاری و مدیریتی مفید کنیم؟ مجموعه اعمالی که یک کاربر یک سایت وب در هر مراجعه انجام می دهد نمایش دهنده رفتار، عادتها، خواسته ها و گرایش اوست. منابع اصلی اخذ رفتارها و داده های کاربر عبارتند از: فرم ها ، کوکی ها ، لوگ فایل ها.

لوگ فایل‌ها: سرور اینترنت موقع اتصال هر کاربر یک فایل تشکیل می دهد که در آن اطلاعات ارتباط برقرار شده و کاربر را در آن ثبت می کند . این اطلاعات شامل IP یا URL کاربر، شناسه کاربر، تاریخ و زمان مراجعه، محل ارجاع و نوع مرورگر است.

کوکی‌ها: فایل هایی هستند که در کامپیوتر کاربر برای استفاده در مراجعات بعدی، نگهداری سابقه اتصال، ثبت اعمال انجام شده توسط کاربر ایجاد می شود. این فایل، یک فایل متنی است. در این فایل مدت زمان اتصال کاربر، شناسه کاربری و رمز ورود، فعالیت های کاربر در طول ارتباط نگهداری می شود. یک نسخه از فایل کوکی در انتهای فایل لوگ سرور نگهداری می شود.

فرم‌ها: فرمها صفحاتی هستند که برای گرفتن اطلاعات از کاربر استفاده می شوند. مهمترین نوع فرمها، فرم ثبت نام یا عضویت کاربر در یک سایت که مربوط به نگهداری اطلاعات شخصی مانند نام، نام خانوادگی، سن، جنس، میزان تحصیلات، شغل، میزان درآمد، محل زندگی و یا حتی علاقه ها و نیازها و ...است.

۳- مروری بر الگوریتم های مختلف رتبه بندی صفحات وب:

۳-۱- Page Rank:

یک روش مستقل از پرس و جو است. این الگوریتم رتبه هر صفحه را با اختصاص وزن به پیوندی که به آن صفحه داده شده است به دست می آورد. مقدار این وزن به کیفیت صفحه ای که پیوند در آن قرار گرفته، بستگی دارد. در این صورت پیوندهای صفحات مهم تر وزن بیشتری می گیرند.

$$r(i) = c * \sum_{j \in B(i)} \frac{r(j)}{N(j)} \qquad (1)$$

B (i) صفحات ورودی به صفحه i، n تعداد صفحات وب، N تعداد صفحات خروجی j، r (i) و r (j) به ترتیب رتبه صفحه i و j می باشند. C یک مقدار ثابت است که در اینجا یک در نظر گرفته شده است.

از آنجایی که وب کاملا متصل نیست و دارای دو مشکل پایین افتادن رتبه و دیگری سوراخهای وب است(یک کلاستر کاملاً متصل داخلی در گراف وب که هیچ اتصالی به خارج این کلاستر ندارد را یک پایین افتادن رتبه می نامیم.). اول باید تمام نودهای سوراخهای وب با درجه خروجی 0 را پاک کنیم. و بعد برای حل مشکل پایین افتادن رتبه از یک ضریب میرایی[۵] با نام d استفاده می کنیم.[2]

$$r(i) = \left(\frac{1-d}{n}\right) + d * \sum_{j \in B(i)} \frac{r(j)}{N(j)} \qquad (2)$$

یک صفحه دارای رتبه ی بالا است اگر تعداد صفحات زیادی به آن اشاره کنند یا صفحات اشاره کننده دارای رتبه ی بالایی باشند.[3]

۳-۲- Weighted Page Rank:

این الگوریتم در واقع یک نسخه توسعه یافته از الگوریتم PR است. در این الگوریتم اهمیت یک صفحه علاوه بر لینک های خروجی به لینک های ورودی نیز وابسته است.

$$WPR(i) = (1-d) + d \sum_{j \in B(i)} WPR(j) W^{in}_{(j,i)} W^{out}_{(j,i)} \qquad (3)$$

پارامترهای $W^{in}_{(j,i)}$، $W^{out}_{(j,i)}$ به ترتیب وزن صفحات ورودی و خروجی می باشند.[4]

۳-۲-۱. وزن صفحات ورودی:

این فرمول وزن link(j,i) را بر اساس تعداد لینک های ورودی به صفحه i ، نسبت به تعداد لینک های ورودی به تمام صفحات مرجع مربوط به صفحه j به صورت زیر محاسبه می کند.

$$W^{in}_{(j,i)} = \frac{Ii}{\sum_{p \in B(j)} Ip} \qquad (4)$$

I_i و I_p به ترتیب تعداد لینک های ورودی به صفحه i وp می باشند.

۳-۲-۲. وزن صفحات خروجی:

مقدار $W^{out}_{(j,i)}$ در فرمول 3 وزن link(j,i) را بر اساس تعداد لینک های خروجی به صفحه i ، نسبت به تعداد لینک های خروجی به تمام صفحات مرجع مربوط به صفحه j به صورت زیر محاسبه می کند.

$$W^{out}_{(j,i)} = \frac{Oi}{\sum_{p \in B(j)} Op} \qquad (5)$$

O_i و O_p به ترتیب تعداد لینک های خروجی به صفحه i وp می باشند.

این الگوریتم فقط بر روی لینک ها کار می کند. ممکن است صفحاتی که در اول لیست آورده می شوند با پرس‌وجوی کاربر نامرتبط باشد. برای حل این مشکل برای صفحات چهار قانون تعریف می کند.

- صفحات بسیار مرتبط(VR)[۶]: شامل اطلاعات بسیار مهمی در مورد پرس‌وجو می باشد.
- صفحات مرتبط[۷](R): با پرس‌وجو مرتبط است اما اطلاعات مهمی در مورد پرس‌وجو ندارد.
- صفحات با ارتباط ضعیف[۸](WR): اطلاعات مفیدی در مورد پرس‌وجو ندارند در صورتی که شامل کلمات کلیدی می باشند.
- صفحات نا مرتبط[۹](IR): نه تنها اطلاعات مفیدی از پرس-وجو ندارند بلکه کلمات کلیدی را هم ندارند.

مقادیر VR>R>WR>IR می باشد که مقادیر پیشنهادی این الگوریتم به ترتیب 0,0.1,0.5,1 است. که با مقادیر به دست آمده از PR و WPR جمع می شود و نتایجی بهتری را در لیست صفحات برای کاربر می آورد.

۳-۳- Weighted Link Rank:

در الگوریتم PR به لینک ها یک اهمیت مساوی داده شده است در این الگوریتم لینک ها بر اساس سه مشخصه موقعیت نسبی در صفحه[۱۰] ، تگی که از آن لینک موجود است[۱۱] و طول متن های کلیک شونده[۱۲] وزن دهی می شوند.

$$R(i) = \frac{1-d}{n} + d \sum_{j \in B(i)} \frac{W(j,i) R(j)}{\sum_k W(j,k)} \qquad (6)$$

$$W(j,i) = L(j,i)\big(c + T(j,i) + AL(j,i) + RP(j,i)\big) \qquad (7)$$

L (j,i) اگر لینکی از صفحه i به j وجود داشته باشد برابر یک است در غیر این صورت صفر است. C یک مقدار ثابت است که یک وزن پایه به هر لینک می هد. (برابر یک در نظر گرفته شد).

T(j,i) مقداری که بر اساس لینک در تگ است اگر در <h1>آمده باشد ارزش بالاتری نسبت به <h2> دارد همچنین برای و <b> برقرار است.

AL(j,i) طول کلمات قابل کلیک که بر میانگین طول کلمات قابل کلیک در کاراکتر تقسیم می‌شود(برابر ۱۰۰ در نظر گرفته شد) .در واقع کلمات قابل کلیک طولانی (دارای جزئیات یا توضیحات)ارزش بیشتری نسبت به کلمات کوتاه مانند" اینجا" و "خانه" دارند.

RP(j,i) موقعیت نسبی لینک نه در صفحه بلکه در کد است. لینک هایی که در اول کد قرار دارند ارزش بیشتری دارند.

نتایج بهتری نسبت به PR به دست آورده است.[5]

۳-۴- Weighted Page Content Rank:

در الگوریتم های PR ممکن است صفحاتی که در اول لیست هستند به پرس و جوی کاربر مربوط نباشند و صرفا بر اساس تعداد لینک هایی که دارند رتبه-ی بالایی گرفته باشند. برای رفع این مشکل از جنبه‌هایی محتوایی که در پایین اشاره شده است استفاده می‌شود.

این الگوریتم از تکنیک های ساختار کاوی و محتوا کاوی هم زمان استفاده می کند و نظم صفحات در لیست نتیجه را بهبود می دهد. بطوریکه کاربر می تواند صفحات مهم و مربوط را به راحتی در لیست پیدا کند. در واقع در کنار استفاده از جنبه های ساختاری وب که در WPR استفاده می شد با مقایسه کلمات پرس و جوی کاربر با صفحه وب جواب های بهتری را پیدا می کند.

از ساختار کاوی وب برای محاسبه اهمیت صفحه استفاده می کند و محتوا کاوی وب برای پیدا کردن اینکه یک صفحه چقدر مربوط است مورد استفاده قرار می گیرد. اهمیت به معنای محبوبیت صفحه است می توان آن را بر اساس تعداد لینک های داخلی و خارجی و بیرونی صفحه محاسبه کرد. رابطه به معنای تطبیق صفحه با پرس و جو است و اگر یک صفحه حداکثر تطبیق را با پرس و جو داشته باشد بیشتر مربوط است.

$$PR(i) = (1-d) + d\sum_{j\in B(i)} PR(j) W^{in}_{(j,i)}\ W^{out}_{(j,i)} * (CW + PW) \quad (8)$$

در فرمول فوق B صفحاتی می باشند که به i اشاره می کنند. Win وزن داخلی لینک و $Wout$ وزن خارجی صفحات است اما CW وزن محتوای صفحه i و PW وزن احتمالی صفحه i است که در ادامه توضیح داده شده است.[6]

۳-٤-۱. محاسبه ارتباط:

ارتباط انتخاب یک صفحه را از نظر دو عامل محاسبه می کند: یکی نشان دهنده احتمال پرس و جو در صفحه و دیگری حداکثر تطبیق از پرس و جو به این صفحه است.

احتمال وزن:

احتمالی از حضور کلمات پرس‌وجو در صفحه وب است. که نسبت تعداد کلمات پرس‌وجو در سند حاضر بر تعداد کل کلماتی که از پرس‌وجو کاربر استخراج شده است.

$$probability\ weight(PWi) = \frac{Y_i}{N} \quad (9)$$

Yiتعداد کلمات پرس‌وجو در سند i ام و N تعداد کل کلمات پرس‌وجو است.

وزن محتوا:

این وزن از محتوای صفحه وب با توجه به شرایط پرس‌وجو استفاده می کند. این عامل نسبت مجموع فرکانس بالاترین رشته پرس‌وجو ممکن در سفارش بر مجموع فرکانس تمام رشته پرس‌وجو سفارش است.

$$Content\ weight(CWi) = \frac{X_i}{M} \quad (10)$$

M مجموع رتبه همه رشته های پرس‌وجوی ممکن معنی دار در سفارش، Xi مجموع رتبه بالاترین رشته پرس‌وجو در سفارش است.

۳-۵- Page Rank Based on Number of Visit of Link:

الگوریتم‌هایی که تا الان بیان شد چه بر اساس ساختار یا محتوای وب هیچکدام رفتار کاربر را نشان نمی‌دادند این الگوریتم رفتار کاربر را بر اساس اینکه یک لینک را چند بار رؤیت کرده است نشان می‌دهد.

$$PR(i) = (1-d) + d\sum_{j\in B(i)} \frac{L_i PR(j)}{TL(j)} \quad (11)$$

Li تعداد رؤیت از صفحه i به j و $TL(j)$ تعداد کل رؤیت لینک های خروجی از صفحه j است.[7]

۳-۵-۱. چگونگی محاسبه تعدادرؤیت‌ها:

برای محاسبه تعداد رؤیت لینک های خروجی از یک صفحه وب به اسکریپت سمت مشتری نیاز است. زمانی که یک صفحه درخواست می‌شود اسکریپت طرف مشتری از سمت سرور وب بارگذاری می‌شود. زمانی که یک رخداد بر روی هایپر لینک اتفاق می افتد اطلاعات رخداد و صفحه هایپرلینک را برای سرور می فرستد.

در قسمت سرور یک پایگاه داده از لوگ فایل ها برای ذخیره شماره صفحه هایپرلینک و تعداد ویزیت‌ها استفاده می‌شود. تعداد رؤیت‌ها هر وقت که یک صفحه رؤیت می‌شود افزایش می‌یابد. این اطلاعات برای رتبه بندی بهتر به موتور جست و جو داده می شوند.

۳-۶- Weighted Page Rank Based on Number of Visit of Link:

همانطور که قبلا گفته شد WPR امتیاز بالاتری به صفحات مهم تر می دهد. در این جا فقط وزن صفحات ورودی در نظر گرفته شده است و با الگوریتم VOL ترکیب شده است.

$$WPR_{vol}(i) = (1-d) + d\sum_{j\in B(i)} \frac{L_i WPR_{vol}(j) W_{in}(j,i)}{TL(j)} \quad (12)$$

پارامترهای فرمول فوق در قسمت های قبل بیان شده اند و لزومی برای بیان دومرتبه‌ی آنها نیست. این الگوریتم در اکثر مواقع جوابی بهتر از WPR می دهد.[8]

۳-۷- An Effective Content based Web Page Ranking:

الگوریتم PR به دلیل دادن امتیاز یکسان به صفحات ورودی و خروجی کارایی خوبی ندارد. ممکن است صفحاتی در اول لیست آورده شوند که ارتباط نزدیکی با پرس‌وجوی وارد شده نداشته باشند برای رفع این مشکل از ترکیب وزن محتوا با وزن لینک های ورودی استفاده می‌شود.

$$PR(i) = (1-d) + d\sum_{j\in B(i)} PR(j).WL(j,i).Wc \quad (13)$$

در فرول فوق Wc همان وزن محتوا است. WL هم وزن صفحات ورودی است.[9]

۳-۸- Weighted Page Rank Algorithm Based on Number of Visit of Link in Time Duration:

الگوریتم WPRVOL رتبه ی بالاتری به صفحاتی که مهم تر هستند می دهد. اما در این روش فاکتور زمان اضافه شده است. در این الگوریتم تعداد دفعات رؤیت یک صفحه خاص در یک محدوده ی زمانی محاسبه می‌شود. و رتبه ی آن بر همین اساس به دست می آید. با یک بروز رسانی برون خط شروع می کنیم. به هر وب سایت بر اساس موتور جست و جو یک مقدار اولیه به عنوان رتبه اختصاص داده می‌شود. سپس در یک محدوده‌ی زمانی معین به تدریج که رؤیت صورت می‌گیرد رتبه‌ها به‌روز می‌شوند. در نتیجه وب سایت هایی که میزان رؤیت آنها تغییر کرده است در اول لیست قرار می گیرند.

$$WPR_{vol}(i) = \{[(1-d)+d\sum_{j\in B(i)}\frac{L_u WPR_{vol}(j)W_{in}(j,i)}{TL(j)}]*t+k\} \quad (14)$$

در فرمول فوق t نشان دهنده سال تحت بررسی، k یک مقدار ثابت است. بقیه پارامترها قبلا آورده شده اند.

در این الگوریتم صفحات قدیمی کمتر بالای لیست می آیند. و یک کاربر نمی تواند با رؤیت های مکرر رتبه‌ی خود را بالا ببرد زیرا در اینجا نسبتی از رؤیت در یک محدوده‌ی زمانی محاسبه می‌شود. و این الگوریتم جنبه‌های ساختاری و کاربرد کاوی وب را شامل می‌شود و نتیجه خوب با دقت بالا را می دهد.[16]

۳-۹- Page Rank with Topic Bias and Time Factor :

با توجه به مشکل غنی تر شدن اغنیا در الگوریتم PR می توان گفت: رتبه ی PR برای بعضی از صفحات نشانه ی اهمیت نیست بلکه نشانه ی سن است. [10]

برای حل این مشکلات روشی ارائه شده است. در الگوریتم TSPR[۱۳][14] وقتی کاربر پرس‌وجو را وارد می کند شباهت بین موضوع-های اصلی وب و موضوع پرس‌وجو به دست می آورد.

۱- محاسبات برون خط:

رتبه‌ی صفحه را در موضوعات مختلف به دست می‌آورد. مثلا در بحث تجاری رتبه ی آ، در بحث فرهنگی رتبه ی ب.

۲- محاسبات برخط:

برای گرفتن رتبه ی جامع بر اساس مفهوم پرس‌وجو. ابتدا احتمال موضوعات مختلف از پرس‌وجو را محاسبه می کند. و سپس نمره‌ی کلی را به دست می آورد. برای مثال یک موضوع که مربوط به فرهنگ است بدون در نظر گرفتن موضوع فرهنگ، احتمال آن بالا است. سرانجام احتمال آن نیز در موضوع فرهنگ بالا است.

مشکل این الگوریتم پیچیدگی محاسباتی بالای آن است.

در الگوریتم PRTT ابتدا چندین موضوع از هر وب سایت مشخص می‌شود. اگر کلمات کلیدی که در فضاهای زیر بیشتر اتفاق افتاده باشند به عنوان موضوع وب سایت در نظر گرفته می شوند. <META>,<TITTLE>,<A herf> ودر جاهایی که تاکید دارند مثلا در حروف پررنگ یا کج.

در نهایت رتبه با استفاده از شباهتی که بین صفحه با بقیه ی صفحاتی که به آن لینک دارند محاسبه می‌شود. کلمات پرس‌وجوی کاربر به عنوان موضوع‌های جست‌وجو است.

M ضریب تقسیم، عامل تقسیم برای توزیع ارزش نسبت معینی از PR صفحات به آن صفحات است.

W_iوزن موضوعی بین صفحه مرجع و صفحه i است. کاربر دنبال کلمه کلیدی k است:

۱- اگر k بین موضوعات صفحه باشد. برای صفحات خروجی که k جزء کلمات کلیدی آنها باشد شمارنده افزایش می یابد(مقدار اولیه صفر است.). و W_i برابر نسبت عکس شمارنده می‌شود در غیر این صورت مقدار آن صفر می‌شود.

۲- اگر k بین موضوعات صفحه A نباشد.

$$N_i = |Topic(A) \cap Topic(Pi)|$$
$$W_i = \frac{N_i}{\sum N_i} \quad i = 1,2,3 \dots . n \quad (15)$$

برای رفع مشکل زمان از تاریخ انتشار استفاده می کند. موتور جست‌وجو به صورت دوره‌ای وب را می‌خزد. در هر دوره که صفحه آورده شده باشد شمارنده آن افزایش می یابد و سپس زمان به صورت زیر تعریف می‌شود.

$$t = \frac{1}{n+\eta} \quad (16)$$

η یک عامل تعادل است.

در نهایت در این الگوریتم رتبه از فرمول زیر به دست می آید.

$$PR(i) = dT_i[\sum_{j\in B(i)} PR(j)\left(\frac{m}{N_v}+(1-m)W_j\right)+(1-d)] \quad (17)$$

پارامترها در پاراگراف‌های بالا آورده شده اند.

۳-۱۰- Ratio rank:

بهبود یافته الگوریتم PR است از ساختار کاوی وب استفاده می کند. همچنین از وزن لینک‌های ورودی و خروجی و تعداد رؤیت لینک‌ها استفاده کرده است. و از کاربران نسبت اهمیت پارامترها را سوال می کند.

$$PR(i) = (1-d)+d\sum_{j\in B(i)}\frac{(L_i*x*W_{in}(j,i)+y*W_{out}(j,i))PR(j)}{TL(j)} \quad (18)$$

x نسبت وزن لینک‌های ورودی و y نسبت وزن لینک‌های خروجی است. مقادیرشان بین صفر تا یک است. اما همیشه مقدار x بیشتر از y است زیرا اهمیت لینک‌های ورودی بیشتر از لینک‌های خروجی است.[11]

طی آزمایشات انجام شده برای پیدا کردن بهترین نسبت x و y مقدار x برابر ۰.۷ و y برابر ۰.۳ در نظر گرفته شد.[12]

۳-۱۱- Enhanced-Ratio Content Rank

این الگوریتم ترکیبی از ساختار کاوی و محتوا کاوی وب است. به طوریکه اهمیت و رابطه صفحه وب از طریق وزن‌های لینک‌های ورودی و خروجی و تعداد رؤیت‌های صفحه و یک پارامتر وزن محتوایی با توجه به پرس و جوی کاربر محاسبه می‌شود.

$$PR(i) = (1-d) + d * W_c \sum_{j \in B(i)} \frac{(L_i * 0.7 * W_{in}(j,i) + 0.3 * W_{out}(j,i)) PR(j)}{TL(j)} \quad (19)$$

در فرمول ۱۹، Wc همان وزن محتوایی است که در فرمول۱۰آمده است.

این الگوریتم به دلیل استفاده وزن محتوایی تا حدودی مشکل رانش موضوع که در روش‌های ساختار کاوی هست برطرف می کند.[13]

۳-۱۲- Weighted Page User Content Rank:

الگوریتم WPUCR از هر سه روش وب کاوی استفاده می کند. این الگوریتم از وزن لینک‌های ورودی و خروجی و وزن محتوایی به همراه نظر کاربر استفاده می کند. نظر کاربر در اینجا به معنی تعداد دفعاتی که صفحه پیموده می‌شود است.

$$WPUCR(i) = F * \{(1-d) + d \sum_{j \in B(i)} WPUCR(j) W^{in}_{(j,i)} \; W^{out}_{(j,i)} * (CW + PW)\} \quad (20)$$

در فرمول فوق F تعداد دفعاتی است که صفحه پیموده شده است البته نشان دهنده نظر و علاقه ی کاربر است. این روش به دلیل استفاده از هر سه جنبه وب کاوی جواب دقیق تر و کامل تری را می دهد.[15]

٤- نتیجه گیری:

الگوریتم‌هایی که تنها بر اساس روش‌های ساختار کاوی وب می باشند مانند PR داری معایبی از جمله: صفحاتی که در صدر لیست می آیند ارتباط کمی با پرس‌وجوی کاربر دارند، به تمام لینک‌ها وزن یکسانی داده شده است در صورتی که لینک‌های ورودی اهمیت بیشتری دارند و غنی تر شدن اغنیا است.

الگوریتم‌های محتوایی دارای مشکل پهنش رتبه و پیچیدگی محاسباتی بالا می باشند. برای حل مشکل اول می توان آنها را با روش‌های ساختار کاوی ترکیب کرد. مقایسه الگوریتم ها در جدول ۱ است. می توان مزایا معایب و روش کاوش همچنین نوع ورودی های آنها را دید.

اما در ترکیب سه روش وب کاوی جواب‌های بهتری به دست می آید و تقریبا معایب ذکر شده بر طرف می شوند. در نهایت بهترین الگوریتم برای موتور جست و جو عمدتا الگوریتم‌هایی می باشند که هر سه جنبه ی وب کاوی را داشته باشند و پیچیدگی محاسباتی پایینی داشته باشند. از جمله این روش‌ها می توان Enhanced-Ratio Content Rank و Page Rank with Topic Bias and Time Factor نام برد.

جدول(۱) مقایسه الگوریتم های رتبه بندی صفحات وب

الگوریتم	تکنیک کاوش	ورودی ها	مزایا	معایب
Page Rank	ساختارکاوی وب	گراف وب	محاسبه رتبه بر اساس ساختار لینک(محبوبیت صفحه)	وزن برابر برای لینکهای ورودی و خروجی، غنی تر شدن اغنیا، صفحات نا مرتبط با پرس و جو
Weighted Page Rank	ساختارکاوی وب	گراف وب	میزان رتبه بالاتر از PR	غنی تر شدن اغنیا، صفحات نا مرتبط با پرس و جو
Weighted Link Rank	ساختارکاوی وب و محتوا کاوی وب	موقعیت نسبی لینک در صفحه، تگی که از آن لینک موجود است، طول متن های کلیک شونده	به لینک ها اهمیت مساوی نمی دهد.	صفحات نا مرتبط با پرس و جو، محبوبیت صفحات را در نظر نمی گیرد.
Weighted Page Content Rank	ساختارکاوی وب و محتواکاوی وب	گراف وب، وزن محتوایی و احتمال وزن	صفحات مرتبط با درخواست کاربر	پهنش رتبه

Page Rank Based on Number of Visit of Link	ساختارکاوی وب و کاربردکاوی وب	گراف وب، تعداد رؤیت لینک	اهمیت به نظر کاربر	غنی تر شدن اغنیا، صفحات نا مرتبط با پرس و جو، بالا بردن رتبه با رؤیت مکرر صفحه خود
Weighted Page Rank Based on Number of Visit of Link	ساختارکاوی وب و کاربردکاوی وب	گراف وب، تعداد رؤیت لینک	اهمیت به نظر کاربر و وزن دهی لینک های ورودی و خروجی	غنی تر شدن اغنیا، صفحات نا مرتبط با پرس و جو، بالا بردن رتبه با رؤیت مکرر صفحه خود
Enhanced Content Page Rank	ساختارکاوی وب و محتواکاوی وب	گراف وب، وزن محتوایی	صفحات مرتبط با پرس وجو کاربر، پیچیدگی محاسباتی پایین	در نظر نگرفتن وزن لینک های خروجی، غنی تر شدن اغنیا
Weighted Page Rank Based on Number of Visit of Link in time duration	ساختارکاوی وب و کاربردکاوی وب	گراف وب، تعداد رؤیت لینک، زمان	اهمیت به نظر کاربر و وزن دهی لینک های ورودی وخروجی، ناتوانی در بالا بردن رتبه با رؤیت مکرر صفحه خود	صفحات نا مرتبط با پرس و جو
Topic Sensitive Page Rank	ساختارکاوی وب و محتواکاوی وب	گراف وب، پرس‌وجوی کاربر، موضوع های پایه سایت	در نظر گرفتن پرس وجو کاربر	زمان اجرا بالا، غنی تر شدن اغنیا
Page Rank with Topic Bias and Time Factor	ساختارکاوی وب و محتواکاوی وب	وزن موضوعی، ضریب تقسیم، زمان	زمان اجرا بهتر از TSPR، در نظر گرفتن پرس وجو کاربر و زمان	مشکلی ندارد، محاسبه پارامتر زمان از روش های مختلف
Ratio Rank	ساختارکاوی وب و کاربردکاوی وب	گراف وب، تعداد رؤیت لینک ها، نسبت اهمیت	اهمیت به نظر کاربر و وزن دهی لینک های ورودی وخروجی،	در نظر نگرفتن پرس وجو کاربر
Enhanced Ratio content Rank	ساختارکاوی وب و کاربردکاوی وب و محتوا کاوی وب	وزن محتوایی، گراف وب، تعداد رؤیت لینک ها	اهمیت به نظر کاربر و وزن دهی لینک های ورودی وخروجی، در نظر گرفتن پرس وجو کاربر	مشکلی ندارد
Weighted Page User Content Rank	ساختارکاوی وب و کاربردکاوی وب و محتواکاوی وب	تعداددفعاتی که صفحه پیموده می‌شود،گراف وب، وزن محتوایی، احتمال وزن	اهمیت به نظر کاربر، در نظر گرفتن پرس وجو کاربر و وزن دهی لینکهای ورودی وخروجی	مشکلی ندارد

مراجع:

[1]عین آبادی ،حسین ،" **روشهای مختلف و بکاوی و قابلیت‌های آنها**"، پنجمین کنفرانس سراسری سیستم های هوشمند،ICS05_078، ۱۳۸۲

[2] محمد زارع بیدکی، علی ، **رتبه بندی و خزش مؤثر در وب** ،رساله برای دریافت درجه دکتری در رشته مهندسی کامپیوتر- نرم افزار، دانشگاه تهران، ۱۳۸۸

[3] Brin, S., and Page, L. "The anatomy of a large-scale hypertextual web search engine". In Proceedings of 7th World Wide Web Conference, 1998.

[4] Xing ,Wenpu., Ghorbani Ali., "Weighted Page Rank Algorithm", Proceedings of the Second Annual Conference on

Communication Networks and Services Research (CNSR '04), IEEE, 2004.
[5] BaezaYates ,Ricardo., Davis, Emilio.," Web Page Ranking using Link Attributes" ACM 1581139128/ 04/0005, New York ,17–22- 2004.
[6] Sharma P.; Bhadana P. "Weighted Page Content Rank For Ordering Web Search Result", International Journal of Engineering Science & Technology, Vol. 2(12), PP. 7301-7310, 2010.
[7] Kumar G.; Duhan N.; Sharma A. K. "Page Ranking Based on Number of Visits of Links of Web Page". International Conference on Computer & Communication Technology (ICCCT), 2011.
[8] Tyagi, Neelam., Sharma, Simple., "Weighted Page Rank Algorithm based on number of visits of links of web page", International Journal of Soft Computing and Engineering (IJCSE)-ISSN: 2231-2307, Volume-2, Issue-3, July 2012
[9] SHALYA, NIDHI., SHUKLA, SHASHWAT.," An Effective Content Based Web Page Ranking Approach" International Journal of Engineering Science and Technology (IJEST)- ISSN : 0975-5462, Vol. 4 No.08 August 2012.
[10] Huang, Wei., Li, Bin.," An Improved Method for the Computation of PageRank" International Conference on Mechatronic Science, Electric Engineering and Computer, 1-61284-722-1/11/$26.00 ©2011 IEEE
[11] Singh, Ranveer., Sharma, Dilip Kumar.," RatioRank: Enhancing the Impact of Inlinks and Outlinks" 3rd IEEE International Advance Computing Conference (IACC), 2013.
[12] Singh, Ranveer., Sharma, Dilip Kumar.," Enhanced-RATIORANK: Enhancing Impact of Inlinks and Outlinks" IEEE Conference on Information and Communication Technologies ,2013.
[13] Joshi, Rutusha., Gupta, Vinit Kumar.," Improving Pagerank Calculation by using Content Weight" International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064 ,2012.
[14] T.H. Haveliwala." Topic-Sensitive PageRank". IEEE transactions on knowledge and data ,2003.
[15] Bhardwaj, Ekta., Kumar, Shiv., Tomar, Kuldeep.," Enhancing Page Rank Algorithm" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 5, ISSN: 2321-8169, May 2015.
[16] Gambhir, Anushree., Goyal, Arushi., Singh, Sumit., Srivastava, Yatharth.," Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Pages in Time Duration" international Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463, Vol. 3 Issue 7, July-2014

[1] Web structure mining
[2] Web content mining
[3] Web usage mining
[4] Popular
[5] Damping factor
[6] Very relevant page
[7] Relevant page
[8] Weak relevant page
[9] Irrelevant page
[10] Relative position in page
[11] Tag where the link is contained
[12] Length of anchor text
[13] Topic sensitive page rank