تاریخ دریافت مقاله: ۹۷/۰۸/۰۶
تاریخ پذیرش مقاله: ۹۸/۰۷/۲۰

ترکیب الگوریتم HITS با الگوریتم Distance Rank برای بهبود نتایج در موتورهای جستجو

رعنا میلانی اباجلو
گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران.
پست الکترونیکی: r.milani۹٤abj@gmail.com

فرهاد سلیمانیان قره چپق*
گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
پست الکترونیکی: bonab.farhad@gmail.com

چکیده:

امروزه موتورهای جستجوگر از روش‌های وب‌کاوی برای نشـــان دادن نتایج بهتر استفاده می‌کنند که در لیست نتایج خود پیوندهای زیادی از صفحات وب را به کاربران نمایش می‌دهند و برای بهینه و محدود کردن لیســـت نتایج موتورهای جســـتجو از الگوریتم‌های رتبه‌بندی اســـتفاده می‌شـــود. در ایـــن مقاله یـــک روش جدید کـــه ترکیبی از الگوریتم HITS بـــا الگوریتم Distance Rank اســـت برای بهبود نتایج در موتورهای جســـتجو ارائه‌شده است که در روش پیشنهادی از فرایند اصلی الگوریتم Distance Rank برای بهبود الگوریتم HITS اســـتفاده شـــده اســـت. مشکل اصلـــی الگوریتم HITS این اســـت کـــه رتبه‌بندی صفحات وب براســـاس میزان ارتباط آن‌ها با پـــرس وجوی کاربر اســـت. اما در الگوریتم Distance از فاصله لگاریتمی میان صفحات به منظور رتبه‌بندی اســـتفاده می‌شـــود. ارزیابی روش پیشـــنهادی بر روی سه مجموعه داده شامل گراف اســـتاندارد، گراف تصادفی، گراف دانشـــگاه آزاد اسلامی واحد ارومیه انجام گرفته که نتایج نشان می‌دهد که روش پیشنهادی در مقایسه با الگوریتم‌های دیگر عملکرد بهتری دارد و توانســـته اســـت رتبه‌بندی متفاوت و بهتری نسبت الگوریتم پایه HITS و ســـایر الگوریتم‌های رتبه‌بندی مانند Distance Rank و PR و WPR داشـــته باشـــد. همچنیـــن الگوریتم‌های پیشـــنهادی برمبنای معیارهای AP، P@n و NDC مورد ارزیابی قرار گرفت که نتایج نشـــان می‌دهد که روش پیشـــنهادی به ترتیب مقدار ۱ و ۱ و ۸/۱ را به دست آورده است.

واژه‌هـــای کلیدی: موتورهای جستجو، الگوریتم HITS، الگوریتم Distance Rank

۱. مقدمه

در دنیای اطلاعاتی امروز که هزاران گیگابایت اطلاعات در بیـــن کاربران رد و بدل می‌شـــود، فضـــای وب یکی از بهترین فضاهای استفاده است. از سویی پیدایش و پیشرفت وب در دهه اخیر موجب تحول شگرفی در فراگیری اخبار

* نویسندهٔ مسئول

و اطلاعات علمی در سراسر دنیا شده است؛ این امر باعث افزایش اطلاعات در جهان شده و مشکل پیدا کردن اطلاعات ارزشمند و معتبر از بین حجم وسیع اطلاعات را در پی دارد [۱-۲]. به همین خاطر جستجو مطالب امروزه به امر مهمی تبدیل‌شده است. موتورهای جستجو ازجمله راه‌حل‌های مناسب برای حل این مسئله هستند که شرکت‌هایی همچون گوگل به‌طور مدام در حال یافتن بهترین و سریع‌ترین روش برای کاربران خود هستند. موتورهای جستجو برخلاف تصور کاربرانش که فکر می‌کنند این موتورها مطالب درخواستی‌شان را از بین تمام مطالب موجود در جهان جستجو می‌کند، درواقع آن‌ها این مطالب را از بین مطالب ذخیره‌شده در پایگاه خود جستجو کرده و در اختیار کاربرانش در آن لحظه جست‌وجو قرار می‌دهند [۲-٤]. تعداد بینندگان یک وبگاه از مهم‌ترین عاملی که صفحات وب را دارای اهمیت می‌کند. درواقع موتورهای جستجو بینندگان را به وبگاه می‌آورند بدون توجه به این‌که وبگاه چه خدمتی ارائه می‌دهد یا چه چیزی می‌فروشد. اکثر مردم فقط ۱۰ وبگاه اول معرفی‌شده توسط موتور جستجو را مرور می‌کنند؛ بنابراین موتورهای جستجو و الگوریتم‌های بازیابی اطلاعات و الگوریتم رتبه‌بندی صفحات وب اطلاعات نقش حیاتی را در دنیای امروزی اطلاعات دارد [۵،٦]. الگوریتم‌های رتبه‌بندی صفحات وب، مجموعه‌ای از دستورالعمل‌ها است که موتور جستجوگر با اعمال آن‌ها بر پارامترهای صفحات موجود در پایگاه داده‌اش، تصمیم می‌گیرد که صفحات مرتبط را چگونه در نتایج جستجو مرتب کند. درواقع، رتبه‌بندی ارزش یک صفحه توسط موتور جستجو را مشخص کرده و صفحاتی باکیفیت بالا مبتنی بر درخواست و سلیقهٔ کاربر را پیدا می‌کند [۲].

الگوریتم‌های زیادی برای رتبه‌بندی صفحات وب در سال‌های اخیر ارائه شده است [۵،۷-۹]. این الگوریتم‌ها یکی از اجزای اصلی موتورهای جستجو هستند و هدف آن‌ها تهیه یک رتبه برای هر صفحه وب است و همچنین فضای جستجو را به میزان زیادی کاهش می‌دهند. یکی از الگوریتم‌های رایج برای رتبه‌بندی صفحات وب HITS است [۱] که برای هر پرس‌وجو تحلیل پیوندها انجام می‌شود. الگوریتم HITS یکی از الگوریتم‌های معروف برای رتبه‌بندی صفحات وب بر اساس میزان ارتباط آن‌ها با پرس‌وجوی کاربر است. این الگوریتم از دسته روش‌های وابسته به پرس‌وجو است و برای هر پرس‌وجو تحلیل پیوندها انجام می‌شود که می‌بایست گراف خاص پرس‌وجو به نام گراف همسایگی ساخته شود. از مسائل عمدهٔ رتبه‌بندی بر اساس پیوند که الگوریتم HITS یکی از این الگوریتم‌ها می‌باشد دارای مشکل غنی‌تر شدن اغنیاء[1] می‌باشد. قرار گرفتن همیشه صفحات محبوب[2] در صدر لیست ارائه‌شده به کاربر، باعث می‌شود تا کاربر فقط صفحات خاصی را ببیند و درنتیجه صفحات تازه متولدشدهٔ باکیفیت بالا که کسی به آن‌ها اشاره نمی‌کند نتوانند در دید کاربران قرار گیرند. این مشکل باعث می‌شود صفحات محبوب مرتباً محبوب‌تر شده و تعداد پیوند به آن‌ها افزایش یابد. لذا موتورهای جستجو با ارائه نکردن عادلانه اطلاعات به کاربران، باعث صرف وقت زیاد و درنتیجه کندی تولید علم و دانش خواهند شد. مهم‌ترین چالش در موتورهای جستجو رتبه‌بندی صفحات در پاسخ به پرس‌وجوی کاربر می‌باشد. از طرف دیگر الگوریتم Distance Rank مبتنی بر یادگیری تقویتی است[۱۰] و از فاصله لگاریتمی میان صفحات به‌عنوان توبیخ دریافتی استفاده می‌کند. هدف آن کمینه کردن کل توبیخ‌های دریافتی می‌باشد که باگذشت زمان آگاهی این الگوریتم بیشتر شده و عمل پیمایش را بهتر انجام می‌دهد. الگوریتم Distance Rank در رتبه‌بندی نسبت به بقیه مؤثرتر می‌باشد و حساسیت کمتری به مشکل غنی‌تر شدن اغنیاء نشان می‌دهد. در این مقاله از فرایند اصلی Distance Rank که مبتنی بر یادگیری تقویتی است برای پوشش نقطه ضعف الگوریتم HITS استفاده خواهیم کرد. روش پیشنهادی با سایر الگوریتم‌ها رتبه‌بندی از

1- Rich-get-richer
2- Popular

جمله Distance Rank ،Page Rank ،Wighted Page Rank HITS، بر روی سه مجموعه شامل گراف استاندارد، گراف تصادفی، گراف دانشگاه آزاد اسلامی واحد ارومیه اجرا شده است که نتایج روش پیشنهادی امیدوار کننده است.

ساختار کلی مقاله به‌شرح زیر سازماندهی شده است: در ادامه مقاله ابتدا در بخش دوم کارهای مرتبط با رتبه‌بندی صفحات وب را مورد بررسی قرار خواهیم داد. در بخش سوم مقاله در ابتدا الگوریتم HITS تشریح می‌شود سپس الگوریتم Distance rank تشریح می‌شود. در بخش چهارم، روش پیشنهادی بیان شده است. در بخش پنجم مقاله نتایج حاصل از بررسی و شبیه‌سازی و مجموعه داده‌ها و ارزیابی روش پیشنهادی و معیارهای ارزیابی آورده شده است. در بخش ششم مقاله نتیجه‌گیری کلی ارائه می‌شود.

۲. کارهای مرتبط

در این بخش به پژوهش‌های انجام‌شده درزمینهٔ الگوریتم‌های رتبه‌بندی صفحات وب می‌پردازیم. رضا فتوح و همکاران در [۱۱] به ارزیابی کارایی الگوریتم‌های رتبه‌بندی صفحه برای استخراج صفحات وب پرداخته‌اند که برای حل مشکلات الگوریتم‌های رتبه‌بندی Page Rank(PR) و Weighted Page Rank (WPR) یک الگوریتم جدید به نام Weighted Page Content Rank ارائه شده است. نتایج تجزیه‌وتحلیل نشان می‌دهد که الگوریتم WPR دارای عملکرد بهتری نسبت به الگوریتم Page Rank می‌باشد. در پژوهش دیگر زارع بیدکی و همکاران در[۱۲] به الگوریتم ترکیبی وفقی جهت رتبه‌بندی صفحات وب پرداخته‌اند که در این طرح روش‌های متعدد رتبه‌بندی بیان شده است که نتایج به‌دست‌آمده بهبود چشمگیری را در روش ترکیبی ارائه‌شده در مقایسه با بقیه الگوریتم‌ها نشان داد. فرصتی و همکاران یک الگوریتمی مبتنی بر ساختار پیوندی صفحات و اطلاعات استفاده کاربران برای پیشنهاد صفحات وب در [۱۳] ارائه داده‌اند. آن‌ها پس از معرفی معیار وزن‌دهی، الگوریتمی ترکیبی که از اطلاعات پیمایش کاربران و پیوند بین صفحات به‌منظور پیشنهاد صفحات به کاربران استفاده می‌کند، ارائه کردند. همچنین یک الگوریتم یادگیری مبتنی بر خصیصه‌های ذاتی جهت رتبه‌بندی صفحات وب در [۱٤] ارائه‌شده است. الگوریتم ارائه‌شده با اعمال وزن روی خصیصه‌های صفحات و مقایسه وزن اعمال‌شده باارزش خصیصه صفحات، سعی در یافتن رتبه مناسب به‌منظور دسترسی راحت‌تر کاربران به صفحات باکیفیت از لحاظ محتویات داخلی شده است. جهت ارزیابی مجموعه داده‌ای LETOR به‌کاربرده شده است. نتایج آزمایش‌ها حاکی از آن است که الگوریتم ارائه‌شده در مقایسه با الگوریتم‌های یادگیری که تاکنون ارائه‌شده است نتایج قابل‌توجهی را ارائه می‌دهد. در پژوهش دیگر کلانتری و همکاران در [۱۵] یک الگوریتم رتبه‌بندی صفحات وب مبتنی بر رویکرد یادگیری را ارائه داده‌اند. روش پیشنهادی با استفاده از PHP شبیه‌سازی و برای ارزیابی نتایج از معیار P@N استفاده شد. نتایج ثبت‌شده و ارزیابی نشان می‌دهد که روش پیشنهادی در این طرح دارای کارایی و دقت بالاتری نسبت به روش‌های موجود هستند و همچنین روش پیشنهادی بر مبنای رفتار و قضاوت کاربران و منصفانه است.

سورچی چاوالا[3] یک رویکرد جدید برای تولید مکان یکنواخت منبع‌های کلیک شدهٔ دارای رتبهٔ بهینه با استفاده از الگوریتم ژنتیک بر اساس جلسات خوشه‌بندی‌شدهٔ جستجوی وب جهت جستجوی مؤثر در وبگاه‌های شخصی را ارائه داده است [۱٦] که روش پیشنهادی نتایج امیدوار کننده از خود نمایش گذاشته است. در پژوهش دیگر دویکا سینگ[4] و دایا گاپاتا[5] در [۱۷] یک الگوریتم جدید رتبه‌بندی به نام رتبه‌بندی صفحه بر اساس اولویت کاربر، ارائه و پیشنهاد دادند که از لحاظ ارتباطی کارآمد است زیرا آن عواملی را به کار می‌برد تا صفحات دارای مطالب مرتبط را تعیین کند و همچنین رفتار کاربر در نظر

3- Suruchi Chawla
4- Devika Singh
5- Daya Gupta

گرفته می‌شـود. پاندا[6] و سـوت[7] در [۱۸] بیان می‌کنند که امروزه، شبکه گستردهٔ جهانی یک رسانه محبوب و جذاب برای انتشار اطلاعات است. در این مقاله، در مورد الگوریتم رتبه‌بندی صفحه بحث شـده اسـت که می‌تواند رتبه‌بندی بالاتـری را برای صفحـات مهم وب ایجاد کنـد. در [۱۹] جسـتجوی شـخصی یک حوزهٔ تحقیقی ضروری توسط رانی[8] و سـورانا[9] را بیان کردهانـد که هدف اصلیاش این اسـت که عدم اطمینان و شـک یا تردید واژه‌های جستجو را تشـخیص دهد. در این مقاله، طرحی پیشنهاد داده‌شده اسـت که اولویت‌های مفهومی و ادراکـی کاربر از طریق کلیک کاربران دیگر روی‌داده‌های حاصل از جسـتجو در وب مورد بررسی قرار می‌دهد.

در سـال ۲۰۱۷، سـن[10] و همکاران نسـخه کارآمدی رتبه‌بندی سـاده صفحه با زمان، فاکتور هدایت و مترادف ارائه دادند[۲۰]. این مقاله نسخه بازبینی شده‌ای از رتبه‌بندی صفحـه را با اسـتفاده از همان دسـتورالعمل کلاسـیک و برخی از مواد دیگر اضافه شـده ماننـد زمان، پیوندها، هدایـت، مترادف ارائه می‌دهد کـه آن را برای کاربری که در الگوریتـم پرکاربـرد Rank Ranking محدودیت‌هایـی دارد، وسوسـه‌انگیزتر می‌کند. لاکشـمی[11] و همکاران در سال ۲۰۱۸ مقاله‌ای تحت عنوان هدایت پویا حاصل از نتایج پرس وجو بر اساس نمایه‌گذاری درهم‌سازی با استفاده از الگوریتم بهبود یافته رتبه‌بندی PageRank ارائه دادند[۲۱]. در روش ارائـه شـده در این مقاله یک راه حل براسـاس دسـته‌بندی پرس وجوهای وب به طور پویا و با استفاده از سـاختار نمایه‌گذاری درهم‌سازی داده‌ها پیشنهاد شده اسـت و صفحات وب حاصل با استفاده از الگوریتم بهبود رتبه‌بندی PageRank رتبه‌بندی می‌شوند. با استفاده از این روش، نتایج بسیار غیرمرتبط و نتایج بسیار مهم بر اساس محتوا و تعـداد پیوندها به عنوان نتایـج برتر برای پرس وجو داده شـده، کاهش یافته‌اند. در سال ۲۰۱۹، ستی[12] و دیکسـیت[13] یک سازوکار جدید رتبه‌بندی صفحه براساس الگوهای جسـتجوی کاربر ارائه دادنـد[۲۲]. در این مقاله، یک سـازوکار جدید رتبه‌بندی صفحـه مبتنی بر الگوهای جسـتجوی کاربر و پیوندهای بازدید شده، پیشنهاد شده است. نتایج شبیه‌سازی شده نشـان می‌دهد که سازوکار رتبه‌بندی پیشنهادی بهتر از سازوکار PageRank معمولی در ارائه نتایج رضایت بخش به کاربر عمل می‌کند.

در جدول(۱)، مزایا و معایب هر مدل نشـان داده شده اسـت. مدل‌های مختلفی بـرای رتبه‌بنـدی صفحات وب پیشـنهاد شده اسـت. اما هر مدل می‌تواند مزایا و معایبی داشته باشد.

۳. الگوریتم‌های پایه

در این بخش ما به معرفی دو الگوریتم HITS و Distance Rank خواهیم پرداخت. البته سعی شده است در این بخش هـر یک از الگوریتم‌ها به صورت نظـری و ریاضی مورد بررسی قرار گرفته و سپس در بخش بعدی(٤) به جزئیات روش پیشنهادی پرداخته خواهد شد.

الگوریتـم HITS یک الگوریتم وابسـته بـه پرس‌وجو می‌باشد که در سـال ۱۹۹۸ توسط کلینبرگ[14] ارائه گردید [۱]. ازآنجایـی که این الگوریتم یـک روش مبتنی بر پرس وجو اسـت بـرای هر پرس وجـو تحلیل پیوندهـا انجام می‌شود. برای تحلیل پیوندها، ابتدا گراف خاص پرس وجو به نام گراف همسایگی ساخته می‌شود که در حالت ایده ال شامل صفحات مرتبط یا موضوع پرس وجو می‌باشد. در شکل (۱) شبه کد الگوریتم HITS نشان داده‌شده است.

در این الگوریتم برای سـاخت گراف همسـایگی، ابتدا یک مجموعه از اسناد مرتبط با پرس‌وجو، به‌وسیله موتور جسـت‌وجو واکشی می‌شـوند. به این مجموعه، مجموعه ریشـه گفته می‌شـود. سـپس مجموعه ریشـه به‌وسیله

6- Pande
7- Sote
8- Rani
9- Sorana
10- Sen
11- Lakshmi
12- Sethi
13- Dixit
14- Kleinberg

جدول ۱: مزایا و معایب مدل‌های پیشنهاد شده برای رتبه‌بندی صفحات وب

مراجع	مدل	مزایا	معایب	مدل‌های مقایسه‌ای
[۱۱]	Weighted Page Content Rank	این الگوریتم اطلاعات مهم و مرتبط با پرس وجو را ارائه می‌دهد.	اختصاص وزن به پیوندهای نامرتبط	Page Rank(PR) و Weighted Page Rank (WPR)
[۱۲]	الگوریتم ترکیبی وفقی	کشف پیوندهایی با بیشترین بازدید و بیشترین تعداد مشاهده	افزایش زمان و عدم تشخیص بهترین پیوندها	-
[۱۳]	الگوریتم مبتنی بر وزندهی	الگوریتم بر پایه وزن هر پیوند و نرمال‌سازی آن‌ها کار می‌کند و مجموعه‌هایی از صفحات که مرتبط به پرس وجو هستند را جمع‌آوری می‌کنند.	اختصاص امتیاز به پیوندهایی که خود کار ایجاد شده‌اند و به پرس وجو کاربر مرتبط نیستند.	الگوریتم‌های مبتنی بر وزن
[۱۴]	الگوریتم مبتنی بر وزندهی	اختصاص وزن به پیوندهایی که مرتبط با کلمات کلیدی پرس وجو هستند.	نمایش صفحات نامرتبط با پرس و جوی کاربر به‌دلیل وزن یکسان صفحات	Page Rank(PR)
[۱۵]	الگوریتم رتبه‌بندی صفحات وب مبتنی بر رویکرد یادگیری	با استفاده از رفتار کاربران و استفاده از یادگیری از الگوریتم PageRank در حالت بهینه برای رتبه‌بندی صفحات استفاده می‌شود.	عدم تکرار ناکافی به منظور مرحله یادگیری و آموزش الگوریتم برای تشخیص پیوندهای مهم	Page Rank(PR)
[۱۶]	الگوریتم ژنتیک خوشه‌بندی	خوشه‌بندی پیوندهای برمبنای نتایج اولیه و تعداد نمونه‌های مشابه در پیوندها انجام می‌شود.	عدم کشف مرکزیت خوشه به منظور یافتن پیوندهای مشابه	-
[۱۷]	الگوریتم اولویت	بازیابی پیوندها برمبنای شاخص‌های مهم از قبیل رتبه وبگاه، تعداد کاربران و تعداد بازدید	عدم شباهت‌های ساختاری و میزان ارزشمندی امتیازدهی و سپس عمل رتبه‌بندی به تمامی صفحات	HITS
[۱۸]	الگوریتم اولویت	شناسایی کلمات مهم در پیوندها برمبنای نتایج اولیه و تعداد مشاهدات وبگاه	یافتن صفحاتی که عدم ارتباط پیوندهای مفید میان آن‌ها پایین است.	Page Rank(PR)
[۱۹]	الگوریتم اولویت	در این الگوریتم همه صفحات به همراه سایر صفحاتی که به عنوان پرچسب یا پیوند به این صفحات پیوند داده شده‌اند و همچنین این صفحات به صفحات دیگری پیوند داده شده‌اند به عنوان یک مجموعه در نظر گرفته می‌شوند.	عدم پردازش نشانی پیوندها و نشانی‌های موجود در صفحات وب	Page Rank(PR)
[۲۰]	بهبود الگوریتم رتبه‌بندی	گروه ها یا دسته‌های حاصل را می‌توان براساس ویژگی‌ها به زیردسته‌های دیگری تقسیم نمود.	عدم مشابهت‌های ساختاری رشته‌ای در پیوندها	-
[۲۱]	بهبود الگوریتم رتبه‌بندی	رتبه‌بندی پیوندها برمبنای تعداد رجوع به لینک‌ها و تعداد مشاهدات	یافتن لینک‌های جعلی و نامرتبط با پرس وجوی کاربران	-
[۲۲]	الگوی جستجوی کاربران	برمبنای پرس وجوی کاربران، دنباله‌ای از ضروری‌ترین پیوندها نمایش داده می‌شود.	یافتن صفحاتی که عدم ارتباط پیوندهای مفید میان آن‌ها پایین است.	Page Rank(PR)

همسایگانش تکمیل می‌گردد. همسایه‌ها، مجموعه‌ای از اسناد هستند که یا از اسناد موجود در مجموعه ریشه به آن‌ها پیوند داده‌شده است و یا به اسناد موجود در مجموعه ریشه پیوند داده‌اند. ازآنجا که تعداد اسنادی که به اسناد موجود در مجموعه ریشه پیوند داده‌اند ممکن است عدد بزرگی شود، این عدد محدود و برای تعداد این اسناد حدی در نظر گرفته می‌شود. به این مجموعه جدید، مجموعه پایه یا گراف همسایگی گفته می‌شود. کلینبرگ [۱] صفحات وب را در دو فرم مرجع(Authorities) و قطب(Hub) تعریف می‌کند. به این صورت که یک صفحه مرجع دارای محتوای

authorities $\mathbf{a_p}$ and hubs $\mathbf{h_p}$ of the pages ae stored in the vectors **a** and **h**
V← collection of n pages
N← number of iterations
$\mathbf{z} \leftarrow (\mathbf{1}, \mathbf{1}, \dots, \mathbf{1}) \in \mathfrak{R}^{|V|}$
$\mathbf{a_0} \leftarrow \mathbf{z}$
$\mathbf{h_0} \leftarrow \mathbf{z}$
for i = 0 to N do
{apply **Eq. 5. 27 to** $(\mathbf{a_{i-1}}; \mathbf{h_{i-1}})$ and draw the new authority vector $\hat{\mathbf{a}}_\mathbf{i}$}
{apply **Eq. 5. 28 to** $(\hat{\mathbf{a}}_\mathbf{i}; \mathbf{h_{i-1}})$ and draw the new hub vector $\hat{\mathbf{h}}_\mathbf{i}$}
$\mathbf{a_i} \leftarrow \hat{\mathbf{a}}_\mathbf{i}$
$\mathbf{h_i} \leftarrow \hat{\mathbf{h}}_\mathbf{i}$
end for

شکل ۱: شبه کد الگوریتم HITS[۱]

مرتبط و یک صفحه قطب دارای تعدادی پیوند به صفحات مرجع می‌باشد؛ به‌عبارت‌دیگر یک صفحه با مرجع بالا به‌وسیله تعدادی از صفحات با قطب بالا اشاره‌شده است و همچنین یک صفحه با قطب بالا به تعدادی صفحه با مرجع بالا اشاره می‌کند. اگر فرض کنیم (E,G=V) باشد الگوریتم HITS در دو مرحله زیر عمل می‌کند:

• نمونه‌گیری: در این مرحله مجموعه‌ای از صفحات مرتبط برای پرسش معین جمع‌آوری‌شده است.

• تکرار: در این مرحله مرجع و قطب با استفاده، از مرحله نمونه‌گیری پیدا می‌شوند.

$$H_p=\sum_{q\in I(p)} A_q \quad , \quad A_p=\sum_{B(p)} H_q \qquad (1)$$

که در رابطه (۱) Hq نمره قطب یک صفحه و Aq نمره یک صفحه می‌باشد. I(p) مجموعه‌ای از صفحات مرجع p و B(p) مجموعه‌ای از صفحات که مرجع به اشاره می‌کنند، می‌باشد. این روش مشکلاتی هم به همراه دارد که در زیر به چند مورداشاره می‌کنیم:

• همه اسناد در همسایگی صفحات مرتبط از یک موضوع نبوده و در دسته‌های موضوعی مختلفی قرار دارند.

• تشخیص میان قطب و مرجع آسان نیست، چون در HITS یک صفحه هم می‌تواند قطب خوب و همزمان مرجع خوبی باشد.

• بعضی از وبگاه‌ها به خاطر بالا بردن رتبه مرتباً به یکدیگر اشاره می‌کنند؛ که به آن پهنش رتبه می‌گویند.

الگوریتم Distance Rank مبتنی بر یادگیری تقویتی است که اولین بار ۲۰۰۸ در [۱۰] ارائه گردید. این الگوریتم

```
  //D    is Distance Rank vector
D.←    {∞ , ∞ , ∞ , …}
    t=.;        //iteration number
while δ>ε
          α←e^-β*t
          t=t+۱
           For every page j∈v
                    D_t[j]←(۱-α)*D_t-۱[j]+ min_i (γ*D_t-۱[i]+ log (O[i]))
                    i∈B(j), .≤α≤۱ ,   .≤γ≤۱
       Normalize  D
       δ←//D_t-۱-D_t//
End while
```

شکل ۲: شبه کد الگوریتم Distance Rank

از فاصله لگاریتمی میان صفحات به منظور رتبه‌بندی استفاده می‌کند. منظور از فاصله میان دو صفحه i و j وقتی i به j اشاره کند، لگاریتم درجه خروجی i (تعداد پیوندها) می‌باشد. الگوریتم Distance Rank به‌صورت شکل (۲) محاسبه می‌شود که ε نشان‌دهنده خطا است. بیدوکی[۱۵] در [۱۰] برای درک بهتر این مفهوم تعاریف زیر را ارائه کرده است:

۱) اگر صفحه i به صفحه j اشاره کند وزن پیوند میان i و j برابر است با logO(i) که O(i) نشان‌دهنده درجه خروجی صفحه i می‌باشد.

۲) فاصله لگاریتمی میان i و j عبارت است از وزن کوتاه‌ترین مسیر میان آن‌ها (جمع وزنه‌ای پیوندهای در مسیر) که با dij نشان داده می‌شود. بنابراین به‌جای فاصله عادی، تعریف جدیدی از فاصله به نام میانگین کلیک ارائه شده است. به‌عبارت‌دیگر وزن یک پیوند به‌جای logO(i)، ۱ می‌باشد.

تعاریف بالا نشان می‌دهند که اگر صفحه بارگذاری شده i دارای فاصله d_i باشد، فاصله هرکدام از بچه‌هایش از ریشه (صفحات اشاره‌شده توسط) i درصورتی‌که ورودی دیگری نداشته باشند، با رابطه (۲) محاسبه می‌شود.

$$d_j=d_i+\log O(i) \qquad (2)$$

که در رابطه (۲) di فاصله فرزند j در صفحه i از ریشه می‌باشد و logO(i)، به ازای دریافتی حاصل از انتقال از i به j می‌باشد. در روش فاصله اگر یک صفحه دارای تعداد زیادی پیوند باشد فاصله کمتری نسبت به بقیه دارد و

15- Bidoki

اگر صفحات اشـــاره‌کننده به آن فاصله کمی داشته باشند، فاصله این صفحه نیز کم خواهد بود. برای نشان دادن این مطلب در [۲۳] علاوه بر تعاریف ۱ و ۲ تعریف ســـومی را هم به شکل زیر ارائه کردند:

• اگر d_{ij} نشـــان‌دهنده فاصله لگاریتمـــی میان صفحه i و j باشـــد، میانگین فاصله صفحـــه j از بقیه صفحات با رابطه (۳) محاسبه می‌شود که N نشان‌دهنده تعداد صفحات وب است.

$$d_j = \frac{\sum_{i=1}^{N} d_{ij}}{N} \qquad (۳)$$

بعد از محاســـبه بردار میانگیـــن فاصله تمام صفحات، صفحات مطابق فاصله‌هایشـــان به‌صورت صعودی مرتب می‌شـــوند و لذا صفحات با فاصله کمتر از رتبه بیشـــتری برخوردار خواهنـــد بود [۱۰] که جهت محاســـبه میانگین فاصلـــه برای هر صفحه، یک وابســـتگی میـــان فاصله آن صفحه و صفحاتی که به آن اشاره می‌کنند وجود دارد. با فرض این‌که فقط صفحه i به j اشاره می‌کند، بیدوکی برای محاســـبه فاصله صفحه j (d_j) با استفاده از تعاریف ۱، ۲ و ۳ رابطه (٤) را ارائه کرده است. در رابطه (٤) چون N خیلی بزرگ است، از عبارت $\frac{d_{ii}}{N}$ چشم‌پوشی کردند.

$$d_j=\frac{\sum_{k=1}^{N} d_{kj}}{N}=\frac{\sum_{k\neq i}^{N}(d_{ki}+d_{ij})+d_{ij}}{N}=\frac{\sum_{k\neq i}^{N} d_{ki}}{N}+d_{ij}= \qquad (۴)$$
$$\frac{\sum_{k=1}^{N} d_{ki}-d_{ii}}{N}+d_{ij} \xrightarrow{Eq:} d_j=d_i-\frac{d_{ii}}{N}+d_{ij}\simeq d_i+d_{ij}=d_i+\log(O(i))$$

اگر O(i) نشان‌دهنده تعداد پیوندهای خروجی از صفحه i و B(j) نشـــان‌دهنده مجموعه صفحات اشـــاره‌کننده به j باشند، برای محاســـبه مقدار Distance Rank رابطه (٥) را ارائه کرده‌اند.

$$d_j = \min_i(d_i + \log(O(i))) \; i \in B(j) \qquad (۵)$$

و برای کامل‌تر شـــدن رابطه (٤)، آن را رابطه (٥) توســـعه دادند.

$$d_{j_{t+1}} = (1-\alpha) * d_{j_i} + \alpha * \min_j(\log(O(i)) + \gamma * d_{j_j}) \; i \in B(j), 0 \le \alpha \le 1, 0 \le \gamma \le 1 \qquad (۶)$$

در رابطه (٦) α نرخ یادگیری و γ ضریب نزول می‌باشد بـــه dj_{t+1} فاصله صفحه j در زمان t+1 و d_{it} و d_{jt} به ترتیب نشـــان‌دهنده فاصلـــه صفحـــه i و j در زمان t می‌باشـــند. ضریب نزول برای تنظیم اثرات صفحات قبلی که در مسیر رســـیدن به j قرار دارند، استفاده می‌شـــود. در یک مسیر مانند، l→k→i→j اثر روی با فاکتور γ اعمال می‌شـــود و نرخ یادگیری α با اســـتفاده از رابطه (۷) محاسبه می‌شود که t زمان (شـــماره تکرار) است و ثابت β جهت تنظیم نرخ یادگیری اســـتفاده می‌شـــود. در ابتدا به دلیل این‌که مقدار فاصله هیچ‌یک از صفحات در دســـترس نیست، α مساوی یک بوده و باگذشـــت زمان به‌صـــورت نمایی و نزولی به صفر میل می‌کند.

$$\alpha=e^{-\beta*t} \qquad (۷)$$

رابطـــه (۷) نشـــان می‌دهد که فاصلـــه در هرلحظه به فاصله خود و ورودی‌هایش در لحظه قبلی وابســـته است. به‌عبارت‌دیگـــر عامـــل کاربر در هرلحظـــه انتخاب خود را براســـاس آگاهی‌های قبلی و وضعیـــت فعلی محیط انجام می‌دهد.

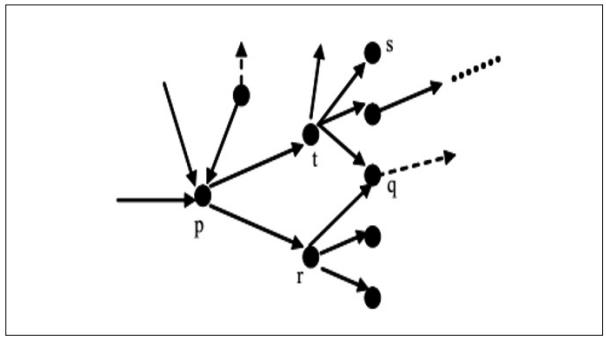
۴. روش پیشنهادی

از مسائل عمدهٔ رتبه‌بندی بر اساس پیوند که الگوریتم HITS یکی از این الگوریتم‌ها می‌باشد دارای مشکل غنی‌تر شـــدن اغنیاء[16] می‌باشـــد. قـــرار گرفتن همیشـــه صفحات محبوب[17] در صدر لیست ارائه‌شده به کاربر، باعث می‌شود تا کاربر فقط صفحات خاصی را ببیند و درنتیجه صفحات تـــازه متولدشـــدهٔ باکیفیت بالا که کســـی به آن‌ها اشـــاره نمی‌کند نتوانند در دید کاربران قرار گیرند. این مشکل باعث می‌شـــود صفحات محبوب مرتباً محبوب‌تر شـــده و تعداد پیوند به آن‌ها افزایش یابد. لذا موتورهای جستجو با ارائه نکردن عادلانه اطلاعات به کاربران، باعث صرف وقت زیاد و درنتیجه کندی تولید علم و دانش خواهند شـــد. مهم‌ترین چالش در موتورهای جستجو رتبه‌بندی صفحات در پاسخ به پرس‌وجوی کاربر می‌باشـــد. در حال حاضر روش‌های

16- Rich-get-richer
17- Popular

جدول ۲: نمونه نرمال شده قطب و مرجع HITS صفحات وب در ۲ تکرار در الگوریتم پیشنهادی

صفحه/وزن	H0	A0	H1	A1	H2	A2	NH2	NA2
A	۱	۱	۱	۵	۶	۲۵	۰,۱۹۳۵	۰,۳۵۷۱
B	۱	۱	۲	۶	۱۱	۳۱	۰,۳۵۴۸	۰,۴۴۲۹
C	۱	۱	۳	۳	۱۴	۱۴	۰,۴۵۱۶	۰,۲۰۰۰



شکل ۳: الگوریتم Distance Rank به دست آوردن میانگین فاصله pq مساوی با dp+log2+log3

رتبه‌بندی موجـــود مانند HITS از مشـــکلاتی مانند غنی‌تر شدن اغنیاء رنج می‌برند. در این بخش الگوریتمی ترکیبی از Distance Rank و HITS برای حل مشکلات فوق ارائه‌شده اســـت. همان‌طور که بیان شـــد الگوریتم Distance Rank مبتنی بر یادگیری تقویتی است و از فاصله لگاریتمی میان صفحات به‌عنـــوان توبیخ دریافتی اســـتفاده می‌کند. هدف آن کمینـــه کردن کل توبیخ‌های دریافتی می‌باشـــد. در این روش یک کاربر به‌صورت موج‌سوار واقعی مدل می‌شود. بدین معنی که در ابتدا کاربر آگاهی کمی از سیســـتم دارد و باگذشـــت زمان آگاهی او بیشتر شده و عمل پیمایش را بهتر انجام می‌دهد. الگوریتم Distance Rank در رتبه‌بندی نســـبت به بقیه مؤثرتر می‌باشـــد و حساســـیت کمتری به مشکل غنی‌تر شدن اغنیاء نشـــان می‌دهد. شکل (۳) نمونه صفحه جدید q و محاسبه میانگین فاصله توسط الگوریتم Distance Rank را نشان می‌دهد.

بنابرایـــن در این بخـــش روش پیشـــنهادی ترکیبی از الگوریتـــم HITS و Distance Rank ارائـــه خواهیم داد. در روش پیشـــنهادی ابتدا الگوریتم HITS اجرا خواهد شـــد و بعـــد از اجرای الگوریتم HITS مقادیـــر صفحات وب را در دو فرم مرجع و قطب به دســـت خواهـــد آمد و همان‌طور که گفته شـــد الگوریتم HITS مشـــکل غنی‌تر شدن اغنیاء را دارد و باید نتایج این الگوریتم با الگوریتم Distance Rank که حساسیت کمتری به مشـــکل غنی‌تر شدن اغنیاء نشان می‌دهد ترکیب می‌شـــود. نحوه ترکیب این الگوریتم چنین خواهـــد بود میزان مرجع و قطب به‌دســـت‌آمده برای تمام صفحات نرمال می‌شـــود یعنی تمام مقادیر مرجع و قطب صفحات براســـاس میزان اعتبارشان نســـبت مجموع کل صفحات مقداردهی می‌شود که در جدول (۲) نمونه نرمال شده برای سه صفحه‌نمایش داده‌شده است.

در جدول (۲) در ابتدا روش پیشـــنهادی نرمال کردن قطب و مرجع الگوریتم HITS را انجام می‌دهد و بعد از نرمال شـــدن نتایج برای ترکیب با نتایج الگوریتم DistanceRank آماده می‌شود و دلیل این نرمال کردن درروش پیشنهادی، ترکیب بهینه و درست براساس درصد مقادیر قطب و مرجع از مقادیر به‌دست‌آمده است با نتایج تولیدشده در الگوریتم DistanceRank می‌باشد که برای این از رابطه جدید روش پیشنهادی به‌صورت رابطه (۸) ارائه می‌شود:

$$dj_{t+1} = (1-\alpha)*dj_i+\alpha*\min_j(\log(O(i))+\gamma*dj_j)+1/(\text{Mean}(NHt,NAt)) \quad (۸)$$

در رابطه (۸) در روش پیشـــنهادی ترکیبی از نتایج هر دو الگوریتم HITS و DistanceRank در رتبه‌بندی صفحات در روش پیشنهادی استفاده شده است. دلیل تقسیم در این رابطه به این دلیل می‌باشد که در الگوریتم DistanceRank بعـــد از محاســـبه بـــردار میانگین فاصله تمـــام صفحات، صفحات مطابق فاصله‌هایشـــان به‌صورت صعودی مرتب می‌شـــوند و لذا صفحات بافاصله کمتر از رتبه بیشـــتری برخوردار خواهند بود که جهت محاســـبه میانگین فاصله برای هر صفحه، یک وابســـتگی میان فاصله آن صفحه و

صفحاتی که به آن اشاره می‌کنند وجود دارد.

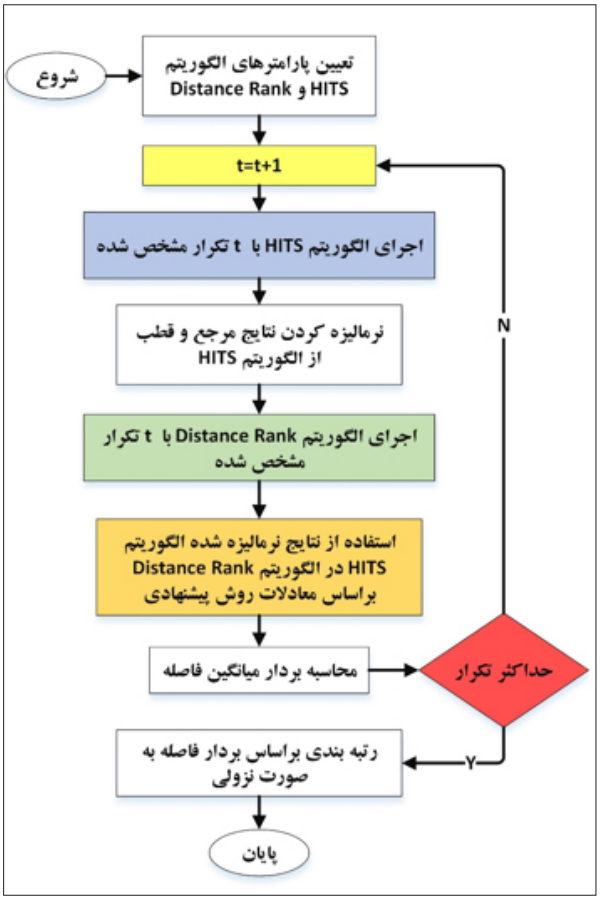
بنابراین هر چه صفحات در الگوریتم HITS دارای مرجع و قطب بالاتری باشـند با تقسیم در بردار محاسبه مقادیر کمتری به دست خواهند آورد و صفحات رتبه‌های بالاتری کسـب می‌کند. یکی نکاتی درروش پیشنهادی وجود دارد مقادیـر t یا تعداد تکرار می‌باشـد که تعـداد تکرار در هر الگوریتم یکسان در نظر گرفته شده است، بنابراین الگوریتم DistanceRank در تکـرار t+1 از نتایج حاصل از الگوریتم HITS استفاده می‌کند. همچنین در رابطه (۸) α نرخ یادگیری و γ ضریب نزول می‌باشد به dj_{t+1} فاصله صفحه j در زمان t+1 و d_{it} و d_{jt} به ترتیب نشـان‌دهنده فاصله صفحه و در زمان می‌باشـند. ضریب نزول برای تنظیم اثرات صفحات قبلی که در مسیر رسیدن به j قرار دارند، استفاده می‌شود. در یک مسـیر مانند، l→k→i→j اثـر l روی j با فاکتور γ اعمال می‌شـود و نرخ یادگیری α با استفاده از رابطه (۷) محاسبه می‌شود که t زمان (شماره تکرار) است و ثابت β جهت تنظیم نرخ یادگیری استفاده می‌شود. در ابتدا به دلیل این‌که مقدار فاصله هیچ‌یک از صفحات در دسترس نیست، α مسـاوی یک بوده و باگذشـت زمان به‌صورت نمایی و نزولـی به صفر میل می‌کند. روندنمای روش پیشـنهادی به‌صورت شکل (٤) می‌باشد.

۵. ارزیابی و نتایج

در ایـن بخش بـه ارزیابی و نتایج روش پیشـنهادی خواهیم پرداخت. در این بخش ابتدا محیط شبیه‌سـازی و مجموعـه داده را معرفی خواهیم کرد. سـپس به ارزیابی روش پیشـنهادی خواهیم پرداخت و نهایتاً به مقایسـه و ارزیابی روش پیشنهادی با مدل‌های دیگر خواهیم پرداخت.

۵.۱. محیط شبیه‌سازی و مجموعه داده

روش پیشنهادی که ترکیب الگوریتم HITS با الگوریتم Distance Rank برای بهبود نتایج در موتورهای جستجو، در محیط سی‌شـارپ و بر روی سیسـتمی بـا پردازنده پنج هسـته‌ای اینتل باقدرت پردازشـی ۲/۳۰ گیگاهرتز و



شکل ۴: روندنمای الگوریتم پیشنهادی

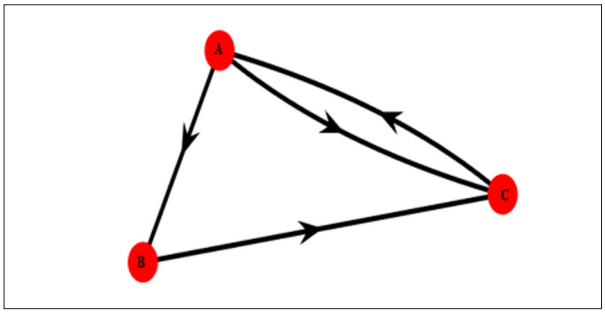
٦ گیگابایـت حافظه موقت شبیه‌سازی‌شـده اسـت. برای ارزیابی روش پیشـنهادی از سـه مجموعه داده که اولین مجموعه مربوط گراف اسـتاندارد که در [۱۱] معرفی‌شده است این مجموعه داده به‌صورت ماتریس و مجموعه پیوند در این گراف وارد سی‌شـارپ شده است. دومین مجموعه داده مربـوط به یـک گراف تصادفی بزرگ اسـت که این گراف به‌صورت تصادفی در محیط سی‌شـارپ تولیدشده است و لازم به ذکر اسـت، داده‌های مورداستفاده در این مجموعه برای شبیه‌سـازی رتبه‌بندی صفحات وب در این مقاله به‌صورت سیستم تصادفی تولید می‌شود و اعتبار و صحت داده‌های ورودی با توابع توزیع محیط سی شارپ بررسی می‌شود تا این‌که یک گراف تصادفی بزرگ مناسب برای رتبه‌بندی تولید شـود و بتوانیم روش پیشنهادی بر روی یک گراف وب بزرگ مورد ارزیابی قرار دهیم. آخرین مجموعه داده مربوط صفحات دانشگاه آزاد اسلامی واحد

ارومیه اســت که برای نمایش ارتبــاط بین این صفحات و درک بهتر پیوندهای ورودی و خروجی بین این صفحات، گراف آن را با توجه به پیوندهای بین این صفحات رســم کردیم.

۲.۵. ارزیابی روش پیشنهادی

در این بخش، برای ارزیابی روش پیشــنهادی از سه مجموعه داده که اولین مجموعه مربوط گراف استاندارد که در مقاله [۱۱] معرفی‌شده است این مجموعه داده به‌صورت ماتریس و مجموعه پیوند در این گراف وارد سی‌شــارپ شده است. دومین مجموعه داده مربوط یک گراف تصادفی بزرگ اســت که این گراف به‌صــورت تصادفی در محیط سی‌شارپ تولیدشده اســت و لازم به ذکر است، داده‌های مورداستفاده در این مجموعه برای شبیه‌سازی رتبه‌بندی صفحــات وب در این فصل به‌صورت سیســتم تصادفی تولید می‌شــود و اعتبار و صحت داده‌های وردی با توابع توزیع محیط ســی شارپ بررسی می‌شــود تا این‌که یک گراف تصادفی بزرگ مناسب برای رتبه‌بندی تولید شود و بتوانیم روش پیشنهادی بر روی یک گراف وب بزرگ مورد ارزیابی قرار دهیم. آخرین مجموعه داده مربوط صفحات دانشگاه آزاد اسلامی ارومیه است که برای نمایش ارتباط بین این صفحات و درک بهتر پیوندهای ورودی و خروجی بین این صفحات؛ گراف آن را با توجه به پیوندهای بین این صفحات رسم شده است. بنابراین به منظور ارزیابی روش پیشــنهادی ما صفحه وب گراف استاندارد را که در شکل (۳) آورده شــده است را با روش پیشنهادی موردبررسی قراردادیــم. برای نمایش ارتباط بیــن این صفحات و درک بهتر پیوندهای ورودی و خروجی بین این صفحات، گراف آن را بــا توجه به پیوندهای بین ایــن صفحات به‌صورت شکل (۵) رسم کردیم.

بعد از حاصل شــدن داده‌های و صفحات مطابق شکل (۵) شکل گراف استاندارد، صفحات وب این مجموعه داده را می‌توان با الگوریتم HITS و Distance Rank و الگوریتم ترکیبی پیشنهادی رتبه‌بندی کرد. جدول (۳) خروجی روش



شکل ۵: گراف استاندارد و نحوه ارتباط بین صفحات

پیشــنهادی و سایر الگوریتم‌ها و نحوه رتبه‌بندی صفحات را نشــان می‌دهد؛ که گراف استاندارد به‌عنوان ورودی در نظر گرفته‌شــده اســت و نحوه ارتباط صفحات به‌صورت یک ماتریس صفر و یک ذکرشــده و صفحات موردنظر را رتبه‌بندی کرده و نمایش می‌دهد.

همان‌طور که در جدول (۳) مشــاهده می‌شود. در این جدول نتایج مقایسه رتبه‌بندی صفحات روش پیشنهادی با سایر الگوریتم‌ها بر روی گراف استاندارد را نشان می‌دهد که نتایج حاصل نشــان می‌دهد هــر دو صفحه A و B در الگوریتم WPR و Pagerank یکســان هستند. صفحه A چون قطب مناسبی است در الگوریتم Hits بالاترین مقدار قطب را گرفته اســت و صفحه C چون یک مرجع مناسب هست بیشترین مقدار مرجع است. الگوریتم Distance بر اساس رتبه‌بندی که انجام داده کمترین مقدار را به صفحه A اختصاص داده است. روش پیشنهادی هم چون ترکیبی از روش Distance و Hits ســعی کــرده یــک حد میانه بین این دو الگوریتم باشــد و صفحه A در رتبه‌بندی اول قراردادهاست و توانسته از الگوریتم Distance برای بهبود نتایج الگوریتم HITS در این مجموعه داده استفاده کند. البته روش پیشــنهادی توانســته از بیــن دو صفحه C و B یک رتبه‌بندی مناسب (صفحه C یک مرجع مناسب است) انجام دهد چراکه صفحه C نیز با رتبه‌بندی بهتر نسبت به صفحه B در رتبه بعد از صفحه A قرارگرفته است.

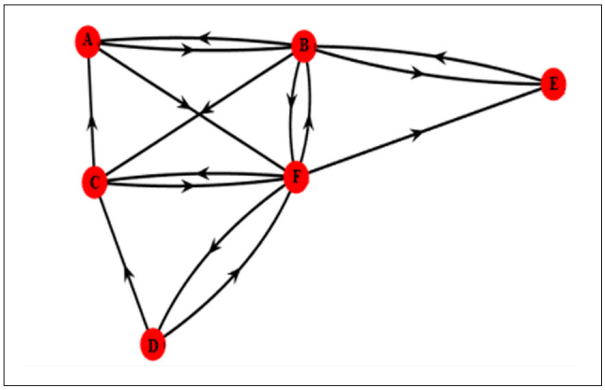
همچنیــن در این بخش مــا از مجموعه داده‌های گراف تصادفی بزرگ که در محیط سی‌شــارپ تولیدشده است، اســتفاده می‌کنیــم؛ بنابراین بــه منظــور ارزیابی روش

جدول ۳: مقایسه رتبه‌بندی صفحات روش پیشنهادی با سایر الگوریتم‌ها بر روی گراف استاندارد

رتبه/صفحه	Pagerank[۱۱]	WPR [۱۱]	HITS		Distance	روش پیشنهادی
			قطب	مرجع		
A	۰,۵۹	۰,۵۹	۰,۵۷۴	۰,۱۱۱	۲,۵	۲,۷
B	۰,۴۲	۰,۴۲	۰,۳۵۷۱	۰,۳۳۳۳	۳,۱۹	۳,۳۷۱۸
C	۰,۵۹	۰,۷۶	۰,۰۷۱۴	۰,۵۵۵۶	۲,۸۵	۳,۰۳۱۸

جدول ۴: تولید ماتریس تصادفی از صفر و یک برای ایجاد گراف تصادفی بزرگ در محیط اجرا

صفحه	A	B	C	D	E	F
A	۰	۱	۰	۰	۰	۱
B	۱	۰	۱	۰	۱	۱
C	۱	۰	۰	۰	۰	۱
D	۰	۰	۱	۰	۰	۱
E	۰	۱	۰	۰	۰	۰
F	۰	۱	۱	۱	۱	۰



شکل ۶: گراف تصادفی بزرگ و نحوه ارتباط بین صفحات

پیشـــنهادی ما صفحه وب گراف تصادفی بزرگ را که در جدول (٤) در محیط سی‌شـــارپ تولید شده است. ماتریس تصادفی در این نمونه شـــامل شش صفحه وب هست که بانام‌های مستعار نام‌گذاری شده است که اگر صفحه وب به صفحه وب دیگر اشاره داشته باشد مقدار آن برابر با یک و در غیر این صورت مساوی صفر خواهد بود. بنابراین این ماتریس به‌صورت تصادفی در محیط نرم‌افزار سی‌شارپ برای شش صفحه وب تولیدشده که آن را در ادامه با روش پیشنهادی موردبررسی و ارزیابی قراردادیم.

در جدول (٤) پیوندهای وردی و خروجی صفحات وب به‌صورت تصادفی ایجادشده است که برای نمایش ارتباط بین این صفحات و درک بهتر پیوندهای ورودی و خروجی بین این صفحات، گراف آن را با توجه به پیوندهای بین این صفحات به‌صورت شکل (٦) رسم کردیم.

بعد از حاصل شدن داده‌ها و ارتباط بین صفحات مطابق جدول (٤) و شکل (٦) گراف تصادفی بزرگ، صفحات وب این مجموعه داده را می‌توان با الگوریتم HITS و Distance Rank و الگوریتم ترکیبی پیشنهادی رتبه‌بندی کرد. جدول (٥) خروجی روش پیشـــنهادی نحوه رتبه‌بندی صفحات را نشـــان می‌دهد؛ که گراف تصادفی بزرگ به‌عنوان ورودی در نظر گرفته‌شده است و نحوه ارتباط صفحات به‌صورت یک ماتریس صفر و یک ذکرشـــده و صفحات موردنظر را رتبه‌بندی کرده و نمایش می‌دهد.

جدول (٦) مقایسه رتبه‌بندی صفحات روش پیشنهادی با ســـایر الگوریتم‌ها بر روی گراف تصادفی بزرگ را در حالت اجرا در محیط سی‌شـــارپ نشان می‌دهد. همان‌طور که در جدول (٦) مشـــاهده می‌شـــود. در این جدول نتایج مقایســـه رتبه‌بندی صفحات روش پیشـــنهادی با ســـایر الگوریتم‌هـــا بر روی گـــراف تصادفـــی و نتایج حاصل را نشـــان می‌دهد. الگوریتم پیشـــنهاد در رتبه‌بندی صفحات، صفحه A را که هم قطب و هم مرجع خوبی دارد را به‌عنوان صفحه اول انتخاب کرده اســـت و صفحه B را که از لحاظ هم مرجـــع دارای کمترین مقدار اســـت به‌عنـــوان آخرین صفحه رتبه‌بندی قرارداده است. درنتیجه روش پیشنهادی در رتبه‌بندی دیگر صفحـــات مانند F ازنظر الگوریتم HITS یک مرجع نســـبتاً قوی و قطب ضعیـــف و از نظر الگوریتم

جدول ۵: رتبه‌بندی صفحات روش پیشنهادی بر روی گراف تصادفی بزرگ

صفحه/اجرا	تکرار۱	تکرار۲	تکرار۳	تکرار۴	تکرار۵
A	۸,۲۱۸۸	۶,۲۸۴۵	۵,۲۲۵۶	۴,۶۸۷۹	۴,۳۸۴۵
B	۱۵,۰۳۱۴	۱۶,۳۲۶۵	۱۵,۷۳۳۴	۱۵,۲۶۲	۱۵,۰۲۰۶
C	۸,۸۳۶۶	۶,۵۶۸۵	۵,۵۶۴۴	۵,۰۲۹	۴,۷۵۸۱
D	۹,۸۵۲۳	۷,۵۷۸	۶,۵۷۰۴	۶,۰۳۲۴	۵,۷۶۱۶
E	۱۱,۸۷۲	۱۰,۰۳۶۴	۸,۹۷۱۳	۸,۴۸۶۹	۸,۲۱۶۹
F	۹,۸۵۵۲	۷,۴۵۵۱	۶,۳۶۶۵	۵,۸۷۸۳	۵,۶۰۸۳

جدول ۶: مقایسه رتبه‌بندی صفحات روش پیشنهادی با سایر الگوریتم‌ها بر روی گراف تصادفی بزرگ

صفحه/رتبه	HITS		Distance	روش پیشنهادی
	قطب	مرجع		
A	۰,۲۳۷۹	۰,۲۷۱	۱,۸۸	۴,۴۳۸۵
B	۰,۱۰۶۴	۰,۰۲۲۹	۱,۸۸	۱۵,۰۲۰۶
C	۰,۱۹۴۱	۰,۲۳۶۵	۱,۷۶	۴,۷۵۸۱
D	۰,۱۴۴۶	۰,۱۷۷۶	۱,۷۶	۵,۷۶۱۶
E	۰,۲۳۹۷	۰,۰۴۳۹	۲,۱۱	۸,۲۱۶۹
F	۰,۰۷۵۵	۰,۲۴۸۱	۲,۱۱	۵,۶۰۸۳

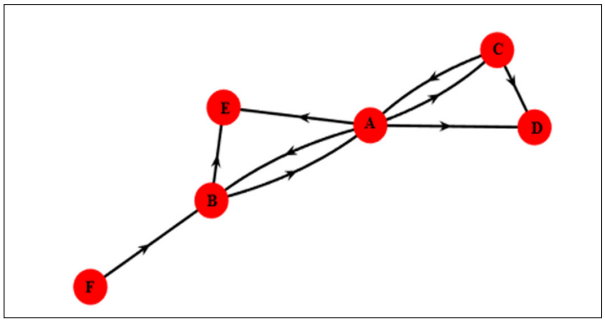
Distance همین صفحه دارای رتبه متوسطی هست را در رتبه دوم قرارداده است و توانسته است یک حد میانه بین دو الگوریتم برقرار کند و بیشتر ضعف رتبه‌بندی الگوریتم Hits را بهبود دهد.

در این مقاله، ما از مجموعه داده‌های دانشگاه آزاد اسلامی ارومیه استفاده می‌کنیم. بنابراین به منظور ارزیابی روش پیشنهادی ما صفحه وب مربوط دانشگاه آزاد ارومیه را که در جدول (۷) آورده شده است را با روش پیشنهادی موردبررسی قراردادیم. در این جدول برای هر یک صفحات یک نام مستعار انتخاب کردیم و همچنین عنوان و نشانی دقیق هر صفحه را در مقابل هر اسم مستعار مشخص کردیم.

همچنین با توجه اطلاعات جدول (۷) و ارتباط مشخص‌شده بین صفحات دانشگاه آزاد اسلامی ارومیه

جدول ۷: صفحات بررسی‌شده با روش پیشنهادی

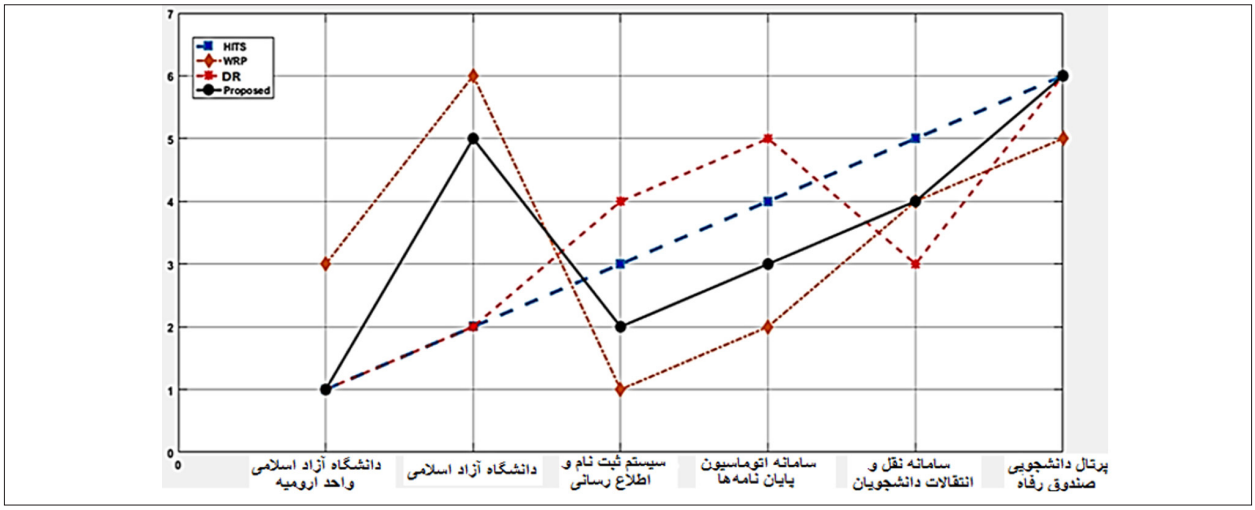
	آدرس صفحه	عنوان صفحه
A	http:// www.iaurmia.ac.ir/	دانشگاه آزاد اسلامی واحد ارومیه
B	http:// www.iau.ac.ir/	دانشگاه آزاد اسلامی
C	http:// amozesh.iaurmia.ac.ir/login.aspx	سیستم ثبت‌نام و اطلاع‌رسانی دانشگاه آزاد
D	http:// thesis.iaurmia.ac.ir/	سامانه اتوماسیون پایان‌نامه‌های دانشجویان
E	http:// transfers.stu...iau. ir/	سامانه نقل و انتقالات دانشجویان
F	http:// bp.swf.ir/	درگاه دانشجویی صندوق رفاه



شکل ۷: گراف دانشگاه آزاد اسلامی ارومیه و نحوه ارتباط بین صفحات

و دانشگاه آزاد اسلامی و سیستم ثبت‌نام و اطلاع‌رسانی دانشگاه آزاد و سامانه اتوماسیون پایان‌نامه‌های دانشجویان و سامانه نقل و انتقالات دانشجویان و درگاه دانشجویی صندوق رفاه، برای نمایش ارتباط بین این صفحات و درک بهتر پیوندهای ورودی و خروجی بین این صفحات، گراف آن را با توجه به پیوندهای بین این صفحات به‌صورت شکل (۷) رسم کردیم.

جدول (۷) رتبه‌بندی روش پیشنهادی بر روی گراف دانشگاه آزاد اسلامی ارومیه را نشان می‌دهد. خروجی روش پیشنهادی نحوه رتبه‌بندی صفحات را نشان می‌دهد؛ که گراف دانشگاه آزاد اسلامی ارومیه به‌عنوان ورودی در نظر گرفته‌شده است و نحوه ارتباط صفحات به‌صورت



شکل ۸: مقایسه روش پیشنهادی رتبه‌بندی با پرس‌وجوی «دانشگاه آزاد اسلامی واحد ارومیه»

جدول ۸: رتبه‌بندی صفحات روش پیشنهادی بر روی گراف دانشگاه آزاد اسلامی ارومیه

صفحه/اجرا	تکرار۱	تکرار۲	تکرار۳	تکرار۴	تکرار۵
A	۸,۰۲۵۹	۵,۷۰۶	۵,۳۸۴	۴,۷۷۵۷	۴,۶۹۵۲
B	۷,۹۴۲۵	۷,۴۸۲۶	۶,۲۳۴۷	۶,۰۲۸۳	۵,۷۰۸۴
C	۱۰,۷۶۱۶	۸,۰۱۸۶	۶,۶۹۴	۶,۴۸۸۲	۶,۱۷۳۴
D	۱۰,۱۹۰۵	۸,۰۷۲۴	۶,۹۶۸	۶,۳۳۹۷	۶,۲۵۱۵
E	۱۰,۱۹۰۵	۷,۳۷۲۴	۶,۹۶۸	۶,۳۳۹۷	۶,۲۵۱۵
F	۴۵,۰۰۶۸	۴۶,۸۸۵۷	۴۶,۵۵۲۹	۴۶,۲۶۴۸	۴۶,۱۳۲۴

یک ماتریس صفر و یک ذکرشده و صفحات موردنظر را رتبه‌بندی کرده و نمایش می‌دهد.

شکل (۸) نتایج به‌دست‌آمده در [۲۱] که با پرس‌وجوی دانشگاه ارومیه و نتایج رتبه‌بندی روش پیشنهادی و الگوریتم‌های WPR، Distance Rank [۲۱] و HITS را با پرس‌وجوی دانشگاه آزاد اسلامی واحد ارومیه مقایسه می‌کند. با توجه به شکل (۸) می‌توان به این نتیجه رسید که روش پیشنهادی با توجه به این‌که صفحه دانشگاه آزاد اسلامی واحد ارومیه در هردو الگوریتم Distance Rank و HITS داری رتبه بالای می‌باشد، این الگوریتم هم همانند دو الگوریتم دیگر عمل می‌کند و صفحه A را به‌عنوان رتبه یک در نظر می‌گیرد و همچنین صفحه F را که داری کمترین رتبه در هردو الگوریتم Distance Rank و HITS می‌باشد، آن صفحه را به‌عنوان آخرین صفحه رتبه‌بندی کرده است.

۳.۵. مقایسه و ارزیابی

برای ارزیابی و مقایسه الگوریتم‌های پیشنهادی از معیارهای ارزیابی AP، P@n و NDC که توسط کالرو جارولین[18] و جانا ککالنین[19] در [۳۹] ارائه گردیده است استفاده کردیم. معیار P@n در رابطه (۹) نشانگر نسبت تعداد اسناد مرتبط در n سند نخست ارائه‌شده به n است. هدف اصلی از این معیار آن است که دقت سیستم از دید کاربران محاسبه شود.

$$P@n= \#ofrelevant\ docs\ in\ top\ n\ results\ /n \quad (۹)$$

معیار P@n از دقت کافی برخوردار نیست. چراکه در این معیار فقط مرتبط بودن یا نبودن یک سند در نظر گرفته می‌شود. رابطه (۱۰) معیار NDCG را نشان می‌دهد. در این معیار که برای n نتیجه نخست استفاده می‌شود، r_j نشان‌دهنده درجه مرتبط بودن سند j با پرس‌وجوی مربوطه می‌باشد.

$$NDCG@n=\sum_{j=1}^{n}\frac{2^{r_j}-1}{\log(1+j)} \quad (۱۰)$$

AP یا میانگین دقت که برای هر پرس‌وجو محاسبه می‌شود، مقدار آن با میانگین P@n برای تمام اسناد مرتبط با پرس‌وجوی موردنظر برابر می‌باشد. در رابطه (۱۱) اگر i– امین سند مرتبط با پرس‌وجو باشد، rel (i) برابر ۱ و در

18- Kalervo Jarvelin
19- Jaana Kekalanien

جدول ۹: مقایسه الگوریتم با معیار P@n

	n=2	n=3	n=4
روش پیشنهادی	۱	۱	۰,۸۲
Distance Rank [10]	۰,۵	۰,۳۳۳	۰,۵
User attention time [23]	۱	۰,۶۶۶	۰,۷۵
Page Rank [24]	۰,۵	۰,۳۳۳	۰,۵
Wighted Page Rank [25]	۰,۵	۰,۶۶۶	۰,۵
HITS [1]	۰,۵	۰,۶۶۶	۰,۷۵

غیر این صورت صفر خواهد بود و N نشان‌دهنده‌ی تعداد نتایج ارائه‌شده برای یک پرس‌وجو می‌باشد.

$$AP=\frac{\sum_{i=1}^{N} P@i \cdot rel(i)}{\#\text{total relevant docs for this query}} \qquad (۱۱)$$

با استفاده از معیارهای اشاره‌شده، روش پیشنهادی را با الگوریتم‌های Dsitance، Weightend Page Rank، HITS، Rank page، Rank و الگوریتم مبتنی بر زمان توجه کاربران ارائه شده در [٦] مقایسه کردیم و که نتایج آن با معیار P@n که نمایانگر دقت الگوریتم در n نتیجه نخست می‌باشد به‌صورت جدول (۹) حاصل گردیده است.

با توجه به جدول (۹) دقت روش پیشنهادی در هر سه حالت (دقت در دو نتیجه نخست، دقت در سه نتیجه نخست و دقت در چهار نتیجه نخست) مناسب بوده است و با معیارهای NDCG و AP نتایج ارزیابی به‌صورت جدول (۱۰) حاصل گردیده است.

در جدول (۱۰) که نمایانگر میانگین دقت و درجه مرتبط بودن نتایج هریک از الگوریتم‌ها می‌باشد، روش پیشنهادی از میانگین دقت (AP) بالایی نسبت به سایر الگوریتم‌ها عملکرد بهتری از خود به نمایش گذاشته است.

۶. نتیجه‌گیری

روش پیشنهادی برای ترکیب الگوریتم HITS با الگوریتم Distance Rank برای بهبود نتایج در موتورهای جستجو شبیه‌سازی شده است. نتایج مقایسه الگوریتم با سایر الگوریتم‌ها بر روی سه مجموعه از داده‌ها نشان

جدول ۱۰: مقایسه روش پیشنهادی با معیارهای NDCG@n و AP

	NDCG@n	AP
روش پیشنهادی	۸,۱	۱
Distance Rank [10]	۶,۰۵	۰,۶۹۹
User attention time [23]	۶,۶۸۱	۰,۹۱۶
Page Rank [24]	۶,۰۵	۰,۶۹۹
Wighted Page Rank [25]	۶,۲۷۹	۰,۸۰۵
HITS [1]	۶,۴۳۱	۰,۸۰۵

داد که روش پیشنهادی در مقایسه با الگوریتم‌های دیگر عملکرد بهتری دارد و توانسته است رتبه‌بندی متفاوت و بهتری نسبت HITS و سایر الگوریتم‌های رتبه‌بندی مانند Distance Rank و PR و WPR داشته باشد. همچنین برای ارزیابی و مقایسه الگوریتم‌های پیشنهادی از معیارهای ارزیابی AP، P@n و NDC استفاده کردیم. می‌توان نتیجه گرفت که با ترکیب یک الگوریتم HITS با الگوریتم Distance Rank برای بهبود نتایج در موتورهای جستجو می‌توان یک الگوریتم قدرتمند برای رتبه‌بندی نتایج موتور جستجو داشت. که نتایج موتورهای جستجو را مطابق با حد میانه میان دو الگوریتم HITS با Distance Rank رتبه‌بندی می‌کند و تعداد خطاهای رتبه‌بندی را تا جای ممکن کاهش می‌دهد.

مراجع

1. Kleinberg, Jon M. «Authoritative sources in a hyperlinked environment.» ACM-SIAM Symposium on Discrete Algorithms. 1998.
2. Gomathi, C., M. Moorthi, and K. Duraiswamy. "Web access pattern algorithms in education domain." Computer and information science 1, no. 4 2008.
3. Joy Shalom Sona, Prof. Asha Ambhaikar "A Reconciling Website System to Enhance Efficiency with Web Mining Techniques: International Journal Of Scientific & Engineering Research Volume 3, 1 ISSN 2229-5518, Issue 2, February-2012.
4. Lin, Chun-Wei, and Tzung-Pei Hong. "A survey of fuzzy web mining." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3, no. 3, 2013.
5. Suthar, Parth, and Bhavesh Oza. "A survey of web usage mining techniques." Int. J. Comput. Sci. Inf. Technol. (IJCSIT) 6, no. 6, 2015.
6. Herrouz, Abdelhakim, Chabane Khentout, and Mahied-

dine Djoudi. "Overview of web content mining tools." arXiv preprint arXiv: 1307.1024, 2013.
7. Du, Ranran, and Yongbin Lin. "Comment ranking by search engine." U.S. Patent Application 10/242,105, filed March 26, 2019.
8. Ananthi, J. "A survey web content mining methods and applications for information extraction from online shopping sites." International Journal of Computer Science and Information Technologies (IJCSIT) 5, no. 3, 2014.
9. Jia, Chen, Yunyan Du, Siying Wang, Tianyang Bai, and Teng Fei. "Measuring the vibrancy of urban neighborhoods using mobile phone data with an improved PageRank algorithm." Transactions in GIS 23, no. 2 (2019): 241-258.
10. Bidoki, Ali Mohammad Zareh, and Nasser Yazdani. "DistanceRank: An intelligent ranking algorithm for web pages." Information Processing & Management 44, no. 2, 2008.

۱۱. فتوحی، رضا، و روح اله عبدی پور، «ارزیابی کارایی الگوریتم‌های رتبه‌بندی صفحـه برای اسـتخراج صفحـات وب»، همایش ملی مهندسـی کامپیوتر و توسـعه پایدار با محوریت شـبکه‌های کامپیوتری، مدل‌سـازی و امنیت سیستم‌ها، مشهد، موسسه آموزش عالی خاوران، ۱۳۹۲.

۱۲. زارع بیدکـی، علی‌محمد، آزادنیـا محمد، «الگوریتـم ترکیبی وفقی جهت رتبه‌بنـدی صفحـات وب»،کنفرانس ملی سـالانه انجمـن کامپیوتر ایران، ۱۳۸۶.

۱۳. فرصتـی، رعنا، و محمدرضا میبدی، «الگوریتمی مبتنی بر سـاختار پیوندی صفحات و اطلاعات اسـتفاده کاربران برای پیشنهاد صفحات وب»، دومین کنفرانـس داده‌کاوی ایران، تهران، دانشـگاه صنعتی امیرکبیر، موسسـه پژوهشی داده‌پردازان گیتا، ۱۳۸۷.

۱۴. قامشـلو، نگین، و فرهـاد مردوخی، «ارائه یک الگوریتـم یادگیر مبتنی بر خصیصه‌هـای ذاتی جهت رتبه‌بندی صفحـات وب»، دومین کنفرانس ملی فنـاوری، انرژی و داده با رویکرد مهندسـی برق و کامپیوتر، کرمانشـاه، انجمن IEEE شاخه دانشجویی کردستان، ۱۳۹۵.

۱۵. کلانتری، میترا، «یک الگوریتـم رتبه‌بندی صفحات وب مبتنی بر رویکرد یادگیری»، دومین همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، تهران ، گروه پژوهشی بوعلی، ۱۳۹۴.

16. Chawla, Suruchi. «A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search.» Applied Soft Computing 46, 2016.
17. Gupta, Daya, and Devika Singh. "User preference based page ranking algorithm." In Computing, Communication and Automation (ICCCA), 2016 International Conference on, pp. 166-171. IEEE, 2016.
18. Sote, A. M., and SR Pande. "Application of Page Ranking Algorithm in Web Mining." In International Conference on Advances in Engineering & Technology–2014. 2014.
19. Rani, S. Geetha, and M. Sorana Mageswari. "A link-click-concept based ranking algorithm for ranking search results." Indian Journal of Science and Technology 7, no. 10, 2014.
20. Sen, Tuhena, Dev Kumar Chaudhary, and Tanupriya Choudhury. "Modified Page Rank Algorithm: Efficient Version of Simple Page Rank with Time, Navigation and Synonym Factor." In Computational Intelligence and Networks (CINE), 2017 3rd International Conference on, pp. 27-32. IEEE, 2017.
21. Lakshmi, L., P. Bhaskara Reddy, and C. Shoba Bindu. "Dynamic Navigation of Query Results Based on Hash-Based Indexing Using Improved Distance PageRank Algorithm." In Artificial Intelligence and Evolutionary Computations in Engineering Systems, pp. 213-221. Springer, Singapore, 2018.
22. Sethi, Shilpa, and Ashutosh Dixit. "A novel page ranking mechanism based on user browsing Patterns." In Software Engineering, pp. 37-49. Springer, Singapore, 2019.
23. Xu, Songhua, Yi Zhu, Hao Jiang, and Francis CM Lau. "A User-Oriented Webpage Ranking Algorithm Based on User Attention Time." In AAAI, vol. 8, pp. 1255-1260. 2008.
24. Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: Bringing order to the web". Stanford InfoLab, 1999.
25. Xing, Wenpu, and Ali Ghorbani. "Weighted pagerank algorithm." In Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on, pp. 305-314. IEEE, 2004.