تحلیل ساختار و الگوریتم ذخیره و بازیابی اطلّاعات در پایگاه‌های استنادی وبی

عبدالرّسول خسروی*
عضو هیئت علمی،
دانشگاه علوم پزشکی بوشهر

رحمت‌الله فتّاحی
استاد گروه کتابداری و اطلاع‌رسانی،
دانشگاه فردوسی مشهد

دریافت: ۱۳۸۸/۱۰/۱۵ | پذیرش: ۱۳۸۹/۰۱/۲۲

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۵۲۰۶-۱۷۳۵
شاپا(الکترونیکی) ۵۵۸۳-۲۰۰۸
نمایه در LISA، SCOPUS و ISC
http://jist.irandoc.ac.ir
دوره ۲۶ | شمارهٔ ۲ | ۱۹۹ - ۲۲۲
زمستان ۱۳۸۹

چکیده: از عمده تلاش‌هایی که طی سالیان اخیر برای پاسخگویی به نیازهای علمی محققان، پژوهشگران و دانشجویان در محیط وب صورت گرفته است، می‌توان به توسعه پایگاه‌های استنادی اشاره داشت که به منزله یک رویکرد جدید در ذخیره و بازیابی اطلاعات به شمار می‌آید. قابلیت‌های درخور توجه این پایگاه‌ها، امکانات بسیاری را در اختیار پژوهشگران چه در زمینه جستجو و بازیابی اطلاعات و چه در عرصه علم سنجی قرار داده است. هدف این مقاله که به روش کتابخانه‌ای و مطالعه موردی (مشاهده مستقیم پایگاه‌ها) انجام گرفته، بررسی ساختار و قابلیت‌های جستجو و بازیابی اطلاعات در پایگاه‌های استنادی می‌باشد. این امر با تحلیل ساختار نمایه‌سازی پایگاه‌های استنادی و اینکه این پایگاه‌ها از چه سیستم‌ها والگوریتم‌هایی برای وزن‌دهی، نمایه‌سازی و برقراری پیوند میان استنادها و تحلیل هم استنادی نویسنده استفاده می‌کنند مورد بررسی قرار گرفته است. نتایج این بررسی نشان داد که این نوع پایگاه‌ها از ساختار شبکه‌ای برخوردارند و از دو سیستم نمایه‌سازی استنادی خودگردان و پاب سرچ و نیز الگوریتم مورد استفاده آن‌ها بر اساس مدل‌های وزن‌دهی و بردار فضایی و خوشه‌بندی سلسله‌مراتبی انباشتگی است، که در متن مقاله بطور مشروح به آن‌ها اشاره شده است.

کلیدواژه‌ها: *نمایه‌سازی استنادی، پایگاه‌های استنادی وبی، بازیابی اطلاعات، نظام نمایه‌سازی استنادی خودگردان[1]، سیستم[2] پاب سرچ، الگوریتم‌های ذخیره اطلاعات*

* پدیدآور رابط: khosravi2422@gmail.com

1. Autonomous citation index- 2. PubSearch

مقدمه

سال‌ها پژوهشگران و دانشجویان برای یافتن اطلاعات خود از سه راهبرد مهم جستجو استفاده قرار کرده‌اند. راهبرد نخست از طریق دنبال‌کردن دستی و ساده منابعی که در تحقیقات یا موضوعات مشابه توسط دیگران مورد استفاده قرار گرفته است یا به عبارتی استنادها، بوده است. استنادها در آثار علمی جایگاه ویژه‌ای دارند. در واقع یک مقاله علمی زمانی معتبر است که به آثار و متون مرتبط با آن موضوع استناد نماید. این روش، هرچند هنوز توسط برخی از پژوهشگران مورد استفاده قرار می‌گیرد اما به دلیل محدودیت‌های فراوان از قبیل محدودیت زمانی، روزآمد نبودن، عدم دستیابی نویسنده به همه منابع موجود و برخی از موارد دیگر عملاً بطور کامل جوابگوی نیازهای پژوهشگران نیست.

راهبرد دوم مورد توجه، جستجوی کلیدواژه‌ای موضوعی در نمایه‌نامه‌ها و چکیده‌نامه‌ها و پایگاه‌های اطلاعاتی به منظور یافتن منابع مرتبط می‌باشد. این روش نیز بدلیل عدم آشنایی کامل نویسندگان و محققان با راهبردهای جستجو و برخی از کاستی‌های دیگر بطور کامل پاسخگوی نیازهای محققان نبوده است. هرچند که پایگاه‌های اطلاعاتی به دلیل روزآمدی و اطلاعات نو و پویا خدمات ارزنده‌ای به محققان ارائه داده و می‌دهند.

پژوهشگرانی که این دو راهبرد نیاز آن‌ها را برآورده نمی‌ساخت به راهبرد سوم یعنی استفاده از کارشناسان موضوعی روی می‌آوردند. این روش بدلیل تسلط موضوعی محقق و روزآمدی اطلاعات، از شیوه‌های موثر مورد استفاده می‌باشد، اما به دلیل محدودیت و کمبود کارشناسان و متخصصان موضوعی، استفاده از این شیوه برای همگان میسر نبوده است.

بدنبال گسترش ابزارها و شیوه‌های بازیابی اطلاعات، امروزه بسیاری از پایگاه‌های اطلاعاتی راهبرد چهارمی را پیش روی پژوهشگران قرار داده‌اند که تحت عنوان نمایه استنادی مورد توجه قرار گرفته است.

استناد بیانگر نوعی استفاده از دانش پیشین است. نوشته علمی فقط بر خود متکی نیست، بلکه بر بسیاری از منابع پیشین استوار است. نمودار شدن منبعی در سیاهه یک اثر بازگوکننده این مطلب است که در ذهن نویسنده ارتباطی میان اثر وی و مقاله‌ای که در فهرست مآخذ خود به آن استناد کرده است وجود دارد. بدین وسیله پایگاه‌های استنادی با مبنا قرار دادن استنادها شروع به نمایه‌سازی و فعالیت‌های مرتبط با آن‌ها نمودند.

می‌توان گفت مهمترین مزیت نمایه‌های استنادی در این پایگاه‌ها، توانایی ردگیری و یافتن مقالات هسته و در حقیقت شناسایی تولیدکنندگان مقاله‌های علمی است. هرچند می‌توان به لحاظ تاریخی، اساس و سابقه طولانی مقوله استناد را در میان مسلمانان نیز یافت که به واسطه علم‌الحدیث

و بحث شناسایی روات و سلسلهٔ احادیث از قدیم علماء و بزرگان ما اهمیت زیادی بـه ایـن مباحـث می‌دادند (حری، ۱۳۸۵)، ولی ضوابطی مدون که بتوان آن‌ها را مبانی نظری چنین امری تلقی کرد، وجود ندارد.

بحث ایجاد پایگاه‌های استنادی وب پایه برای اولین‌بار در سـال ۱۹۹۵ بـه عنـوان تنهـا پایگـاه جهانی ISI توسط گارفیلد شروع شد (نوروزی، ۲۰۰۵). امروزه پایگاه‌ها و نمایه‌های استنادی چنـان رشدی نموده است که به عنوان یکی از ارکان پژوهش و تولید علـم در هـر کشـوری مـورد توجـه قرار گرفته است. به‌رغم جایگاهی که این پایگاه‌ها در بین محققـان دارنـد، آنچـه کـه مـورد توجـه کتابداران و نمایه‌سازان است نحوه و ساختار ذخیره و بازیابی اطلاعات موجود در این پایگاه‌هاست و این‌که چنین پایگاه‌هایی از چه الگـوریتم‌هـایی بـرای ذخیـره و بازیـابی اطلاعـات خـود اسـتفاده می‌کنند، چگونه فیلدهای مختلف به هم ربط پیدا نموده و سبب برجسته شدن این نوع پایگاه‌هـا در ردگیری استنادها گردیده است. از این رو، هدف اصلی این مقاله، بررسی ساختار ذخیره و بازیـابی اطلاعات در پایگاه‌های استنادی است تا ضمن تحلیل ساختار و الگوریتم ذخیره و بازیابی اطلاعات در پایگاه‌های استنادی وبی به این نکته بپردازد که پایگاه‌های استنادی امروزه از چـه الگـوریتم‌هـا، راهبردها و شیوه‌هایی برای ذخیره و بازیابی اطلاعات کتابشناختی استفاده می‌کنند.

بر این اساس، پرسش‌های زیر مورد توجه پژوهش حاضر می‌باشد:

۱. پایگاه‌های استنادی وبی از چه قابلیت‌هایی برای جستجو و بازیابی اطلاعات برخوردارند؟
۲. پایگاه‌های استنادی وبی از چه سیستم‌ها و الگوریتم‌ها و روش‌هایی برای ذخیـره و بازیـابی (ردگیری) اطلاعات استنادی استفاده می‌کنند؟

روش مورد مطالعه

این پژوهش با استفاده از روش مطالعه موردی، مشاهده مستقیم و تحلیل پایگاه‌ها و نیز مطالعه کتابخانه‌ای انجام گردید. بدین صورت که در روش مطالعه موردی، به مشاهده مستقیم پایگاه‌ها، و تحلیل قابلیت‌های جستجو و بازیابی اطلاعات در زمینه چگونگی نمایه‌سازی استنادی پرداخته شـد. امکانات هر یک از پایگاه‌ها، نحوه نمایش اطلاعـات و قابلیـت‌هـای بازیـابی مـورد شناسـائی قـرار گرفت. سپس برای تحلیل الگوریتم‌ها و نظام‌های نمایه‌سازی از روش مطالعـه کتابخانـه‌ای اسـتفاده گردید. از طریق بررسی متون، نظام‌های ذخیره و بازیابی و هچنین الگـوریتم‌هـای مـورد اسـتفاده و روش‌های نمایه‌سازی مورد شناسائی و بررسی قرار گرفت.

پایگاه‌های مورد بررسی در این پژوهش شامل سه پایگاه استنادی سایت سیر[1]، آی اس آی[2] و اسکوپوس[3] می‌باشد

1. CiteCeer 2. ISI Web of Knowledge 3. Scopus

پیشینه پژوهش

مطالعات انجام گرفته در داخل کشور نشان می‌دهد که در خصوص الگوریتم‌های ذخیره و بازیابی اطلاعات تحقیقات و نوشته‌های منسجم و جامعی وجود ندارد. اما به طور کلی در خارج از ایران مطالعات گوناگونی در زمینه الگوریتم‌های پایگاه‌های اطلاعاتی انجام گرفته است.

کامرون[1] (۱۹۹۷) یک پایگاه استنادی و کتابشناختی جهانی مبتنی بر وب پیشنهاد داد که این پایگاه قادر باشد که به هر اثر علمی نوشته شده و استنادهای آن ارتباط برقرار کند. وی پیشنهاد کرد این پایگاه در سطح جهانی و از طریق اینترنت در دسترس همگان باشد.

گیلز[2] (۱۹۹۸) نتایج آزمایش خود را که بر روی سایت سیر انجام داده بود، در مقاله‌ای تحت عنوان «سایت سیر: سیستم نمایه‌سازی استنادی خودکار» به تشریح کار این سیستم پرداخته و بطور مفصل الگوریتم‌ها و کارکردهای این سیستم اشاره داشته است. عمده ویژگی‌های این سیستم را صرفه‌جویی در زمان، خودکاربودن و مرور محتوی استنادها می‌داند. همچنین در این مقاله توانائی‌های این سیستم را تشریح می‌کند.

بولاکر[3] (۱۹۹۸) در مقاله‌ای تحت عنوان«سایت سیر: واسط خودکار برای بازیابی خودکار و تعیین انتشارات مورد علاقه» سایت سیر را به عنوان یک واسط کمکی خودکار در ارتقاء فرآیند یافتن انتشارات علمی بر روی وب بسیار سودمند می‌داند.

هی[4] (۲۰۰۱) نیز در مقاله «پاب سرچ: سیستم بازیابی مبتنی بر استناد وبی[5]» به چگونگی کارکرد این سیستم در خصوص جستجوی خوشه‌ای مدرک و هم استنادی نویسنده و معماری و نیز سایر کارکردها اشاره کرده است.

هی(۲۰۰۲) نیز در مقاله «واکاوی یک پایگاه استنادی برای خوشه‌بندی مدرک» یکی از راه‌های نمایه‌سازی متون در وب را نمایه‌سازی از طریق پایگاه‌های استنادی می‌داند. وی یک فرآیند واکاوی به منظور استخراج دانش خوشه‌ای مدرک از پایگاه‌های استنادی به منظور پشتیبانی از بازیابی اطلاعات در وب را پیشنهاد می‌کند. وی تکنیک‌های واکاوی استفاده شده برای ایجاد خوشه مدرک مبتنی بر نقشه خود–سازماندهی کوهنن[6] (KSOM) و تئوری رزونانس تطبیقی فازی[7] می‌داند. وی تکنیک‌های پیشنهادی را که ترکیب شده و یک سیستم به نام PubSerach برای انتشارات علمی و بی معرفی می‌کند.

هی (۲۰۰۲ب) همچنین در مقاله‌ای دیگر «واکاوی یک پایگاه استنادی برای تحلیل هم استنادی نویسنده» یک فرآیند واکاوی خودکار برای تحلیل هم استنادی نویسنده(ACA) مبتنی بر پایگاه‌های استنادی پیشنهاد می‌کند. سیستم کارآ برای این منظور را PubSearch معرفی می‌کند.

---

1. Cameron
2. Giles
3. Bollacker
4. He
5. PubSearch: a web citation – based retrieval system
6. Kohonen's self organization Map
7. Fuzzy Adaptive Resonance Theory

بررسی متون نشان می‌دهد که پایگاه‌های استنادی مورد توجه پژوهشگران قرار گرفته است. پایگاه‌های استنادی نیز برای فراهم آوردن زمینه‌های استفاده پژوهشگران از سیستم‌های نمایه‌سازی استنادی خودکار استفاده می‌کنند. همچنین برای بازیابی دقیق‌تر از الگوریتم‌های وزن‌دهی و خوشه‌بندی هم استنادی نویسندگان و مدرک استفاده می‌کنند.بر این اساس، مطالعه حاضر، که نخستین مورد در زبان فارسی است، می‌تواند برای کتابداران و نیز متخصصان علوم رایانه در ایران سودمند باشد.

نمایه استنادی

نمایه استنادی، شامل فهرستی از مقالات و نمایه‌های مرتبط با آن مانند نویسندگان و کلیدواژه‌های موضوعی هریک از مقالات منتشرشده‌ای است که به آن مقالات استناد کرده‌اند. به عبارت دیگر، در مورد یک مقاله خاص، نمایه استنادی مشخص می‌کند که این مقاله به کدام منابع پیش از خود استناد کرده و همچنین بعدها توسط چه مقالات دیگری که پس از آن منتشر شده‌اند، مورد استناد قرار گرفته است. در واقع، از نمایه‌های استنادی بصورت شبکه استنادی و جایگاه هر مقاله به منزله یک گره در این شبکه یاد می‌شود. این نوع نمایه، علاوه بر بازنمون محتوایی، ارتباط یک مدرک را با مدرک دیگر از طریق استنادها مشخص می‌کند. از طریق ردگیری استنادها می‌توان ارزش مقاله‌های علمی را به لحاظ میزان استفاده از آن مشخص کرد (حسن زاده و نوروزی چاکلی، ۱۳۸۷). در این نوع نمایه‌سازی، مقاله استناد شده بوسیله ماخذ استناد مشخص می‌شود (گارفیلد، ۱۹۶۴). نمایه‌های استنادی در اصل برای بازیابی اطلاعات ایجاد شدند (گارفیلد،همان). نمایهٔ استنادی امکان ردیابی گذشته‌نگر (فهرست مقالات استناد شده) و آینده‌نگر (مقالات بعدی که به مقاله‌ای معین استناد کرده‌اند) را فراهم می‌سازد.

از نمایه‌های استنادی به منظورهای مختلفی می‌توان بهره جست؛ مثلاً:

الف) استنادها به بازیابی به منزله سرنخ‌هایی برای شناسائی مقالات مرتبط استفاده می‌شود؛

ب) جایگاه استنادها در انتشارات استنادکننده می‌تواند در قضاوت پیرامون سهم عمدهٔ یک مقالهٔ استناد شده و مناسبت آن با درخواستی معین مفید باشد (گارفیلد۱۹۹۴، سالتون۱۹۷۱)؛

ج) نمایهٔ استنادی با تعیین جایگاه و تعداد دفعات استناد به مقاله‌ای خاص در متون، به منزله شاخص اهمیت آن مقاله به شمار می‌رود؛

د) نمایهٔ استنادی تحلیل عمقی تحولات پژوهشی را امکان‌پذیر و حوزه‌های نوپدید علمی را شناسایی می‌کند.

پایگاه‌های استنادی وبی

امروزه حجم انبوه اطلاعات، به ویژه اطلاعات موجود در شبکه جهانی وب، موجب ظهور شیوه‌های جدیدی در بازیابی اطلاعات گردیده است. یکی از جنبه‌های مهم ظهور و گسترش وب، ایجاد پایگاه‌های استنادی بود که به دلیل ماهیت پویا و ویژگی‌های منحصر به فرد وب، در ساختار آن‌ها تغییراتی نسبت به نمایه‌های چاپی ایجاد شده است.

نمایه‌های استنادی در پایگاه‌های استنادی وبی برای جستجو و بازیابی اطلاعات به عنوان یک منبع ارزشمند تلقی می‌شود. عمده پایگاه‌های شناسائی شده که بیشتر مورد توجه محققان قرار می‌گیرد عبارتند از: سایت سیر[1]، اسکوپوس[2]، آی اس آی[3]. هرچند برخی از پایگاه‌ها و موتورهای جستجوی دیگر هستند که قابلیت‌های پایگاه‌های استنادی را به عنوان امکانات دیگر به قابلیت‌های خود افزوده‌اند مثل: گوگل اسکالر (دانشوران)[4]. بطور کلی پایگاه‌های استنادی وبی مجموعه‌ای از داده‌های استنادی را در مخزن خود جمع آوری نموده و برای ذخیره و نمایه‌سازی آن‌ها از تکنیک‌ها و الگوریتم‌های متفاوتی استفاده می‌کنند.

پاسخ به پرسش‌های پژوهش

۱. پایگاه‌های استنادی وبی از چه قابلیت‌هایی برای جستجو و بازیابی اطلاعات برخوردارند؟

در بررسی وضعیت جستجو در پایگاه‌های استنادی نظیر اسکوپوس، آی اس آی و سایت سیر عمده قابلیت‌های شناسائی شده آن‌ها در دو مقوله جستجو و بازیابی تقسیم گردید، که عمده قابلیت‌ها و امکانات در جدول شماره ۱ و جدول ۲ آمده است. همانگونه که جدول‌های زیر نشان می‌دهند، برخی از امکانات جستجو مورد استفاده همه پایگاه‌های اطلاعاتی است، اما برخی از امکانات به طور خاص تنها در برخی پایگاه‌های استنادی به لحاظ کارکردشان تعبیه شده است.

جدول شماره ۱: بررسی امکانات و قابلیت‌های جستجو در پایگاه‌های مورد بررسی

نام پایگاه	ساده	پیشرفته	نویسنده	مراجع	سازمانی	مرور	تاریخچه جستجو	سیستم هشدار	آنالیزور مجلات
اسکوپوس	√	√	√		√	√	√	√	√
آی اس آی	√	√	√	√	√		√	√	√
سایت سیر	√		√	√					

1. Citeseer 2. Scopus 3. ISI 4. Google Scholar

جدول شماره ۲: بررسی امکانات و قابلیت‌های بازیابی و شیوه نمایش در پایگاه‌های مورد بررسی

نام پایگاه	تاریخ	ربط	تعداد استناد	اولین نویسنده	عنوان منبع	حوزه‌های موضوعی	نوع مدرک	نویسندگان	موسسات	ردگیری استناد	زبان	کشور
اسکوپوس	√	√	√	√	√	√	√	√	√	√	√	√
آی اس آی	√	√	√	√	√	√	√	√	√		√	
سایت سیر			√									

۲- پایگاه‌های استنادی وبی از چه سستم‌ها و الگوریتم‌ها و روش‌هایی برای ذخیره و بازیابی (ردگیری) اطلاعات استنادی استفاده می‌کنند؟

به طور معمول، سازماندهی فایل در نمایه‌های استنادی نیز همچون پایگاه‌های دیگر بر مبنای اصل فایل مقلوب انجام می‌شود (پائو، ۱۳۷۸). در نمایه‌های استنادی دو فایل اصلی وجود دارد: نمایه مآخذ[1]، نمایه استنادی[2]. نمایه مآخذ دربردارنده توصیف کامل کتابشناختی همه مقاله‌های نمایه‌شده می‌باشد. این نمایه بر اساس نام نویسندگان اصلی الفبایی می‌شود. نمایه استنادی فایلی است که براساس آن نام نویسندگان مقاله‌ها (یعنی مقاله‌هایی که توسط مقاله‌های اصلی در نمایه مآخذ مورد استناد قرار گرفته‌اند) تنظیم یافته است. این نمایه به منزله فایل مقلوب نمایه مآخذ عمل می‌کند. ساختار پیشینه‌های مربوط به مدارک مآخذ شامل فیلدهای معمولی یعنی نویسنده، عنوان، مجله و مانند آنهاست. نمایه استنادی فایل را بر اساس عناصر داده‌ای که در مآخذ استناد شده موجود است مقلوب می‌کند. برای جستجوی نمایه استنادی قالبی استاندارد به کار گرفته می‌شود. بطور کلی، ساختار نمایه‌های استنادی از سازماندهی مشخصات کامل انواع اطلاعاتی که باید در فایل‌های جداگانه انباشته شود تشکیل می‌شود که این اطلاعات شامل مقدار داده‌های ذخیره شده در هر پیشینه، ساختار پیشینه، رابطه میان عناصر مختلف داده‌ها، امکان ذخیره فایل، چگونگی ذخیره پیشینه‌ها تشکیل شده است. بسیاری از پایگاه‌های استنادی نظیر ISI برای ذخیره‌سازی استنادها از جدول‌هایی استفاده می‌کنند(هی ۲۰۰۱).

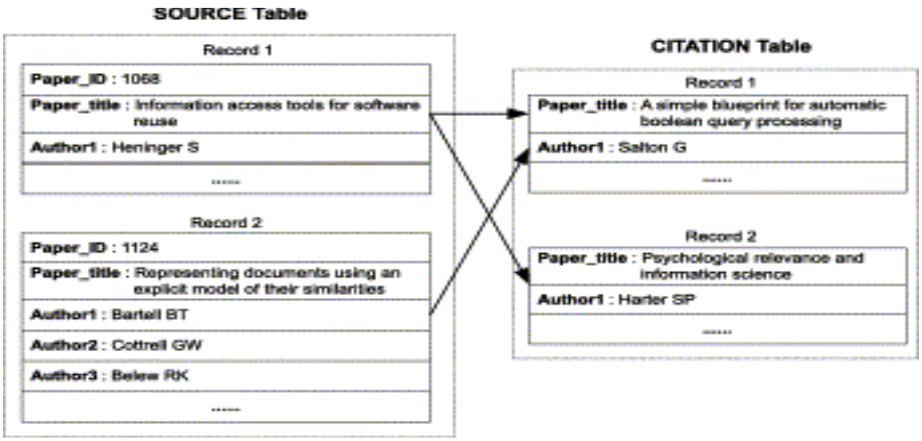
هی (۲۰۰۱) همچنین در مقاله‌ای تحت عنوان«واکاوی پایگاه استنادی برای تحلیل هم استنادی نویسنده» به این نکته اشاره می‌کند که پایگاه استنادی از دو جدول منبع[3] و جدول استنادها[4] برای برقراری ارتباط بین استنادهای منابع استفاده می‌کند. جدول منبع، اطلاعات کتابشناختی مقالات و

1. Source Index
2. Citation Index
3. SOURCE Table
4. CITATION Table

جدول استناد، استنادهای استخراج شده از منابع مورد استفاده در مقالات شامل می‌شود. عمده خصیصه‌های دو جدول داده‌های تعریف شده مشخصی نظیر عنوان مقاله، نام نویسنده، نام مجله، جلد، شماره، صفحه و سال انتشار در بر می‌گیرد. لینک URL نشانی مکان یاب وبی مدرک است که از این طریق امکان دسترسی به متن کامل فراهم می‌کند. ID مقاله در جدول منبع و ID استناد در جدول استناد از کلیدهای اصلی به ترتیب در این دو جدول به شمار می‌روند. Source ID از جدول استناد با Paper ID پیوند می‌خورد، تا نشان دهد که به مقاله خاصی که در جدول استناد ذخیره شده است، استناد کرده است. همانطوریکه در تصویر شماره ۱ نشان داده شده است، اکثر فیلدهای جدول منبع با جدول استناد با هم مشابه هستند. همچنین باید یادآور شد که فقط سه نویسنده اول در پایگاه استنادی ذخیره شده‌اند، زیرا فرض بر این است که نویسنده چهارم دخالت کمتری داشته است.

تصویر ۱. ساختار پایگاه از پایگاه استنادی وبی (منبع: He,2000)

نمونه‌ای از پیشینه‌های ذخیره شده در جدول منبع و استناد در تصویر شماره ۲ نشان داده شده است. همانطور که می‌بینید ارتباطات در این پیشینه‌ها چند به چند است.



تصویر ۲: نمونه‌ای از پیشینه‌های ذخیره شده در منبع و استناد که مبتنی بر مقالات بازیابی شده در حوزه بازیابی اطلاعات(IR) ایجاد شده است

با توجه به ساختار کلی نمایه‌های استنادی، خط مشی پایگاه‌های استنادی در ارتباط با نمایه‌سازی شامل موارد زیر است که این عملیات بصورت خودکار انجام می‌گیرد. وجود یک یا چند سرور دارای برنامه‌های نرم‌افزاری روبات، نیز شیوه‌ها و راهبردها و تکنیک‌های نمایه‌سازی در پایگاه‌های استنادی وبی به منظور اقدامات زیر مورد نیاز است:

- مرور منابع و استنادها و پیوندهای آنان و تجزیه مقالات برای یافتن استنادها؛
- نمایه‌سازی فیلدها؛
- نمایه‌سازی استنادها و مراجع؛
- مرتبط کردن استنادها به هم؛
- انتخاب کلید واژه‌ها بر اساس الگوریتم خاص؛
- ارسال آن‌ها به پایگاه خاص و مرتب‌سازی آن‌ها؛
- برقرای لینک به مقالات و وزن‌دهی آن‌ها؛
- آماده شدن نمایه برای جستجو.

سیستم‌های نمایه‌سازی استنادی خودکار شناسائی شده در پایگاه‌های استنادی

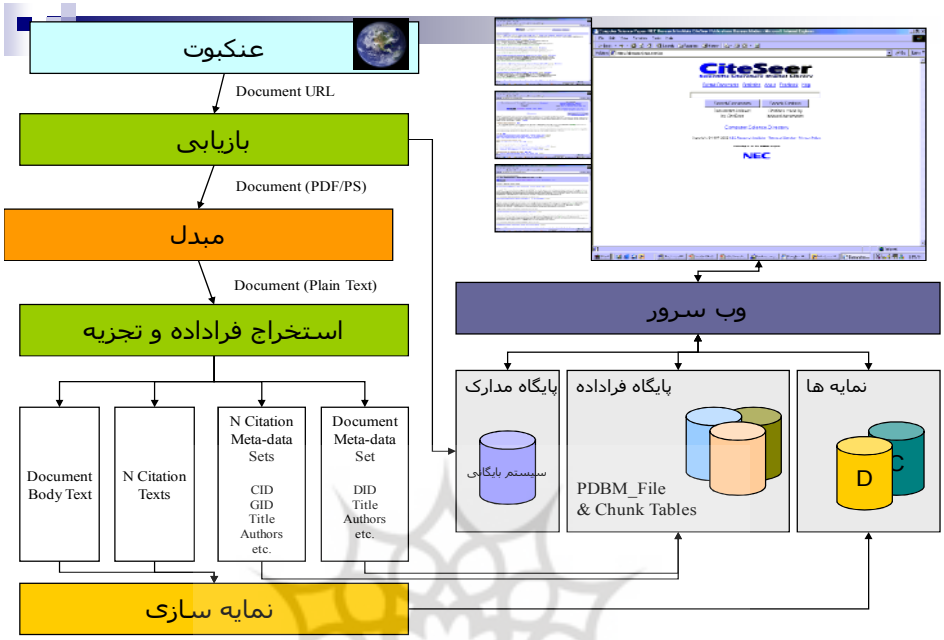
اکثر موتورهای کاوش و پایگاه‌های استنادی به دلیل ماهیت تجاری بودن، الگوریتم‌های نمایه‌سازی خود را به راحتی در اختیار کاربران قرار نمی‌دهند. ما برای شناسائی سیستم‌های نمایه‌سازی خودکار از متون مربوطه استفاده کردیم. آنچه که ازمطالعه تحقیقات و بررسی‌های دیگران بدست آمده است (گیلز،۱۹۹۸، هی،۲۰۰۲) نشان می‌دهد که تاکنون پایگاه‌های استنادی از

دو سیستم برای نمایه‌سازی استفاده می‌کنند. یکی از این سیستم‌ها ACI یا سیستم نمایه‌سازی خودکار که یکی از عمده‌ترین آن‌ها مورداستفاده Citseer است و دیگری PubSearch است که مورد استفاده قرار گرفته است.

۱. نمایه‌سازی استنادی خودگردان[1] (ACI)

بررسی‌های انجام گرفته توسط گیلز (۱۹۹۸)، لورنس[2] (۱۹۹۸) و لورنس(۱۹۹۹) نشان می‌دهد که پایگاه‌های استنادی سایت سیر و آی اس آی از نظام نمایه‌سازی استنادی خودگردان برای ذخیره و بازیابی اطلاعات استفاده می‌کنند. این نظام قادر است از متونی که الکترونیکی هستند، به صورت خودکار یک نمایهٔ استنادی ایجاد کند. موفقیت این نظام به توانایی آن در انجام درست این فعالیت‌ها بستگی دارد. همانطور که در تصویر ۳ نشان داده شده است، نمایه‌سازی استنادی خودگردان، به همین شیوه مقالات را از وب دانلود کرده و در صورت لزوم به صورت متن برمی‌گرداند. سپس برای استخراج استنادها و زمینه‌ای که این استنادها در مجموعه‌ای از مقالات تحت آن زمینه ایجاد شده‌اند، آن مقالات را تقطیع (تجزیه) می‌نماید و این اطلاعات را در یک پایگاه اطلاعاتی ذخیره می‌کند. البته این نظام کلیدواژه‌ها را بر اساس الگوریتم خاص جدا و به آن‌ها مقدار، یا به عبارتی، وزن داده و سپس در نمایه‌های مربوطه قرار می‌دهد. نمایه‌سازی خودگردان در واقع از یک ساختار شبکه‌ای برخوردار است که همه استنادها را به هم مرتبط و وزن‌دهی می‌کند. نظام مشتمل بر مقالهٔ تمام متن و نمایه‌سازی استنادی است و امکان شناسایی مقالات را از طریق جستجوی کلید واژه‌ای و یا پیوندهای استنادی فراهم می‌کند. این نظام همچنین انواع مقالات مرتبط به یک مقالهٔ خاص را از طریق اطلاعات استنادی مشترک و یا تشابه کلمات جایابی می‌کند. آنچه که در ابتدای این فرآیند اهمیت دارد شناسایی مدارک در سیستم نمایه‌سازی استنادی خودگردان است.

1. Autonomous citation index
2. Lawrence



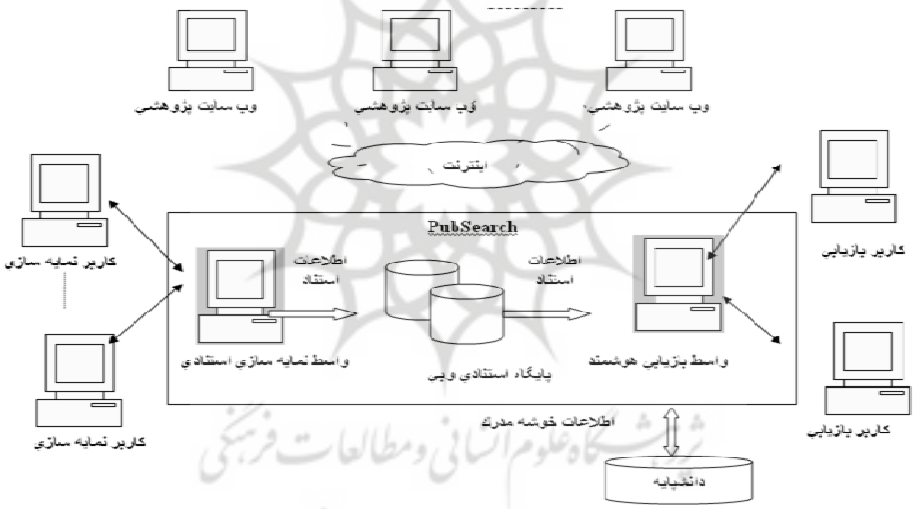
تصویر ۳. نمایش ساختار کلی نمایه‌های استنادی

۲. سیستم نمایه‌سازی Pubsearch

این سیستم شامل سه مولفه یعنی واسط نمایه‌سازی استنادی[1]، پایگاه استنادی وبی[2] و عامل بازیابی هوشمند[3] می‌باشد. دو شیوه برای ایجاد نمایه‌های استنادی بوسیله واسط نمایه‌سازی استنادی استفاده شده است. اولین شیوه شبیه سایت سیر است که از موتورهای جستجو برای مکان‌یابی کلیدواژه‌های انتشارات علمی در وب سایت‌ها استفاده می‌کند و شیوه دیگر استفاده از وب سایت‌های خاصی که توسط کاربران مورد استفاده قرار گرفته است. واسط نمایه‌سازی از هر شیوه که استفاده نمود ابتدا انتشارات علمی از وب را دانلود کرده و سپس فایل‌های پی دی اف[4] و پست اسکریپت[5] را به داده‌های متنی تبدیل و سپس بخش کتابشناختی را بوسیله جستجوی کلیدواژه‌های «Bibliography» و «References» مشخص می‌کند. در نهایت بخش کتابشناختی را تقطیع نموده و اطلاعات استنادی را استخراج و در پایگاه استنادی ذخیره می‌کند. در پایگاه استنادی دو جدول استناد و منبع ایجاد شده است که در جدول منبع، اطلاعات کتابشناختی استخراج شده توسط

1. Citation indexing database
2. Web citation database
3. Intelligent retrieval agent
4. PDF
5. Postscript

واسط نمایه‌سازی استنادی ذخیره میگردد و در جدول استناد، اطلاعات استنادی استخراج شده در آن ذخیره می‌گردد، که در بخش ساختار پایگاه‌های استنادی بطور مفصل اشاره شد. بخش سوم این سیستم یعنی واسط بازیابی هوشمند به تکنیک‌های خوشه‌بندی پایگاه‌های استنادی برای ایجاد خوشه‌های اسناد علاوه بر خوشه‌های نویسنده اشاره دارد (هی، ۲۰۰۱). این سیستم همچنین خوشه‌بندی از مدرک علاوه بر خوشه‌های هم استنادی نویسنده نویسنده را بر اساس کلیدواژه‌های پیداشده در استنادها ایجاد می‌کند. بدین ترتیب ممکن است کاربری کلیدواژه دقیق را وارد نکند، منتها بتواند با یک کلیدواژه دیگر از طریق خوشه‌بندی به نتایج دیگری که بعضا با کلیدواژه‌های وی همخوانی داشته باشد ، دسترسی پیدا کند. نمائی از سیستم در تصویر شماره ۴ نشان داده شده است.



تصویر شماره ٤: تصویرسیستم PubSearch (He,2001)

شناسایی مدارک در نظام نمایه‌سازی استنادی خودگردان

نظام نمایه‌سازی استنادی خودگردان می‌تواند با جستجوی وب، یا با برقراری پیوند مستقیم با ناشران، مقالات را شناسایی و بازیابی کند. جستجوگران در صورت آشنایی با سیستم‌های نمایه‌سازی استنادی خودگردان می‌توانند مستقیماً از پایگاه‌های مربوطه، به مقالات جدید دسترسی

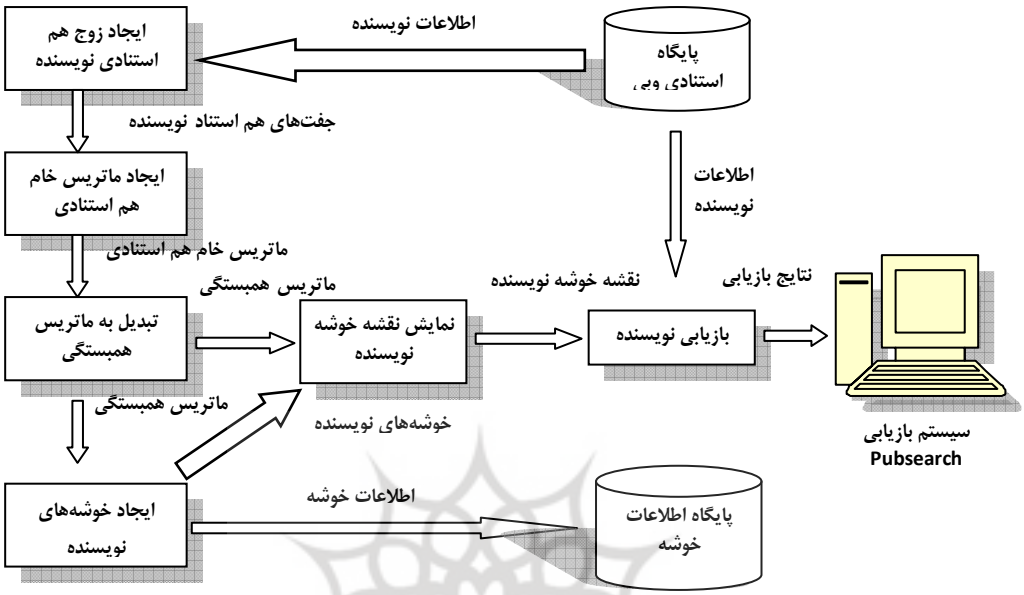
یابند و این امر با نمایه شدن نسبتاً فوری این مقالات محقق می‌شود. مجلات نوعاً برای دسترسی به مقالات پیوسته (آنلاین) طلب هزینهٔ [اشتراک] می‌کنند و بنابراین، یکی از راه‌های نمایه کردن این مقالات انجام توافقاتی با خود ناشران است. نمایه‌سازی استنادی خودگردان با تورق‌پذیر کردن آسان و سریع بستر استنادها و نیز با نمایه‌سازی گزارش‌های فنی، مقالات کنفرانس‌ها و دیگر متونی که اغلب زودتر از مقالات مجلات در دسترس قرار می‌گیرند، به ارزیابی اهمیت تک تک آثار کمک می‌کند.

فرآیند واکاوی هم استنادی نویسنده در پایگاه‌های استنادی

تحلیل هم استنادی نویسنده به عنوان یک شیوه تحلیل ساختار فکری مطالعات علمی به طور گسترده مورد استفاده قرار گرفته است. از تحلیل هم استنادی نویسنده می‌توان برای شناسائی نویسندگان هم استناد که تحقیقات موضوعی مشابه و یکسانی دارند، استفاده نمود. تحلیل هم استنادی نویسنده یعنی اینکه دو نفر که از یک مقاله استفاده کنند و در مقالاتشان به آن استناد نمایند. در گذشته از ابزارهای آماری نظیر SPSS برای این منظور استفاده می‌شد (لارسون[1] ۱۹۹۶، پری و رایس[2] ۱۹۹۸، وایت و مک کین[3] ۱۹۹۸، به نقل از: هی، ۲۰۰۲).

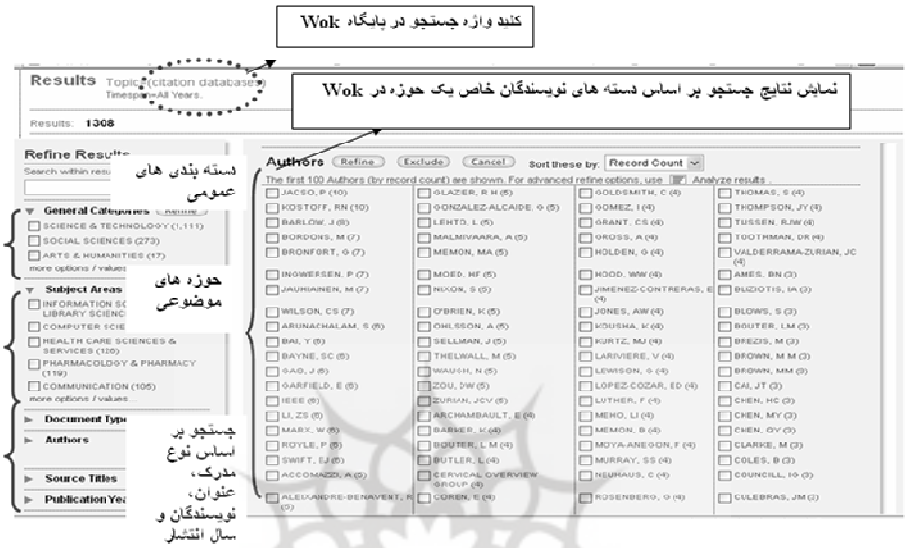
به عنوان مثال، وقتی در یکی از پایگاه‌های استنادی نظیر اسکوپوس یا برخی از پایگاه‌های دیگر جستجوی موضوعی انجام می‌دهید، در قسمت نمایش نتایج جستجو لیست شماری از نویسندگان را مشاهده می‌نمائید که به عنوان نویسندگان کلیدی موضوعات می‌باشد. پایگاه‌های استنادی از فرآیند واکاوی هم استنادی نویسنده برای نمایش و رتبه‌بندی در پایگاه‌های خود استفاده می‌کنند. این فرآیند بطور کامل بر اساس پژوهش (He, 2002) در تصویر شماره ۴ نمایش داده شده است. همچنین بر اساس همان پژوهش، الگوریتم خوشه‌بندی سلسله‌مراتبی انباشتگی[4] (AHC) برای ماتریس همبستگی و ایجاد خوشه‌های هم استنادی نویسنده در پایگاه‌های استنادی استفاده می‌شود. در مرحله آخر اطلاعات خوشه‌ای نویسنده هم استنادی برای بازیابی پرسش‌های کاربران مبتنی بر بازیابی نویسنده در سیستم Pubsearch ترکیب شده است. جستجو از کلیدواژهای موضوعی و نویسندگان در پایگاه‌های استنادی نظیر WOK[5] و SCOPUS نمونه‌ای از نمایش دسته‌های نویسندگان در یک حوزه خاص قابل مشاهده و استفاده است.

1. Larson
2. Perry & Rice
3. White & MacCmin
4. Agglomerative hierarchical clustering (AHC)
5. Web of Knowledge



تصویر شماره ۵: فرآیند واکاوی هم استنادی نویسنده (He, 2002)

در تصویر شماره ۶، نتایج جستجوی کلیدواژه «Citation Indexing» را در پایگاه «Wok »نشان می‌دهد. در این جستجو ۱۳۰۸ مدرک به عنوان نتیجه نمایش داده شده است. در سمت چپ صفحه نتایج، امکان استفاده از دسته‌های موضوعی و نویسنده وجود دارد که با کلیک بر روی«Author» امکان نمایش نویسندگان آن موضوع با تعداد استنادها ایجاد می‌شود که نشان می‌دهد نویسندگان این حوزه موضوعی که در این پایگاه مقاله دارند، چه کسانی هستند و شیوه نمایش به ترتیب الفبائی و تعداد استناد می‌باشد که در این تصویر بر اساس تعداد استناد نمایش داده شده است. برای نمونه در جستجویی که در پایگاه استنادی Wok در خصوص کلیدواژه «Citation indexing» انجام گرفت، بدین ترتیب کاربرانی که از طریق نویسندگان، نتایج را دنبال می‌کنند، می‌توانند نویسندگانی را که در آن حوزه کار کرده‌اند، شناسائی و به ردگیری هر یک از آن‌ها اقدام نمایند. نمونه‌ای از این کارکرد را در تصویر شماره ۵ می‌توان مشاهده کرد.



تصویر شماره ٦: نمایش نتیجه جستجو در پایگاه Wok

با بررسی شیوه نمایش نتایج جستجو در برخی از پایگاه‌های استنادی، مشاهده گردید که در رتبه‌بندی نتایج بر اساس تعداد استنادها، نویسنده سازمانی، نویسنده اول، حوزه‌های موضوعی و نوع مدرک در پایگاه‌های استنادی وضعیت مشابهی دارند. با این توصیف بنظر می‌رسد که پایگاه‌های استنادی از الگوریتم‌ها و روش‌های مشابهی باید استفاده کرده باشند، تا شاهد چنین رتبه‌بندی در نمایش نتایج جستجوی این پایگاه‌ها باشیم. برای این منظور در این پژوهش در مرحله بعدی ما به دنبال بررسی انواع الگوریتم‌های بکار رفته در این پایگاه‌ها بودیم، تا بدانیم که هر یک از پایگاه‌ها برای نمایش انواع رتبه‌بندی نتایج از چه الگوریتم‌هایی برای ذخیره اطلاعات استفاده کرده‌اند.

عمده الگوریتم‌های شناسائی شده مورد استفاده پایگاه‌های استنادی که از بررسی متون بدست آمده است، عبارتند از:

۱- الگوریتم‌های وزن‌دهی[۱]

برای تشریح الگوریتم‌های پایگاه‌های استنادی ،بطور خاص الگوریتم‌های وزن‌دهی از نتایج

۱. بخش عمده‌ای از الگوریتم وزن‌دهی از نتایج مطالعات گیلز استخراج گردیده است.

مطالعات گیلز که بر روی سایت سیر انجام داده است استخراج شد. استفاده از الگوریتم‌های وزن‌دهی برای این است که نتایجی به کاربران ارائه دهد که بتوانند رتبه‌بندی‌هایی را بر اساس بسامد استناد به نویسنده، ردگیری استنادها، ردگیری موضوعات مورد جستجو و ردگیری نویسندگانی که به هم استناد می‌کنند، انجام دهند. عمده مطالب این بخش حاصل نتایج مطالعات گیلز می‌باشد که به روش کتابخانه‌ای گردآوری شده است و بدنبال می‌آید:

گیلز (۱۹۹۸) ۴ روش را که نرم‌افزار نمایه‌سازی استنادی برای آزمون شناسایی موارد مختلف استنادها به مقالات واحد به کار می‌برند، مطرح می‌کند. هرچند در مقاله خود روش چهارم را بدلیل مشکلاتی که دارد یک روش ضعیف قلمداد می‌کند. روش‌های مذکور عبارتند از:

۱. خط مبنای ساده[1]: روش خط مبنای ساده‌ای که در تمامی استنادها قابل استفاده باشد، حداکثر تعداد کلمات هر استناد را که با یک استناد قبلی انطباق داشته و برحسب طول استناد کوتاه‌تر نرمال‌سازی شده باشد، به دست می‌آورد. اگر این تعداد بیشتر از یک حدّ آستانه[2] باشد، استناد جدید همچون استناد قبلی در حکم استناد به مقاله‌ای واحد تلقی می‌گردد و این استناد جدید به همراه استناد قبلی گروه‌بندی می‌شود وگرنه برای آن گروه جدیدی در نظر می‌گیریم.

۲. انطباق واژگانی[3]: الگوریتمی شبیه به الگوریتم خط مبنای ساده است که در ابتدا استنادها را برطبق طول آنها از بلندترین به کوتاهترین مرتب می‌کند.

۳. انطباق عبارتی و واژگانی[4]: نوعی الگوریتم انطباق واژگانی است که توالی دو کلمه را در هر فیلد فرعی به عنوان اصطلاحی واحد در فرآیند انطباق در نظر می‌گیرد؛ به عبارت دیگر این الگوریتم نظمی برای کلمات در نظر می‌گیرد که در الگوریتم‌های پیشین به آن توجهی نمی‌شود. حدهای آستانه‌ای جداگانه‌ای برای انطباق‌های یک کلمه‌ای و دوکلمه‌ای به کار می‌روند.

۴. روش لیکلت[5]: لیکلت نوعی شکل پیچیده فاصله ویرایشی است که در صدد انطباق وزن‌دهی شده بهینه به حروف و مولتی گراف‌ها ست (گروهی از حروف). لیکلت بین دو استناد سنجه فاصله را در نظر می‌گیرد. در این روش استنادها بر حسب طول از بلندترین به کوتاه‌ترین فاصله لیکلت برای هر فقره استنادبدست می‌آید. اگر این فاصله کمتر از حد آستانه باشد، آن استناد یکسان و واحد در نظر گرفته شده و به گروه قبلی اضافه می‌شود و گرنه، گروه جدیدی ایجاد می‌گردد.

همچنین گیلز(۱۹۹۸) و بولاکر(۱۹۹۸) مطالعه و آزمایشی بر روی سایت سیر انجام داده‌اند که نتایج آن‌ها در مورد الگوریتم‌های مختلف وزن‌دهی در زیر نشان داده شده است:

---

1. Baseline Simple
2. Threshold
3. Word Matching
4. Word and Phrase Matching
5. Likelt

۱-۱ . سنجه‌های فاصله‌ای

نتایج بررسی آن‌ها نشان می‌دهد که نرم‌افزار پایگاه‌های استنادی نظیر سایت سیر از سنجه‌های متفاوتی برای محاسبهٔ فاصله (و برعکس، تشابه) بین یک جفت مدرک (رشته‌های متنی) استفاده می‌کنند. اغلب سنجه‌های معروف فاصله بین مجموعه‌های متنی بر مدل‌های تشابه بین گروه‌های حروف موجود در متن مبتنی هستند. یک نوع از این سنجه‌ها، مبتنی بر فاصلهٔ متنی، رشته‌ای و ویرایشی است که فاصله را به عنوان میزان تفاوت بین شبکه‌ای از نمادها در نظر می‌گیرد. مثلاً فاصلهٔ لونشتاین[1] (لونشتاین، ۱۹۶۵ به نقل از: گیلز،۱۹۹۸) اولین فاصلهٔ ویرایشی بسیار معروف است که در آن تفاوت بین دو رشتهٔ متنی عبارت است از: تعداد موارد ورودی، حذفی و جایگزینی حروف مورد نیاز برای انتقال یک رشته به دیگری، نمونهٔ پیچیده و نسبتاً جدید سنجه فاصله‌ای متنی، به سنجه فاصله‌ای لیکلت معروف است.

نوع دیگر فاصلهٔ رشته‌ای متنی مبتنی است بر آمار کلماتی که در مجموعه‌هایی از مدارک، به ویژه مجموعه‌هایی که در حکم قسمتی از پیکره تعداد زیادی از مدارک هستند، مشترک می‌باشند. نوع متداول این سنجه که مبتنی بر بسامد واژگان است، به بسامد اصطلاح x بسامد معکوس مدرک[2] (TDFIF) معروف است. گاهی به جای کل کلمات فقط ریشه‌شان را در نظر می‌گیرند. روش اکتشافی ریشه‌گیری که Porter مبدع آن بود (پورتر،۱۹۸۰)، ریشهٔ واحد اشکال مختلف کلمات یکسان را استخراج می‌کند (مثلاً کلمات «رفتن»، «رفت» و «رفته» همگی از کلمهٔ «رفت» مشتق شده‌اند). بسامد هر ریشهٔ کلمهٔ s در یک مدرک d عبارت است از fds و تعداد مدارک دارای این ریشه کلمهٔ s عبارت است از ns. بیشترین بسامد واژگانی در مدرک d بسامد F_{dmax} نشان داده می‌شود. در یک چنین طرحی از TFIDF وزن کلمه w_{ds} اینگونه محاسبه می‌شود:

$$w_{ds} = \frac{(0/5 + 0/5\frac{f_{ds}}{fd_{mas}})(\log\frac{No}{ns})}{\sqrt{\sum jed((0/5 + 0/5\frac{f_{dj}}{fd_{\max}})^2(\log\frac{No}{nj})^2)}}$$

که در آن No تعداد کل مدارک است. به منظور یافتن فاصلهٔ بین دو مدرک، حاصل‌ضرب نقطه‌ای بردارهای دو کلمه برای این مدارک محاسبه می‌شود. یکی از محدودیت‌های این رویکرد، خشهٔ ذاتی[3] است: کلمات غیر مشترک ممکن است به خاطر برخی مدارک به صورت همسان و مشترک جلوه کنند. بدینگونه مدرک کاذبی دال بر مرتبط بودن آن مدارک به دست آید. یکی

1. Levenshtein
2. tern frequcncy xinverse document frequency
3. inherent noise

دیگر از محدودیت‌های این رویکرد وجود ابهام و چندگانگی در کلمات و عبارات است. مثلاً کلمهٔ «شیر» هم ماده غذایی است، هم حیوان وحشی و هم ابزاری در لوله‌کشی آب. از اینرو بسامدهای ساده کلمات که جداسازی آن‌ها نیازمند تحلیل بستر آنهاست در این مورد به کار نمی‌آیند.

سومین نوع سنجهٔ فاصلهٔ معنایی آن است که از دانش در باب مؤلفه‌ها یا ساختار مدرک استفاده می‌کند. مثلاً در مورد انتشارات پژوهشی می‌توان برای محاسبهٔ تشابه از اطلاعات استنادی بهره جست.

۲-۱. سنجه‌های تشابه

سایت سیر برای محاسبهٔ [میزان] تشابه از سه روش ذیل استفاده می‌کند(گیلز،۱۹۹۸):

بردارهای واژگانی: برای برآورد ارزش ریشهٔ کلمه در هر مدرک طرح T.FIDF را به کار گرفته است که در آن بردار همهٔ ارزش‌های ریشهٔ کلمه، «جایگاه» یک مدرک را در فضای بردار واژگانی نشان می‌دهد. سنجهٔ فاصله‌ای به کار رفته عبارت است از پیش نمایی بردار واژگانی یک مدرک روی مدرک دیگری (حاصل‌ضرب نقطه‌ای بردارها).

فاصلهٔ رشته‌ای: سایت سیر برای محاسبهٔ فاصلهٔ ویرایشی بین سرآیندهای مدارک در یک پایگاه اطلاعاتی از فاصلهٔ رشته‌ای لیکلت بهره می‌برد [1] (یانیلوس،۱۹۹۷). به زبان ساده می‌توان گفت که سرآیند هر مدرک تمام اطلاعات مدرک، قبل از چکیدهٔ آن (و در صورت نبود چکیده، قبل از مقدمهٔ آن) است. سرآیند یک مدرک حاوی عنوان، نام نویسنده و وابستگی سازمانی وی و احتمالاً محل نشر است. لیکلت بر آن است که رشته‌های فرعی موجود در رشته‌ای بزرگتر را بر هم منطبق سازد. نویسندگان مشترک، مؤسسات یا کلمات در عنوان فاصلهٔ لیکلت بین سرآیندها را کاهش می‌دهد. فرض اساسی در حین استفاده از لیکلت آن است که سرآیند مدرک در بردارندهٔ اطلاعات بسیار مهم در باب آن مدرک بوده و حضور کلمات در نظم‌هایی مشابه نشانگر مدارکی با منشاء مشابه است.

استنادها: تک کلمه‌ها (و در حدی کمتر، عبارات تکی) نمی‌توانند موضوع اصلی و یا مفاهیم مطرح در یک مدرک پژوهشی را به قدر کفایت نشان دهند. استنادهای دیگر آثار هم دستچین نویسندگان با فرض مرتبط بودن آنهاست. پس منطقی است که از اطلاعات استنادی برای داوری در مورد ارتباط مدارک استفاده شود. برخی از پایگاه‌های استنادی نظیر سایت سیر از استنادهای مشترک به منظور برآورد این نکته استفاده می‌کند که کدام مدارک موجود در پایگاه اطلاعاتی دانلود شدهٔ مدارک پژوهشی ارتباط بیشتری با مدرک دستچین شدهٔ [2] کاربر دارد. برای این منظور

1. Likelt 2. Document picked

از این سنجه، یعنی »استناد مشترک x بسامد معکوس مدرک«[1] (CCIDF) مشابه با وزن‌های واژگانی کلمه محور TFDIF (سالتون به تقل از گیلز، ۱۹۸۰) است. الگوریتم لازم برای محاسبهٔ ارتباط CCIDF تمامی مدارک موجود در پایگاه اطلاعاتی با مدرک موردنظر A و انتخاب بهترین مدرک M عبارت است از:

۱ـ تخصیص وزنی (W_i) به هر استناد i، برابر با معکوس بسامد استناد در کل پایگاه اطلاعاتی

۲ـ تعیین فهرست استنادها و وزن‌های مربوط به آن‌ها برای مدرک A و درخواست از پایگاه اطلاعاتی برای یافتن مجموعهٔ n مدرک $\{B_j\}: j=1,\ldots,n$ که حداقل یک استناد مشترک با مدرک A دارد.

۳ـ تعیین ارتباط مدرک R_j برای هر $j=1,\ldots n$ به صورت مجموع وزن‌های استنادهای مشترک با استنادهای A.

$$R_j = \sum_{i \in A \cap IEB_j}^{w_i}$$

۴ـ مرتب کردن مقادیر R_j و نشان دادن مدارک B_j با بیشترین مقادیر R_j مدارک M.

البته این الگوریتم نیز بولاکر[2] (1008) نیز به آن اشاره کرده و مورد تائید قرار داده است.

CCIDF هم مثل TFIDF فرض را بر این می‌گذارد که اگر دو مدرک، یک فقره استناد کاملاً غیر متعارف را همزمان دارا باشند، وزن این استناد باید بالاتر از وزن استناد موجود در تعداد بیشتری از متون باشد. هرچند کارکرد CCIDF را رسماً برآورد نکرده‌است ولی کارآمد می‌باشد.

تلفیقی از روش‌ها: با توجه به اینکه که بازیابی مدارک مشابه براساس استناد به صورت ذهنی بهتر از بازیابی براساس بردار واژگانی و یا شیوهٔ لیکلت است، برخی از پایگاه‌های استنادی نظیر سایت سیر که عملکرد CCIDF را به صورت کلی ارزیابی کرده و با تکنیک‌های دیگری مثل بردار واژگانی و فاصلهٔ ویرایشی که پیشتر ذکر شدند، تلفیق کرده تا بتواند از سنجهٔ فاصلهٔ تشابهی که دقیق‌تر از هر روش اعمال شده به صورت مجزاست، استفاده کند. این سنجهٔ تشابه تلفیق شده در واقع برآیند وزن‌دهی شده‌ای از سنجه‌های تشابه پیشگفته که با الگوریتم زیر انجام می‌شود:

۱ـ محاسبهٔ سنجه‌های بردار واژگانی، لیکلت و تشابه استنادی و نرمال‌سازی هر سنجه در قالب یک مقیاس صفر تا یک، که یک نمایانگر مدارک مشابه از نظر معناشناختی و صفر مبین مدارک کاملاً متفاوت (فاصلهٔ معنایی قطعی) است. سنجه‌های مشابهت نرمال شده بین دو مدرک A, B به ترتیب به صورت WV(A,B) [در بردار واژگانی] LI (A,B) [در محاسبه سنجه لیکلت] و CI (A,B) [در تشابه استنادی] نشان داده می‌شوند.

1. common citation x inverse document frequency
2. Bollacker

۲ـ برآورد مشابهت بین یک مدرک مورد نظر A و تمامی n مدارک داوطلب در مجموعهٔ مدارک $[B_j]:j=1\ldots,n$ به کمک سه سنجهٔ مذکور در بند ۱،

۳ـ تخصیص وزن‌های W_{wv}، W_{LI} و W_{CI} بر سنجه‌های مشابهت مربوطه آنها. این ارزش‌ها (مقادیر) وزنی بین صفر و یک بوده و طوری نرمال‌سازی می‌شوند که $W_{WV} + W_{LI} + W_{CI} = 1$ برابر ۱ باشد.

۴ـ یافتن یک سنجهٔ مشابهت تلفیقی S_j بین A و هرکدام از مدارک B_j به صورت حاصل جمع وزن‌دهی شده:

$$S_j = W_{wv}WV(A,B) + W_{LI}LI(A,B) + W_{CI}CI(A,B)$$

۵ـ بازیابی مدارکی با بیشترین مقادیر S_j.

لازم به ذکر است که الگوریتم‌های که گیلز اشاره کرده در اصل از تحقیقات سالتون اقتباس شده است که نه تنها نتایج گیلز بر آن‌ها صحه گذاشته است، بلکه مطالعات بولاکر (1998) هم این الگوریتم‌ها را تایید کرده است. هر جند این الگوریتم‌ها بطور خاص در سایت سیر مورد مطالعه قرار گرفته است، اما جستجو در دیگر پایگاه‌های استنادی هم وضعیت مشابه سایت سیر دارد. بدین صورت پس از جستجو در پایگاه‌های استنادی، نمایش نتایج بر اساس تعداد استنادها، تحلیل استنادی، ردگیری استنادها برای شناسائی نویسندگان فرمت یکسانی دارند. بدین منظور می‌توان نتیجه گرفت که احتمالا الگوریتم‌های مورد استفاده این پایگاه‌ها همپوشانی فراوانی دارند. چون با بررسی‌های نگارندگان الگوریتم‌های دیگری مورد شناسائی قرار نگرفت و عمده الگوریتم‌ها همان‌هائی بود که به آن‌ها اشاره داشت.

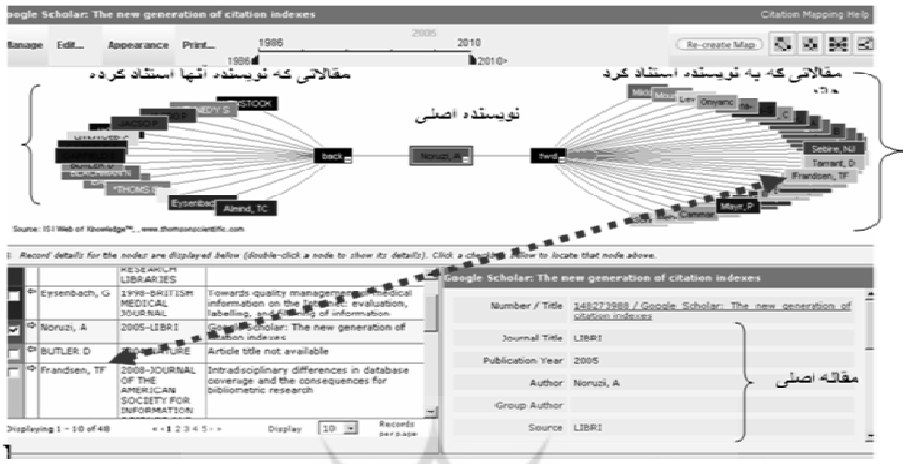
یکی از عمده نتایج بازیابی در پایگاه‌های استنادی نمایش اطلاعات بر اساس تعداد استنادها می‌باشد. یعنی مقالاتی که بیشترین فراوانی دارند یا بیشتر مورد استناد قرار گرفته‌اند، به ترتیب تعداد استناد (از زیاد به کم) مورد نمایش قرار می‌گیرند. در جستجوهایی که در اکثر پایگاه‌های استنادی انجام شد، شیوه نمایش تعداد و بسامد استنادها یکسان بود. بر این اساس می‌توان نتیجه گرفت که اکثر پایگاه‌های استنادی الگوریتم وزن‌دهی یکسانی به کار می‌گیرند. اما همانطور که پیشتر هم اشاره شد، پایگاه‌های اطلاعاتی و موتورهای جستجو به دلیل ماهیت تجاری بودن، الگوریتم‌های ذخیره اطلاعات خود را بطور آشکار در اختیار عموم قرار نمی‌دهند. در عین حال، براساس مطالعات و تحقیقات عمده الگوریتم‌های مورد استفاده این نوع پایگاه‌ها مبتنی بر الگوریتم وزن‌دهی و مدل بردار فضایی است (گیلز،۱۹۹۸). یکی از موارد نمایش نتایج بازیابی در پایگاه‌های استنادی نظیر اسکوپوس و سایت سیر و گوگل اسکالر مشخص کردن نویسندگان کلیدی هر

حوزه هست که باز هم نشان می‌دهد که این پایگاه‌ها برای تحلیل هم استنادی یک نویسنده از الگوریتم مشابهی استفاده می‌نمایند

۲- الگوریتم‌های خوشه‌بندی

پایگاه‌های استنادی برای ایجاد شبکه‌های استنادی از نویسندگان و موضوعات مرتبط رویکردهای متفاوتی اتخاذ نموده‌اند. یکی از این رویکردها تحلیل هم استنادی نویسنده است، که برای تحلیل ساختار فکری مطالعات علمی مورد استفاده قرار گرفته است. برای این منظور از تحلیل هم استنادی نویسنده برای خوشه‌بندی نویسندگان استفاده کرده است. الگوریتم مورد استفاده خوشه‌بندی سلسله‌مراتبی انباشتگی(AHC) است که به عنوان تکنیک واکاوی برای خوشه‌بندی نویسنده و نمایش و سنجش چند بعدی برای نمایش نقشه خوشه نویسنده استفاده شده است (He, 2001, Bollaker, Lawrence & Giles, 1998).

یکی از الگوریتم‌هایی که امروزه برای بازیابی در موتورهای جستجو و برخی پایگاه‌های اطلاعاتی استفاده میشود، الگوریتم خوشه‌بندی است. نتایج جستجو در آن پایگاه‌ها و موتورها نشان می‌دهد که هنگام بازیابی نتایج، خوشه‌های نویسندگان و مدارک قابل مشاهده است. بررسی‌ها نشان می‌دهد پایگاه‌های استنادی نیز از الگوریتم‌های خوشه‌بندی استفاده می‌کنند. (He, 2002). هدف نهایی خوشه‌بندی این است که داده‌های موجود را به چند گروه تقسیم کنند و در این تقسیم‌بندی داده‌های گروه‌های مختلف باید حداکثر تفاوت ممکن را نسبت به هم داشته باشند و داده‌های موجود در یک گروه باید بسیار به هم شبیه باشند. البته کیفیت نتایج خوشه‌بندی به روش اندارگیری شباهت و توانایی و قدرت الگوریتم در کشف الگوهای مخفی میان داده‌ها بستگی دارد. همچنین از این الگوریتم برای نمایش نقشه هم استنادی نویسنده استفاده گردیده است که در تصویر۶ شماره نقشه استنادی نویسندگانی که به گیلز استنادکرده‌اند نمایش داده شده است. به عنوان نمونه در تصویر شماره ۷ وقتی در پایگاه آی اس آی در مورد «Noruzi» جستجو انجام شد، مقاله «Google Scholar: The new generation of citation indexes» به عنوان پراستنادترین مقاله در ربته اول صفحه نمایش ظاهر گردید. برای ردیابی و نمایش نقشه استنادی ا زگزینه 'Citation Map" استفاده گردید و نقشه مقالتی که به این عنوان استناد کرده‌اند و مقالاتی که خود نویسنده به آن‌ها استناد کرده بود نمایش داده شد. همانطوریکه هر(۲۰۰۱) هم اشاره کرد، پایگاه‌های استنادی از الگوریتم «AHC» بریا نمایش نقشه‌های استنادی نویسنده استفاده می‌کنند.



تصویر شماره۷: نمایش نقشه استناد یه یک نمونه مقاله بازیابی شده در پایگاه Wok

الگوریتم خوشه‌بندی بطور کلی به دو دسته تقسیم می گردد: الگوریتم‌های سلسله‌مراتبی و غیر سلسله‌مراتبی. الگوریتم‌های خوشه‌بندی سلسه‌مراتبی مستلزم یک فرآیند ساختاری شبیه درخت است. از میان انواع الگوریتم‌های خوشه‌بندی الگوریتم خوشه‌بندی سلسله‌مراتبی انباشتگی در پایگاه‌های استنادی بیشتر مورد استفاده قرار گرفته است (He, 2002)

بحث و نتیجه‌گیری

انقلاب ناشی از ظهور وب و تأثیر آن بر ذخیره و دسترسی به اطلاعات صرفاً به جهت دسترس‌پذیری به اطلاعات نیست بلکه به جهت کارآمدی در دسترسی به اطلاعات است. در این راستا هر چند بسیاری از پایگاه‌های اطلاعاتی متن کامل در دسترس هستند، ولی آنچه که امروزه به طور خاص مورد توجه پژوهشگران و دانشجویان قرار گرفته پایگاه‌های استنادی است که به لحاظ دسترس پذیرساختن شبکه‌ای از اطلاعات مرتبط توجه بسیاری را به خود جلب نموده است. علت اساسی آن است که این نمایه‌ها ویژگی‌های مهمی دارند که دست به دست هم داده و آن‌ها را برای یک چنین کاربردی بسیار مناسب و منحصر به فرد ساخته‌اند.

بسیاری از پایگاه‌های استنادی که ماهیت مشابهی دارند، از الگوریتم‌های مشابهی استفاده می‌کنند. آنچه که در این مقاله بطور خاص مورد توجه قرار گرفت تحلیل ساختار الگوریتم‌های ذخیره و بازیابی پایگاه‌های استنادی بود که ماهیت تجاری داشتند، نظیر: سایت سیر، اسکوپوس. هرچند، در این پژوهش برخی از پایگاه‌های استنادی نظیر گوگل دانشوران به لحاظ تفاوت در ساختار و ماهیت مورد بررسی قرار نگرفتند. بررسی‌های اولیه نشان داد که اصولا چنین پایگاه‌های

بدلیل ماهیت تجاری الگوریتم‌های ذخیره‌سازی خود را به راحتی در اختیار دیگران قرار نمی‌دهند. اما با بررسی و کنکاش در اینترنت و جستجوی عملی در این پایگاه‌ها و مقایسه نتایج با یکدیگر به این نتیجه رسیدیم که بسیاری از این پایگاه‌ها از دو سیستم نرم‌افزاری برای ذخیره‌سازی و نمایه‌سازی استنادها استفاده می‌کنند. یکی از نرم‌افزارهای که برخی از این پایگاه‌ها نظیر سایت سیراستفاده می‌کنند نرم‌افزار نمایه‌سازی استنادی خودگردان است که بسیاری از این پایگاه‌های برای سازماندهی اطلاعاتی کتابشناختی متون علمی از آن استفاده می‌کنند. نمایه‌سازی استنادی خودکار روند نشر اطلاعات علمی الکترونیکی را تسریع می‌کند. همچنین به واسطه استفاده از آن نرم‌افزار، یک شبکهٔ پیوندی از اطلاعات کتابشناختی متون علمی با دسترس‌پذیری وسیع نویسندگان مقالات ایجاد می‌کند. سایت سیر به طور خودکار در وب مقالات را شناسائی و استنادهای آن‌ها را تقطیع می‌کند و سپس نمایه‌سازی می‌کند. همچنین از موتور جستجوی آلتا ویستاو هات بات و اکسایت استفاده می‌نماید. برخی از پایگاه‌های استنادی دیگر سیستم PubSearch (Hui,2002) مورد استفاده قرار داده‌اند که عمده تفاوت این سیستم با ACI استفاده از الگوریتم خوشه‌بندی نویسنده است که این سیستم را از سیستم‌های دیگر متمایز می‌سازد.

نتایج دیگر این بررسی نشان داد که امروزه پایگاه‌های استنادی دو کارکرد عمده دارند که توجه محققان و نویسندگان را به خود جلب نموده است. یکی ردگیری استنادها برای شناسائی نویسندگان هسته در یک موضوع است که پایگاه‌های استنادی هم از تکنیک تحلیل هم-استنادی نویسنده برای نمایش دسته‌بندی نویسندگان استفاده می‌کنند که برای اجرای این تکنیک، الگوریتم خوشه‌بندی سلسله‌مراتبی انباشتگی نویسنده را به کار می‌برند. دوم شناسائی مقالات پر استناد است که از طریق نمایش نتایج جستجو در پایگاه‌های بر اساس تعداد استنادها دنبال می‌شود. در این راستا، پایگاه‌های استنادی عمدتا از مدل‌های وزن‌دهی و مدل‌بردار فضایی برای ذخیره‌سازی استنادها استفاده می‌نمایند. در ایران نیز پایگاه‌های استنادی شروع به فعالیت نموده‌اند که می‌توانند برای بازیابی بهتر و روان‌تر اطلاعات از ترکیبی از این الگوریتم‌ها استفاده نموده و زمینه دسترسی کاربران را در بهره‌گیری از اهداف کاربردی آن‌ها یعنی شناسائی نویسندگان پراستناد، مقالات پراستناد، حوزه‌های فعال و سایر کارکردها فراهم سازند. هر چند به نظر می‌رسد که سایر پایگاه‌های استنادی به ویژه در حوزه‌های تخصصی قابل راه‌اندازی است.

برای بازیابی کارآمدتر اطلاعات، شناسائی نویسندگان هسته و خوشه‌های موضوعی استفاده از الگوریتم‌ها و سیستم‌ها، کتابداران می‌توانند به متخصصان نرم‌افزاری در طراحی و راه‌اندازی این پایگاه‌ها کمک کنند. همچنین پیشنهاد می‌گردد تا پژوهشگران و دانشجویان تحصیلات تکمیلی به بررسی بیشتر و دقیق‌تر این موضوع پرداخته و امکان سنجی ایجاد دیگر پایگاه‌های استنادی فارسی و ارزیابی پایگاه استنادی جهان اسلام مورد توجه قرار دهند.

فهرست منابع

پائو، میراندا لی. ۱۳۷۸. *مفاهیم بازیابی اطلاعات*. ترجمه رحمت‌الله فتاحی و اسدالله آزاد. مشهد، دانشگاه فردوسی، موسسه چاپ و انتشارات.

حسن زاده، محمد، و عبدالرضا نوروزی چاکلی. ۱۳۸۷. *نمایه‌سازی استنادی و روابط علمی*. رهیافت، شماره ۴۳، پاییز و زمستان.

حری، عباس، و اعظم شاهبداغی. ۱۳۸۵. *شیوه‌های استناد در نگارش‌های علمی*. تهران :انتشارات دانشگاه تهران.

مهراد، جعفر، محمد حسین دیانی، رحمت الله فتاحی، محمدرضا داور پناه، علی گزنی، و رویا مقصودی. ۱۳۸۶. **نمایه استنادی علوم ایران**. شیراز: مرکز منطقه‌ای اطلاع رسانی علوم و فناوری.

Bollacker, K, Lawrence, S., & Giles, C. 2000. Discovering relevant scientific literature on the web. IEEE Intelligent Systems, 15(2) , 42–47.

Bollacker, K., Lawrence, S., & Giles, C. 1998. Citeseer: An autonomous Web Agent for Automatic Retrieval and Identification of Interesting publication. 2nd International ACM conference on Autonomous Agent.pp. 116-123,ACM Press,May.

Garfield, E. 1979. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Wiley, New York.

Garfield, E 1994. The concept of citation indexing: A unique and innovative tool for navigating the research literature. Current Contents, January 3.

Garfield, E 1994. Where was this paper cited? Current Contents, January 31, 1994.

Garfield, E. 1955. Citation indexes for sciences: A new dimension in documentation through association of ideas. Science 122 (3159) : 108-111.

Garfield, E. 1979. *Citation indexing: Its theory and applications in science, technology and the humanities*. New York: Wiley Inter science.

Garfield, E. 1979. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*, John Wiley & Sons, New York.

Giles, C.L. 1998, "Cite Seer: An Automatic Citation Indexing System," K. Bollacker, and S. Lawrence.Digital Libraries: Third ACM Conf. on Digital Libraries, ACM Press, New York, 1998, pp. 89-98

He, Y. 2001. Mining a web citation database for the retrieval of scientific publications over the WWW. M. Eng. Thesis, School of Computer Engineering, Nanyang Technological University, Singapore.

He, Y. 2001b.PubSearch: a web citation – based retrieval system. Siu Cheung Hui. Library Hi Tech. volume 3.pp.274-285

He, Yulan. Siu Cheung Hui(2002). Mining a web Citation Database for author C0- Citation analysis. Information processing and Mangemwnt 38.491-508

Lawrence, Steve.1998. Digital Libraries and Autonomous Citation Indexing. C. Lee Giles, Kurt Bollacker. IEE computer Available at:

Lawrence, Steve.1998. Indexing and Retrieval of Scietific Literature. Kurt Bollacker ,C. Lee Giles. ACM

Nourozi, Alireza. 2005. Google scholar: The new generation of citation indexes. Libri, vol55.pp-170-180

Salton, Gerard and C.S. Yang 1973. On the specification of term values in automatic indexing. Journal of Documentation, 29:351–372.

Salton, Gerard and Chris Buckley 1987. Term weighting approaches in automatic text retrieval. Tech Report 881, Department of Computer Science, Cornell University, 1987.

Porter, M. F 1980. An algorithm for suffix stripping. Program, 14:130–137, 3.

Salton Gerard. 1971. Automatic indexing using bibliographic citations. Journal of Documentation, 27:98–110.

Yianilos, 1997.*The LikeIt Intelligent String Comparison Facility*, Tech. Report 97-93, NEC Research Institute.

Web-based citation databases: An Analysis of the Structures and Algorithms for Information Storage and Retrieval

Abdoulrasoul Khosravi*
Faculty member of Boushehr University

Seyyed Rahmatullah Fattahi
Professor in Ferdowsi University of Mashhad

Information Sciences & Technology

Iranian Research Institute
For Science and Technology
ISSN 1735-5206
eISSN 2008-5583
Indexed in LISA, SCOPUS & ISC
Vol.26 | No.2 | pp: 199-222
Winter 2011

Abstract: Developing citation databases as a new approach to information storage and retrieval is one of the main activities of information professionals in recent years. The aim of such databases is to satisfy researchers and students' information needs on the Web. The various capacities of these databases provide researchers with different facilities both in information searching and retrieval as well as in scientometrics. Performed using case study (direct observation of these databases) and review of the related literature, this study aims at investigating structures as well as information search and retrieval facilities of Google Scholars and Scopus. This necessitates the analysis of the indexing structure, algorithms of the databases regarding weighting, indexing, connecting citations and author co-citations. Findings revealed that these databases had networked structures and used two citation indexing systems, mainly an automated indexing system. Their algorithm system is based on weighting and spatial vector models and cumulative hierarchy clustering which are discussed in detail. The algorithm allows searchers to trace a network of cited and citing sources.

Keywords: citation indexing; web-based citation databases; Information retrieval; Autonomous citation index

* Corresponding Author: khosravi2422@gmail.com