آزمایشگاه فناوری وب

پایان‌نامه کارشناسی ارشد

ارائه یک سیستم پیشنهاد استناد مبتنی بر داده‌های پیوندی

فتانه زرین کلام

استاد راهنما: دکتر محسن کاهانی

بهمن ماه ۱۳۹۰


بسم الله الرحمن الرحيم

تقدیر و سپاس

بدینوسیله بر خود لازم می‌دانم که از زحمات بی‌دریغ استاد گرامی، جناب آقای دکتر کاهانی که راهنمایی‌های ارزشمند ایشان در تمام مراحل انجام این پایان‌نامه راه‌گشای من بوده است تشکر نمایم.

همچنین از تلاش‌های بی‌وقفه و پشتیبانی‌های پدر و مادر عزیزم و همسر مهربانم کمال سپاس‌گزاری را دارم.

چکیده

حجم فراوان و روبه رشد مقالات منتشر شده بر روی وب، فرآیند تصمیم‌گیری و انتخاب مقالات مرتبط با یک زمینه تحقیقاتی را برای پژوهشگران دشوار کرده است. روش رایجی که اغلب پژوهشگران برای جستجوی اسناد مرتبط با یک زمینه تحقیقاتی استفاده می‌کنند، یافتن کلمات کلیدی و استفاده از موتورهای جستجو می‌باشد. با توجه به این‌که پیدا کردن لیست کلمات کلیدی که دربرگیرنده تمام مقالات یک زمینه باشند کاری دشوار است، با استفاده از این روش، نمی‌توان به خوبی تمامی مقالات مرتبط را پیدا نمود.

یک سیستم پیشنهاد استناد، با دریافت متن ورودی، مقالاتی که باید آن متن به آن‌ها استناد کند را پیشنهاد می‌کند، و بدین ترتیب می‌تواند در یافتن مقالات مرتبط با یک موضوع به پژوهشگر کمک کند.

در حال حاضر سیستم‌های پیشنهاد استناد موجود محدود به پیشنهاد از یک منبع داده محلی می‌باشند، این محدودیت، از آنجاییکه در زمینه کتاب‌شناسی هیچ منبع‌داده‌ای حاوی اطلاعات کامل درباره تمام مقالات نمی-باشد، باعث کاهش کیفیت پیشنهادها می‌شود.

در این پایان‌نامه یک سیستم جدید برای پیشنهاد استناد ارائه شده است که در لایه داده خود از داده‌های پیوندی استفاده می‌کند و الگوریتم پیشنهاد آن مبتنی بر ترکیب شباهت رابطه‌ای و شباهت متنی می‌باشد. ارزیابی-های انجام شده نشان می‌دهد که استفاده از داده‌های پیوندی بعنوان لایه داده بدلیل مزایای آن از جمله انتشار داده‌ها با یک قالب استاندارد و برقراری پیوند بین منابع داده مختلف باعث کاهش پیچیدگی جمع‌آوری داده و غنی شدن لایه داده به دلیل استفاده از چندین منبع می‌شود.

همچنین با توجه به آزمایش‌های انجام شده، معیار شباهت رابطه‌ای پیشنهادی، در تشخیص شباهت مقالات موفق است و استفاده از آن در کنار شباهت متنی می‌تواند ضعف استفاده تنها از شباهت متنی را در پیدا کردن مقالات مرتبط کاهش دهد و در نتیجه سبب بهبود کیفیت سیستم پیشنهاد استناد شود.

کلمات کلیدی: بازیابی اطلاعات، پیشنهاد، استناد، تحلیل استناد، شباهت رابطه‌ای، داده‌های پیوندی

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فهرست علائم و اختصارات

SemCiR:	Semantic Citation Recommendation
NDCG:	Normalized Discounted cumulative gain
RDF:	Resource Description Framework.
SPARQL:	SPARQL Protocol and RDF Query Language.
ItIF:	In-text Impact Factor
DSI:	Distance Similarity Index
N3:	Notation3
LOD:	Linked Open Data
HTTP:	Hypertext Transfer Protocol
URI:	Uniform Resource Identifier
W3C:	World Wide Web Consortium

فصل ۱– مقدمه

در این فصل، ابتدا تعریف مساله آورده شده است که در آن علت وجود سیستم‌های پیشنهاد استناد و چالش‌های آن توضیح داده می‌شود. سپس در بخش ۱-۲، راه‌حل پیشنهادی این پایان‌نامه برای برطرف کردن چالش‌های موجود به اختصار شرح داده می‌شود. در ادامه نیز نوآوری‌های روش‌پیشنهادی و ساختار بقیه مطالب این پایان‌نامه ذکر می‌شود.

۱–۱ تعریف مساله

هر پژوهشگری قبل از شروع کاری جدید در زمینه مورد علاقه خود باید از کارهای انجام شده درباره آن موضوع آگاهی کافی داشته باشد. نداشتن دانش کافی نسبت به کارهای گذشته باعث به نتیجه نرسیدن تلاش‌های یک پژوهشگر و یا انجام کاری تکراری می‌شود. با توجه به اهمیت زیاد این موضوع، و نیز رشد روزافزون علم و افزایش تعداد اسناد علمی منتشر شده، نیاز به سیستمی که پژوهشگران را در این امر یاری کند کاملا محسوس است (Bethard et al, 2010; Tang and Zhang, 2009).

امروزه، اغلب پژوهشگران برای یافتن کارهای مرتبط با یک موضوع، از روش‌های رایج، مثل جستجو در گوگل استفاده می‌کنند. ورودی این روش‌ها، اغلب تعدادی کلمه کلیدی، و خروجی آن‌ها اسنادی است که شامل این کلمات کلیدی هستند. بدین ترتیب اگر پژوهشگری در ارتباط با موضوع مورد علاقه خود متنی داشته باشد و کارهای مرتبط با آن را بخواهد، ابتدا باید کلمات کلیدی موجود در متن را استخراج کند. استخراج این کلمات برای پژوهشگری که به تازگی به تحقیق در یک زمینه پرداخته است، کار آسانی نیست. همچنین مشکل دیگر این است که پژوهشگر باید زمان نسبتا زیادی را برای انتخاب گزینه‌های بهتر از بین خروجی‌های موتور جستجو صرف کند.

ضمنا از آنجاییکه هدف بدست آوردن کارهای مرتبط است و نه صرفا کارهایی که شباهت متنی زیادی با متن ورودی دارند، ممکن است تعدادی از کارهای مرتبط در بین جواب‌های جستجو ظاهر نشوند (Henzinger et al., 2003). علت این امر را می‌توان علاوه بر مشکلات رایج در پردازش متن مثل ابهام‌های زبانی و کلمات هم-خانواده (Baeza-Yates, 2004)، موارد اشاره شده در زیر دانست:

۱- ممکن است موضوع دو سند دقیقا یکسان باشد، اما از آنجایی که توسط دو نویسنده مختلف نوشته شده‌اند و کلمات مورد استفاده این دو نویسنده متفاوت است، شباهت متنی دو سند کم بوده و در نتیجه استفاده صرف از شباهت متنی باعث می‌شود مرتبط شناخته نشوند.

۲- دو سند مرتبط لزوما دو سند با شباهت متنی زیاد نمی‌باشند، مثلا سندی که درباره زمانبندی پردازش‌ها به کمک الگوریتم ژنتیک بحث می‌کند، ممکن است به سندی که ایده کلی الگوریتم ژنتیک را توضیح داده است، شباهت متنی کمی داشته باشد. اما این دو سند کاملا به هم مرتبط می‌باشند، چرا که سند دوم، پایه علمی تکنیک مورد استفاده در سند اول را توضیح می‌دهد. در اینجا نیز استفاده صرف از شباهت متنی قادر به تشخیص ارتباط این اسناد نیست.

راه‌حل برطرف کردن چنین مشکلاتی وجود یک سیستم پیشنهاد استناد است که ورودی آن یک قطعه متن و خروجی آن مقالاتی است که باید در آن متن مورد استناد قرار بگیرند، به عبارتی مقالات مرتبط با آن متن است (Strohman et al., 2007; Bethard et al., 2010; Tang and Zhang, 2009).

لایه داده در سیستم‌های پیشنهاد استناد، منابع داده‌ای می‌باشند که اطلاعات مقالات علمی را منتشر می‌کنند، جمع‌آوری این داده‌ها بعنوان لایه داده یکی از مهم‌ترین چالش‌ها در این سیستم‌ها می‌باشد. چرا که منابع داده مختلف اطلاعات خود را به قالب‌های مختلفی منتشر می‌کنند، در نتیجه برای جمع‌آوری اطلاعات هر یک از این منابع داده باید روال مجزایی اجرا شود.

چالش دیگری که در لایه داده سیستم‌های موجود وجود دارد، این است که از آنجاییکه این منابع داده کاملا جدا از هم می‌باشند و داده‌های آن‌ها به صورت محلی و خصوصی می‌باشد، اگر مثلا مقاله‌ای در یک منبع داده، و مراجع آن در منبع داده دیگری منتشر شود، امکان بازیابی اطلاعات مراجع آن مقاله بسادگی وجود ندارد.

معمولا اطلاعات مقالات مختلف توسط منابع‌داده متفاوتی منتشر می‌شوند، مثلا اطلاعات یک مقاله ممکن است توسط *IEEE*[1] منتشر شود و در *Citeseer*[2] و *DBLP*[3] وجود نداشته باشد، و یا برعکس. حتی اگر اطلاعات یک

[1] http://www.ieee.org/index.html
[2] http://citeseer.ist.psu.edu/index
[3] http://dblp.uni-trier.de/

مقاله توسط دو منبع داده منتشر شود، این منابع ممکن است اطلاعات متفاوتی درباره آن مقاله منتشر کنند، برای مثال کلمات کلیدی یک مقاله در *IEEE* منتشر می‌شود ولی در *DBLP* منتشر نمی‌شود. در نتیجه فرض استفاده تنها از یک منبع داده در زمینه کتاب‌شناسی[1] برای برطرف کردن نیازهای اطلاعاتی سیستم‌های پیشنهاد استناد، یک فرض منطقی نمی‌باشد و باعث می‌شود کیفیت پیشنهادها از نظر کامل بودن نتایج کاهش یابد.

۱–۲ راه‌حل

در این پایان‌نامه برای کاهش ضعف ناشی از استفاده تنها از شباهت متنی در پیشنهاد کارهای مرتبط، دو راهکار ارائه شده است.

۱- استفاده از ویژگی‌های رابطه‌ای مثل لیست مراجع و نویسندگان، در کنار ویژگی‌های متنی

۲- استفاده از متن استناد بعنوان یک ویژگی متنی علاوه بر بقیه ویژگی‌های متنی مثل عنوان و چکیده

برای استفاده از این دو راهکار، ابتدا یک معیار جدید برای محاسبه شباهت رابطه‌ای دو مقاله ارائه شده است. سپس نشان داده می‌شود که استفاده از این معیار در کنار شباهت متنی نقش موثری در تشخیص مقالات مرتبط دارد، و با افزودن متن استناد نیز می‌توان کیفیت سیستم‌های پیشنهاد استناد را افزایش داد.

همچنین در این پایان‌نامه برای تولید سیستمی که به راحتی بر روی هر منبع داده‌ای کار کند و قادر باشد از اطلاعات چندین منبع داده استفاده کند، استفاده از وب داده بجای وب اسناد در لایه داده پیشنهاد می‌شود. استفاده از داده‌های پیوندی بدلیل قالب یکسان آن‌ها در بازنمایی داده‌ها مشکل ناهمگونی داده‌های منابع مختلف را برطرف می‌کند، همچنین بدلیل داشتن پیوندهای معنادار بین داده‌های منابع مختلف امکان استفاده از منابع داده مختلف را ساده می‌کند.

۱–۱ نوآوری‌های سیستم پیشنهادی

نوآوری‌های سیستم پیشنهادی را می‌توان در موارد زیر خلاصه کرد:

[1] Bibliography

۱- ارائه الگوریتمی مبتنی بر ترکیب ویژگی‌های متنی و رابطه‌ای برای پیشنهاد استناد

۲- استفاده از متن استناد بعنوان یک ویژگی متنی و نشان دادن تاثیر آن در سیستم‌های پیشنهاد استناد

۳- ارائه یک معیار شباهت رابطه‌ای برای بدست آوردن شباهت دو مقاله

۴- استفاده از الگوریتم ژنتیک برای وزن‌دهی خودکار به ویژگی‌های رابطه‌ای و نشان دادن میزان تاثیر هر یک از آن‌ها در سیستم پیشنهاد استناد

۵- استفاده از داده‌های پیوندی در لایه داده سیستم پیشنهاد استناد

۶- ارائه یک الگوریتم غنی‌سازی داده‌ها با استفاده از چند منبع داده پیوندی

۱-۲ ساختار پایان‌نامه

سازماندهی مطالب این پایان‌نامه در ادامه توضیح داده شده است. در بخش بعدی با توجه به این که کار انجام شده در این پایان‌نامه به سه زمینه "تحلیل استناد"، "سیستم‌های پیشنهاد مقاله" و "سیستم‌های پیشنهاد دهنده مبتنی بر داده‌های پیوندی" مرتبط است، کارهای مرتبط انجام شده در هر زمینه شرح داده شده است. سپس در فصل سوم، راه‌حل پیشنهادی برای تولید یک سیستم پیشنهاد استناد توضیح داده شده است، در فصل چهارم، جزئیات پیاده‌سازی سیستم پیشنهادی، نحوه ارزیابی و نتایج آن آورده شده است. بخش آخر نیز به بیان نتیجه‌گیری و کارهای آینده اختصاص دارد.

فصل ۲– مروری بر کارهای گذشته

سیستم پیشنهادی این پایان‌نامه، یک سیستم پیشنهاد استناد است که باگرفتن متن ورودی مقالات مرتبط با آن متن را پیشنهاد می‌دهد، در لایه داده این سیستم از داده‌های پیوندی استفاده شده است و الگوریتم آن مبتنی بر روش‌های موجود در زمینه تحلیل استناد می‌باشد. در نتیجه در این فصل کارهای مرتبط موجود در هر یک از زمینه‌های مرتبط با سیستم پیشنهادی یعنی **"تحلیل استناد**[1]**"**، **"سیستم‌های پیشنهاد مقاله"** و **"سیستم-های پیشنهاد دهنده مبتنی بر داده‌های پیوندی"**، مورد بررسی قرارگرفته است.

۲–۱ تحلیل استناد

یک استناد در واقع تاییدی است که یک سند از دیگری دریافت می‌کند (Smith, 1981). در حالت کلی، تحلیل استناد به مطالعه ارتباطات بین یک سند و سندی که به آن استناد می‌کند می‌پردازد، در این حوزه از علم مفاهیم مختلفی استفاده می‌شود که در ادامه توضیح داده شده‌اند:

- *گراف استناد*[2]: گرافی که گره‌های آن نشان‌دهنده اسناد و یال‌های آن‌ها نشان‌دهنده استناد بین آن-ها می‌باشد.
- *تعداد استناد یک سند:* تعداد اسنادی که به آن سند استناد می‌کنند و اغلب بعنوان معیاری برای سنجش کیفیت یک سند استفاده می‌شود، به‌طوری‌که اسنادی که تعداد استناد بیشتری دارند از آنجایی که مورد تصدیق افراد بیشتری قرار گرفته‌اند از کیفیت بالاتری برخوردارند. در (Bornaman,2008) بررسی کاملی از تفسیرهای مختلف معیار تعداد استناد، شرح داده شده است.
- معیار تاثیر یک مجله که توسط گارفیلد در (Garfield, 1972) به‌عنوان معیاری برای ارزیابی کیفیت یک مجله معرفی شده است، برابر است با تعداد دفعاتی که در یک بازه مشخص به طور میانگین مقالات موجود در آن مجله مورد استناد قرار می‌گیرند.

[1] Citation analysis
[2] Citation graph

- زوج‌های کتابشناختی[1]: دو سند که حداقل یک مرجع مشترک داشته باشند زوج شناخته می‌شوند (Kessler, 1963). زوج‌های کتاب‌شناختی مبتنی بر این ایده هستند که سندهایی که در موضوع دارای شباهت هستند مراجع مشترک دارند. این معیار اغلب در بدست آوردن میزان شباهت بین دو سند استفاده می‌شوند.
- استناد مشترک[2]: دو سند دارای استناد مشترک می‌باشند اگر سندی وجود داشته باشد که به هر دو این اسناد استناد کند (Small, 1973). مفهوم استناد مشترک این است که سندهایی که دارای شباهت هستند به احتمال زیاد توسط یک سند مشترک مورد استناد قرار می‌گیرند. از این معیار نیز جهت بدست آوردن میزان شباهت نین دو سند استفاده می‌شود.
- متن استناد: قسمتی از متن یک سند است که در آن به سند دیگری استناد شده است. در تحلیل متن استناد[3] ، کلمات صریح و با محتوایی که نویسنده استناد دهنده برای توصیف کار استناد شده استفاده می‌کند هدف مطالعه قرار می‌گیرد (Small, 1982). استناد در مقاله استناد دهنده خلاصه-ای از سند استناد شده و یا حداقل چیزی است که نویسنده استناد دهنده فکر می‌کند در در کار استناد شده مهم می‌باشد. به‌عبارت دیگر، متن استناد شامل کلمات اصلی کار استناد شده است. در نتیجه برای بدست آوردن میزان شباهت بین یک متن و یک سند می‌توان میزان شباهت آن را با متن‌های استناد آن بدست آورد.

اسمیت در (Smith, 1981) کاربردهای زیادی برای تحلیل استناد نام برده است که یکی از آن‌ها که مرتبط با سیستم پیشنهادی این پایان‌نامه نیز می‌باشد، کاربرد تحلیل استناد در بازیابی اطلاعات است. برای مثال برای بهبود عملکرد الگوریتم‌های دسته‌بندی اسناد در (Couto et al., 2010) از دو معیار استناد مشترک و زوج‌های کتابشناختی استفاده شده است، استروهمن و همکارانش نیز در (Strohman et al, 2007) از معیار استناد مشترک در کنار فاکتوهای دیگر برای پیشنهاد اسناد مرتبط با یک متن استفاده کردند.

[1] Bibliographic coupling
[2] Co-citation
[3] Citation context analysis

با توجه به این ایده که متن استناد شامل کلمات اصلی متن استناد شده است، مطالعات زیادی نیز به طور خاص نقش موثر متن استناد را در بازیابی اطلاعات نشان داده‌اند، برای مثال، اُ کرنل در (O'Connor, 1982) به این موضوع اشاره می‌کند که عبارات اسمی[1] موجود در متن استنادهای یک سند، برای بهبود در بازیابی باید اندیس‌گذاری شوند و به نمایش آن سند اضافه شوند. الجابر و همکارانش در (Aljaber et al., 2010) نقش موثر استفاده از متن استناد را در بالا بردن کارایی خوشه‌بندی اسناد نشان دادند.

ریتچی و همکارانش در (Ritchie, 2008; Ritchie et al., 2008) برای بهبود کارایی بازیابی اسناد، علاوه بر متن آن سند، متن‌های استنادی که در سندهای دیگر به آن سند استناد شده بود را نیز اندیس‌گذاری کردند و نشان‌دادند که کارایی بهبود می‌یابد، همچنین آزمایشاتی جهت بدست آوردن طول متن استناد نیز انجام دادند و نتیجه گرفتند که طول ثابت برای انتخاب متن استناد موثر ترین روش است.

۲-۲ پیشنهاد مقاله

سیستم‌های پیشنهاد دهنده مقالات به کاربران، با توجه به علاقه کاربر و زمینه کاری او از بین مقالات موجود در مجموعه داده خود تعدادی را به ترتیب میزان ارتباط با علاقه‌مندی کاربر به او پیشنهاد می‌دهند، کارهای انجام شده در این زمینه را می‌توان از نظر نحوه گرفتن علاقه‌مندی‌های کاربر به دو دسته تقسیم کرد: (۱) کارهایی که با توجه به پروفایل کاربر علاقه‌مندی‌های او را استخراج کرده و به او پیشنهاد می‌دهند و (۲) کارهایی که متنی را دریافت کرده و کارهای مرتبط با آن متن را پیشنهاد می‌دهند، متن ورودی شامل اطلاعاتی است که که کاربر قصد مطالعه بیشتر در آن زمینه را دارد.

۲-۲-۱ پیشنهاد مقاله مبتنی بر پروفایل کاربر

روش‌های موجود در این دسته در یک مرحله پروفایل کاربر را استخراج کرده و در مرحله دیگر مقالاتی که مشابه علاقه‌مندی‌های کاربر باشد را به او پیشنهاد می‌دهند، در ادامه تعدادی از کارهایی که در این دسته می‌باشند توضیح داده شده است.

[1] Noun phrase

باسیو و همکارانش در (Basu et al., 2001) مساله پیشنهاد مقاله را این گونه مطرح کردند که قرار است مقالات یک کنفرانس به اعضای کمیته بازبینی آن داده شود. به طور خاص رویکرد آن‌ها برای پیشنهاد مقاله، تخصیص مقالات به بازبین‌گرها با توجه به خصوصیات آن‌ها می‌باشد. برای مثال مقالات می‌توانند به وسیله عناوین، چکیده و لیست کلمات کلیدی نمایش داده شوند، و اطلاعات بازبین‌گرها با تحلیل مقالات نوشته شده توسط آن‌ها و استخراج اطلاعاتی درباره تخصص آن‌ها بدست می‌آید. بعد از این‌که این اطلاعات بدست آمد، یک ماتریس مقاله-بازبین‌گر تولید می‌شود و هر پیشنهادی می‌تواند با پرس‌جو گرفتن از پایگاه‌داده‌های ماتریس بدست آید. برای مثال پرس و جوی زیر به این منظور نوشته شده است.

SELECT Reviewer.Name, Paper.ID

FROM Paper AND Reviewer

WHERE Reviewer.Descriptor SIM Paper.Abstract

میدلتون و همکارانش نیز در (Middleton et al., 2004) روشی در زمینه پیشنهاد مقالات پژوهشی با هدف جستجو برای مقالات مرتبط پیشنهاد دادند. رویکرد آن‌ها برای رسیدن به هدف خود منطبق کردن هستان‌شناسی-های[1] مقالات و کاربران است. ویژگی‌های یک مقاله با یک بردار که شامل عنوان‌های وابسته به هستان‌شناسی است ارائه می‌شوند. پروفایل کاربر نیز به کمک ابزارهای تعبیه شده‌ای که رفتارهای کاربر در سیستم را دنبال می‌کنند، استخراج می‌شود. برای مثال یک کاربر یک مقاله خاص، این امکان را دارد که مشخص کند در مورد آن مقاله علاقه‌مند، بدون علاقه و یا بی‌نظر است. یک کاربر همچنین می‌تواند مقالات را با دسته‌بندی‌های از پیش تعریف شده توسط سیستم برچسب‌گذاری کند. در نهایت پروفایل‌های علاقه‌مندی‌های کاربر بدست آمده و پیشنهادات با تکنیک‌های بازیابی اطلاعات بدست می‌آید.

بوگرز و بوسچ در (Bogers and Bosch, 2008) توضیح دادند که چگونه می‌توان از سایت *CiteULike*[2] که یک سایت مدیریت مراجع اجتماعی[3] است، برای پیشنهاد دادن مقاله‌های علمی به کاربران بر اساس مجموعه مراجع‌شان که به‌طور ضمنی نشان‌دهنده پروفایل آن‌ها می‌باشد، استفاده کرد. *CiteULike* یک نمونه خاص از

[1] Ontology
[2] http://www.citeulike.org/faq/data.adp.
[3] Social reference management

سایت‌های نشانه‌گذاری اجتماعی[1] است. کاربران می‌توانند در این سایت‌ها عضو شوند و بعد منابع مختلفی را که روی وب استفاده می‌کنند را نشانه‌گذاری کنند. مثلا یک نفر آدرس صفحات وبی را که برایش جالب است را در یک سایت نشانه‌گذاری ذخیره می‌کند و برای هر کدام نیز تگ‌هایی انتخاب می‌کند. در نتیجه در آینده می‌تواند در بین این آدرس‌ها به جستجو (با استفاده از تگ‌ها) بپردازد. معمولا در این سیستم‌ها کاربران می‌توانند از نشانه‌گذاری‌های یکدیگر نیز استفاده کنند. کاربران *CiteULike* می‌توانند مجموعه مراجع مورد نظرشان را در آن ذخیره و با استفاده از تگ‌ها سازماندهی نمایند (این مراجع بطور خاص، مقاله های علمی می باشند) سپس می‌توانند مراجع کاربران دیگر را که دارای تگ‌های مشابهی هستند یا دارای شهرت و عمومیت بیشتری هستند را بازیابی کنند. همین مساله در سطح نویسنده نیز قابل انجام است. یعنی می‌توانند مراجعی را که توسط کاربران دیگر ثبت شده‌اند و نویسنده آن مراجع با نویسنده مراجع مورد نظر کاربر، یکسان است را جستجو کنند.

در (Bogers and Bosch, 2008)، هدف این است که با استفاده از سرویس تحت وب *CiteULike*، بتوان برای کاربر، یک لیست از مراجع مناسب تولید و به او پیشنهاد کرد. بوگر و بوسچ برای رسیدن به این هدف با کمک داده‌های بدست آمده از *CiteULike*، سه الگوریتم مختلف فیلتر همبستگی پیشنهاد کردند: الگوریتم اول، مبتنی بر آیتم است و برای تشخیص مشابهت اشیا از فاصله کسینوسی استفاده می کند، الگوریتم دوم نیز مبتنی بر آیتم است و برای تشخیص مشابهت‌ها از احتمال شرطی استفاده می‌کند. الگوریتم سوم، مبتنی بر کاربر است و از فاصله کسینوسی استفاده می‌کند. در مدل فیلتر همبستگی مبتنی بر کاربر، هر کاربر دارای پروفایلی است که اشیای مورد علاقه‌اش در آن ثبت شده‌اند.

آن‌ها به این نتیجه رسیدند که معمولا روش مبتنی بر آیتم زمانی خوب است که تعداد کاربران به نسبت تعداد اشیا بیشتر باشد و روش مبتنی بر کاربر زمانی خوب عمل می کند که تعداد اشیا از تعداد کاربران بیشتر باشد. در این مقاله با توجه به آزمایشات تجربی انجام شده در مجموع نتیجه گرفته‌اند که الگوریتم مبتنی بر کاربر به میزان قابل توجهی، از دو الگوریتم دیگر بهتر عمل می‌کند و کارایی اش هم قابل قبول می‌باشد.

[1] Social bookmarking

گوری و پوسی در(Gori and Pucci, 2006) یک الگوریتم پیشنهاد مقالات تحقیقاتی ارائه کردند، که مبتنی بر گراف استناد و خصیصه‌های گشت‌زن تصادفی[1] است. این الگوریتم که *PaperRank* نام دارد می‌تواند مجموعه‌ای از مقالات موجود در یک کتابخانه الکترونیکی که از طریق مراجع به یکدیگر پیوند داده شده‌اند را بررسی کرده و به هر کدام یک امتیاز اولویت[2] نسبت دهد.

در واقع کاری که در (Gori and Pucci, 2006) انجام شده است این است که آن‌ها الگوریتم *pageRank* گوگل را برای هدف خودشان تطبیق داده‌اند. به عبارتی *PaperRank* یک نسخه بایاس شده از الگوریتم *PageRank* است که برای استفاده در سیستم‌های پیشنهاددهنده طراحی شده است. الگوریتم گوگل ایده‌اش این است که محاسبه می‌کند که یک کاربر که بطور تصادفی بین لینک‌های صفحات وب جابجا می‌شود، چقدر احتمال دارد که صفحه x را ببیند، یعنی چقدر احتمال دارد که در حین اینکه لینک‌ها را بطور تصادفی دنبال می‌کند، به صفحه x برسد. هرچه میزان این احتمال برای صفحه‌ای بیشتر باشد، آن صفحه در مرتب‌سازی گوگل، اولویت بیشتری می‌گیرد. گوگل برای محاسبه این احتمال از تعداد لینک‌های خروجی صفحات استفاده می‌کند.

در الگوریتم ارائه شده در (Gori and Pucci, 2006) هم ایده‌ای مشابه الگوریتم گوگل استفاده شده است. روش کار بدین ترتیب است که ابتدا گراف استناد مربوط به کل مقالات موجود در پایگاه ایجاد می‌شود. سپس کاربر، یک مجموعه کوچک از مقالات مورد نظرش را بعنوان نمونه انتخاب می‌کند این مقالات پروفایل کاربر را مشخص می‌کنند. بعد سیستم با توجه به این مقالات و استفاده از گراف استناد مقالات مشابه که از پایگاه‌داده‌ها استخراج می‌شود. مشابه الگوریتم گوگل، برای هر مقاله، یک امتیاز مشخص می‌کند. امتیاز مقاله x، مشخص می‌کند که وقتی یک گشت زن تصادفی از یکی از مقالات مورد علاقه کاربر شروع به گشت زنی کند، چقدر احتمال دارد که به مقاله x برسد. هرچه این احتمال بیشتر باشد، امتیاز مقاله x برای پیشنهاد شدن به کاربر افزایش می یابد. برای محاسبه این احتمال و امتیاز از فرمول‌های مربوط به گراف استناد استفاده شده است.

چندراسکان و همکارانش در(Chandrasekaran et al, 2008) و پوداکاتری و همکارانش در (Pudhiyaveetil et al, 2009) سیستمی برای پیشنهاد مقاله با توجه به پروفایل کاربر پیشنهاد دادند، آن‌ها برای

[1] Random walker
[2] Preference score