یک سیستم توصیه برای یافتن متخصصان در انجمن های علمی آنلاین

1. مقدمه

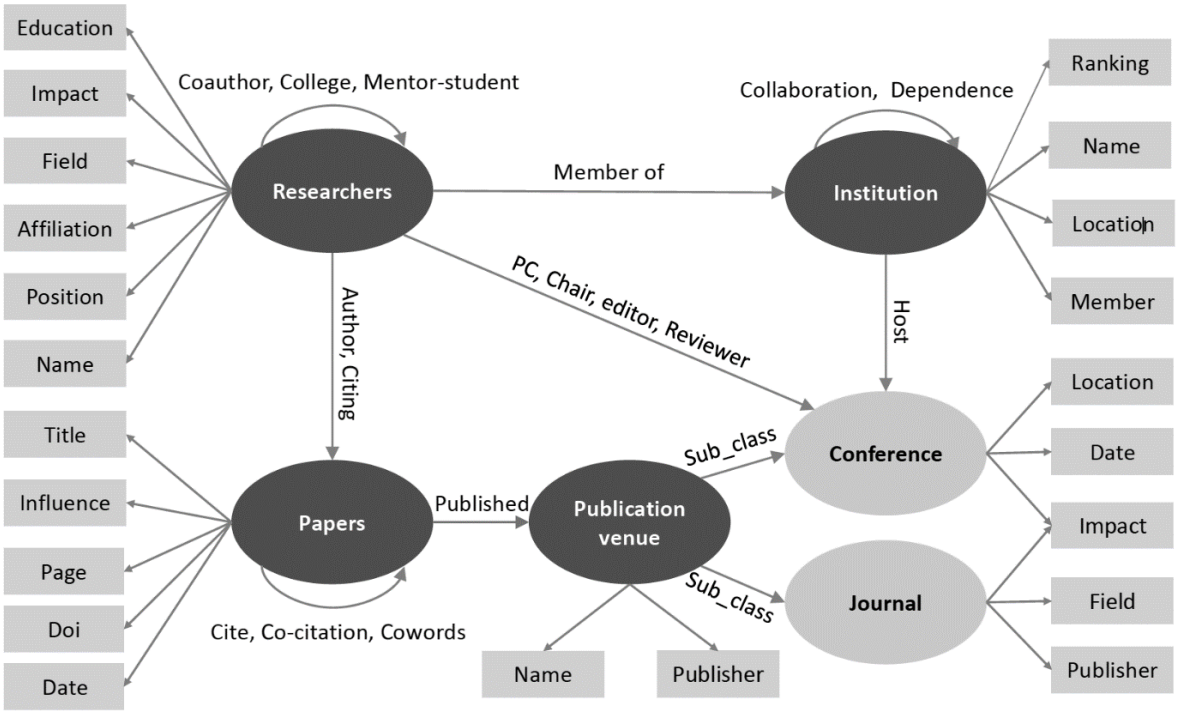
امروز، ما شاهد رشد روزافزون منابع اطلاعاتی در وب هستیم و دسترسی به اطلاعات یک چالش مهم برای کاربران است. موتورهای جستجو نیز برای پاسخگویی به این نیازها توسعه یافته اند اما حجم داده های بازیابی شده توسط سیستم زیاد است و یافتن اطلاعات مربوط به درخواست کاربر بسیار دشوار است [1]. رایج ترین مشکل اکثر سیستم های جستجوی وب عدم توجه به تفاوت بین علایق کاربران و بازیابی نتایج یکسان برای پرس و جوهای مشابه است [1] جستجوی سفارشی به دلیل ارائه نتایج مطابق با علاقه کاربران، نقش مهمی در تأمین اطلاعات مورد نیاز آنها دارد [1، 2]. در جوامع علمی آنلاین، محققان از موتورهای جستجو برای یافتن مقالات علمی، یک متخصص برای همکاری و محل انتشار استفاده می کنند، اما در بسیاری از موارد، به دلیل جستجوی مبتنی بر کلمات کلیدی و عدم توجه به محتوا، به نتیجه نمی رسند نتایج دلخواه در مراحل اولیه. این مقاله به سیستم یافتن خبره (EFS) می پردازد. روند کلی EFS با جمع آوری داده ها و سایر عناصر قابل استفاده برای تعیین زمینه های مهارت آغاز می شود. با شناسایی زمینه های تخصصی، این سیستم ها روش های مختلفی را برای کالیبره کردن متخصصان در یک موضوع خاص اعمال می کنند. روند جستجوی تخصص در EFS اتوماتیک کاملاً مشابه کاری است که بشر انجام می دهد. تنها تفاوت در این است که EFS علاوه بر تأمین نیازهای کاربران، می تواند سریعتر و دقیق تر باشد. EFS توانایی استخراج متخصصان و زمینه های تخصصی مختلف را از مجموعه داده های بزرگ و پیچیده در مقایسه با تجزیه و تحلیل فردی دارد [3]. علاوه بر این، دریافت پاسخ بسیار سریع و آسان، و ارتباطات موثر برای کاربران برای همکاری کاربران بسیار مهم است و از EFS برای یافتن افرادی که برای همکاری خوب هستند استفاده می شود [4، 28]. تلفیق تخصص تیمی از محققان اغلب می تواند نتایج بهتری نسبت به یک اثر فردی داشته باشد. اگر بدانیم که هر محقق تا چه اندازه و در چه زمینه های تخصصی خبره است، به طور بالقوه می توانیم از این دانش برای یافتن محققان با پیشنهاد تخصص و همکاری مناسب استفاده کنیم [4، 5].EFS دانشگاهی برای کارهای مختلف، مانند یافتن بازبینی مقاله، استاد راهنما، کارشناسان مشابه، همکاران دانشگاه و صنعت و همکاران تحقیق توسعه یافته است. این سیستم ها همچنین می توانند به دانشگاه ها در مدیریت دارایی های دانش خود و یافتن شکاف در مناطق خاص کمک کنند [6].

در این کار، ما نشان می دهیم که چگونه داده های فراداده برای شناسایی مناطق تخصصی استفاده می شود و سپس با استفاده از روش های بازیابی اطلاعات و مطالعه تکنیک های موجود در این زمینه، روشی برای پیشنهاد یک متخصص مناسب ارائه می شود. روش پیشنهادی برای مجموعه داده های IEEE اعمال می شود، که به عنوان یک پایگاه داده علمی استاندارد در نظر گرفته می شود. ما همچنین از داده های دنیای واقعی برای ارزیابی عملکرد سیستم استفاده کردیم.

در بخش 2 این مقاله، ما در مورد کارهای تحقیقاتی مربوط به متخصصان جستجو و یافتن سفارشی بحث می کنیم. در بخش 3، مدل پیشنهادی با جزئیات ارائه شده است. پیکربندی، مجموعه داده، پردازش قبل، فن آوری های کاربردی و ارزیابی به ترتیب در بخش 4 شرح داده شده است. در پایان، ارائه ها و همچنین کارهای آینده در بخش 5 خلاصه می شود.

2. آثار مرتبط

اصطلاح \"داده های بزرگ علمی\" که برای منابع علمی که به سرعت در حال رشد هستند اختصاص داده شده است شامل نویسندگان، مقالات، استنادها، کتابخانه های دیجیتال، شبکه های اجتماعی دانشگاهی و غیره است که چالش های جدیدی را در رابطه با مدیریت و تجزیه و تحلیل داده ها ایجاد می کند. انگیزه اصلی کار بر روی این مشکل استخراج دانش به منظور ارائه خدمات آکادمیک بهتر برای محققان است. مقاله ای با عنوان \"Big Scholarly Data A Survey\" با بررسی خصوصیات این نوع داده های بزرگ و همچنین کاربردهای آنها، این موضوع را مورد بحث قرار داد [7]. بر اساس این مطالعه، همانطور که در شکل 1 مشاهده می کنیم، انواع مختلفی از دانشمندان بزرگ از انواع مختلف موجودیتها و انواع بیشماری از روابط بین این موجودات استخراج شده است که آن را به یک سیستم پیچیده تبدیل می کند. شبکه هایی با چنین مشخصاتی شبکه های ترکیبی هستند که به ما امکان می دهد برخی از خصوصیات کلی مدرسه را نشان دهیم



شکل 1. موجودات عمده و روابط آنها در داده های بزرگ علمی [7].

با توجه به رشد سریع اطلاعات در جوامع علمی آنلاین، محققان از موتورهای جستجو برای یافتن نیازهای خود استفاده می کنند. جستجو در این سیستم ها دارای مشکلاتی است و اغلب کاربران در زمان مناسب به نتایج دلخواه خود نمی رسند. از این رو، نیاز به سفارشی سازی جستجو احساس می شود جوامع علمی آنلاین، و جستجوی سفارشی سعی می کند نتایج جستجو را مطابق مشخصات کاربر ارائه دهد [8، 9]. در بخش زیر، ما به طور خلاصه مثالی از کاربرد سیستم های توصیه گر را در حوزه علمی شرح خواهیم داد.

با توجه به رشد عظیم کنفرانس ها و مجلات علمی، یکی از موضوعات مهم انتخاب مناسب ترین مکان برای انتشار مقالات علمی است و این موضوع در برخی از کارهای پژوهشی مورد بررسی قرار گرفته است. در این مورد، یکی از سیستم های اخیر با استفاده از مفاهیم تحلیل شبکه اجتماعی و روش های فیلتر محتوای مبتنی بر محتوا، می تواند مناسب ترین مکان ها را برای چاپ مقاله ای که به کاربر نوشته شده است، توصیه کند. سیستم پیشنهادی در آن مطالعه، هویت نویسنده و عنوان مقاله نوشته شده را در ورودی دریافت می کند و سپس با استفاده از پایگاه داده تعریف شده از اطلاعات کتابخانه، شرکا را شناسایی می کند. سپس پس از اندازه گیری تشابه عنوان مقاله و نشریات هر یک از

شرکای قبلی نویسندگان و شناسایی مقالات، کنفرانس ها یا ژورنال های مشابهی که فرد (افراد) مقاله مشابه خود را به آنها ارسال کرده است، مکان های مرتبط در نظر گرفته می شوند. سپس عملیات را می توان به شرکای نویسنده بازگرداند و کارهای بیشتری انجام داد. ارزیابی با استفاده از داده های دنیای واقعی نیز عملکرد خوب آن را در ارائه توصیه های موثر نهایی نشان داد [10]. امروزه تقاضا برای مدیریت دانش افزایش یافته است. یکی از فاکتورهای مهم مدیریت دانش، یافتن شخصی با تخصص بالا در یک زمینه خاص است. روش مرسوم برای انجام این کار براساس رابطه بین افراد است. از این رو، یک روش سیستماتیک برای شخصی سازی فیلتر اطلاعات مورد نیاز است [11]. در مطالعه ای در این زمینه، عملکرد جستجو با تطبیق اولویت تاریخی کاربران در مجموعه مواردی با الگوهای مشابه با استفاده از فیلترهای مشترک افزایش یافت [12]. از طرف دیگر، برخی از نویسندگان معتقدند که بیشتر مطالعات در زمینه توصیه های علمی به شبکه های همگن نظیر محدود می شود. نتیجه کار نشان می دهد که سطح بیشتری از نزدیکی شبکه های اجتماعی مانند مکان می تواند بر هدف محققان برای همکاری تأثیر بگذارد [13، 14]. بنابراین، فقط استفاده از ویژگی های مبتنی بر شبکه نمی تواند به توصیه خوبی منجر شود. س keyال اصلی برخی از کارهای تحقیقاتی این است که چگونه می توانیم از ویژگیهای متعدد موجود در شبکههای ناهمگن کتابشناختی (که ممکن است به طور ضمنی بر همکاری علمی تأثیر بگذارد) به طور موثر و کاربردی استخراج و استفاده کنیم. به عنوان مثال، تخصص محققان، توانایی آنها و همچنین تعداد دفعات همکاری از ویژگی های مهمی است که ممکن است در اجرای توصیه همکاری تأثیر بگذارد. اهمیت تجزیه و تحلیل شبکه های ناهمگن کتابشناختی و کاربردهای آن در سالهای اخیر به یک روند عمده تبدیل شده است [14، 15]. QuickStep یک سیستم پیشنهادی از مقالات علمی است که از هستی شناسی مباحث مقالات علمی، علوم رایانه و طبقه بندی ایجاد شده توسط Dimoz استفاده می کند. تفسیر معنایی مقالات شامل یافتن مقوله در هستی شناسی مباحث مقالات علمی با استفاده از روش k Nearest Neighbor (kNN) انجام می شود [16]. برخی از محققان ثابت کرده اند که ترکیبی از همسایگی و مسیریابی می تواند نتایج امیدوار کننده ای داشته باشد [17]. برخی از عوامل اضافی مانند دفعات همکاری بین دو محقق نیز ممکن است بر روابط بین آنها تأثیر بگذارد [18]. یک کار تحقیقاتی در این زمینه توسط [19] یک روش ترکیبی از پنج ویژگی سه شبکه ناهمگن را ارائه داده است که شامل شبکه موضوع تحقیق، شبکه همکاری پژوهشگر و شبکه موسسات است. در کار تحقیقاتی ذکر شده، یک روش مبتنی بر مدل زبان معرفی شده است که شباهت تخصصی را از جنبه های مختلف نشان می دهد به منظور افزایش دقت در پیش بینی ها، ویژگی جدیدی ایجاد شده است که ترکیبی از تعداد نویسندگان مقاله خاص و همچنین فرکانس همکاری در کوتاهترین مسیر بین دو گره است. سرانجام، از روش رتبه بندی(Support Vector Machine (SVM برای ترکیب پنج ویژگی در شبکه ناهمگن (3 لایه) استفاده شده است. روش پیشنهادی در سیستم توصیه در بستر ScholarMate مورد استفاده قرار گرفت و سرانجام نتایج آزمایشی رضایت بخشی بدست آمد. همچنین اشاره شد که محدودیت هایی وجود داشت و برای نشان دادن بهتر عملکرد روش پیشنهادی، نتایج تجربی باید بر روی مجموعه داده های بزرگتر آزمایش شود تا شواهد متقاعد کننده تری به دست آید. این روش پیشنهادی برای کمک به افراد در جستجوی همتای خود با یک بستر علمی اجتماعی در چین استفاده شد. گزارش شده است که می توان از ویژگی های بیشتر جنبه های دیگر مانند شباهت معنایی برای افزایش دقت پیش بینی استفاده کرد. به عنوان مثال، برخی از محققان محله محلی را در نظر گرفته و آنها را با شباهت های معنایی ترکیب کرده اند و نتایج آزمایشات تأثیر مناسب آن را بر توصیه های همکار نشان می دهد [18]. در مطالعه دیگری، روش محاسباتی برای کشف مفاهیم مفید شبکه های اجتماعی و چگونگی استفاده از این مفاهیم در طراحی سیستم توصیه متخصص ارائه شده است. روش تجزیه و تحلیل منطقی در جامعه بزرگ منبع آزاد Ohloh اجرا شد. مشاهده شد که شباهت ملیت، محل برگزاری و ترجیحات زبان برنامه نویسی و همچنین شناخت اجتماعی در شکل گیری روابط بین اعضا ضروری است. اگرچه هیچ تضمینی برای همکاری آنها وجود نداشت [20]. بعلاوه، مشخص شد که کار با دیگران در صورت نزدیک بودن به یکدیگر می تواند م moreثرتر باشد و علاوه بر این، کاربران مشاوره با کارشناسان آشنا و مورد اعتماد را بیشتر در نظر می گیرند [4] می توان گفت که یافتن متخصصان مناسب برای کار نه تنها به اقتدار آنها در موضوع مربوط است بلکه کارآیی ارتباطات نیز بسیار مهم بود. بنابراین سیستم های پیشنهادی و شبکه های دانش برای یافتن یک متخصص در سازمان ها و جوامع علمی بسیار مهم هستند و توصیه متخصص مناسب به دلیل

نیاز به استدلال شبکه های ناهمگن پیچیده و همچنین نیاز به در نظر گرفتن میل افراد آسان نیست اگرچه در طی دهه گذشته تلاشهای زیادی در زمینه توسعه روشهایی برای افزایش دقت توصیه ها انجام شده است، اما با توجه به انگیزه افراد هنوز توصیه های سفارشی وجود دارد. در حالی که کارهای قبلی بر شناسایی متخصصان متمرکز بود، سفارشی سازی انتخاب متخصص روش دیگری بود که از طریق برنامه ای از جنبه علوم اجتماعی برای مدل سازی انگیزه کاربران در نظر گرفته شد [21]. در این مطالعه، یک سیستم پیشنهادی پیشنهاد شده است تا نتیجه را از طریق مشخصات انگیزه کاربران و روابط آنها تنظیم کند. الگوریتمی توسط وانگ و همکاران معرفی شده است. به منظور یافتن یک متخصص به نام ExpertRank که تخصص را بر اساس ارتباط اسناد و اعتبار متخصص در جامعه آنلاین ارزیابی می کند [22]. EFS پیشنهادی از سه شاخص برای توصیه متخصص استفاده می کند

1. یک متخصص مبتنی بر افشای اطلاعات به این روش پیدا کنید، متخصصان صریحاً تخصص خود را در پروفایل خود اعلام می کنند. ممکن است زمانبر باشد و با پیشرفت در تخصص کاربران ممکن است نمایه ثابت بماند.

2. بر اساس اسناد و مدارک نوشته شده یا توسط یک متخصص در صورت موجود بودن یک متخصص پیدا کنید می تواند یک شاخص خوب باشد و با استفاده از متن کاوی و تکنیک های بازیابی اطلاعات، می توان نتایج خوبی بدست آورد..

3. یک متخصص پیدا کنید که بر اساس تجزیه و تحلیل شبکه های اجتماعی باشد مطالعات جامعه شناختی نشان می دهد که تأثیر وضعیت اجتماعی نقش مهمی در انتخاب ها دارد

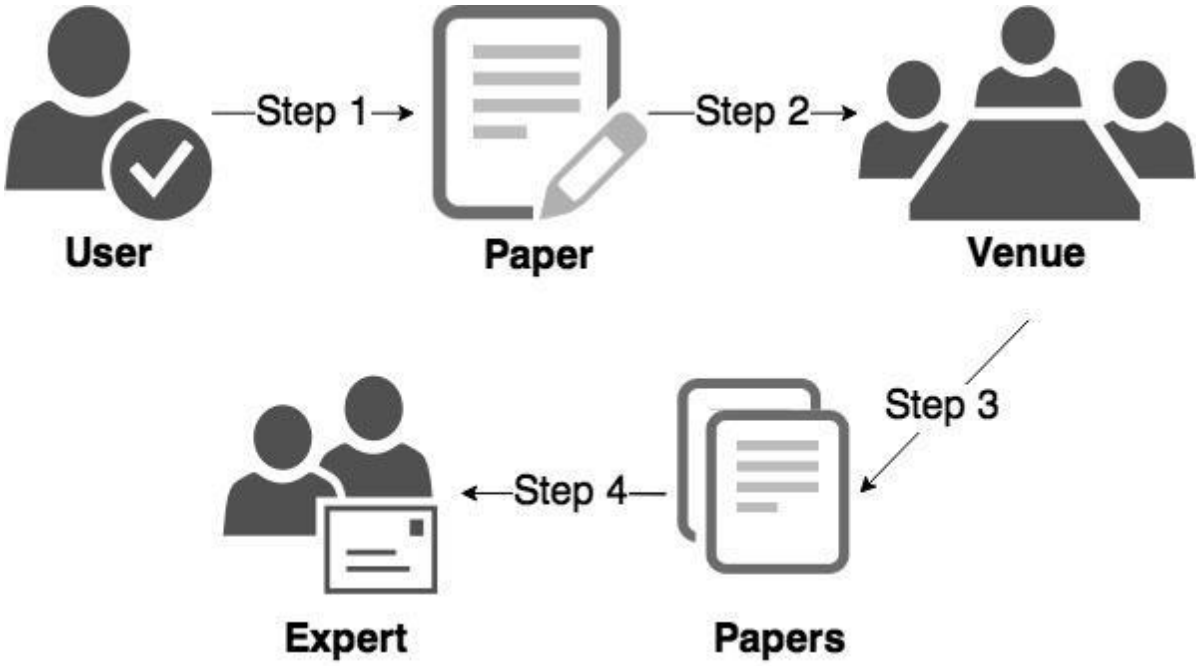
علاوه بر این، فعالیت های اجتماعی متخصص را می توان با کمک مجموعه داده های فیس بوک یا توییتر بررسی کرد. یاهو Answers، Stack Overflow و Quora نیز به دلیل کاربردهایشان برای یافتن افراد متخصص در سیستم های پاسخ به س attentionالات، توجهات را به خود جلب کرده اند. اخیراً شبکه های اجتماعی چندرسانه ای استخراج از اهمیت بیشتری برخوردار شده اند. منابع چندرسانه ای مانند ویدئوهای YouTube ممکن است درباره مهارت افراد اطلاعات ارزشمندی داشته باشند [23، 24]. با این حال، این شبکه های اجتماعی نگرانی اصلی ما نیستند، زیرا ما بر جوامع دانشگاهی و به طور خاص منابع کتابشناسی تمرکز می کنیم. از نظر دانش ما، چند مقاله وجود دارد که این منطقه را شکل می دهد. یکی از مشکلات اساسی رویکردهای ارائه شده، فقدان مجموعه داده های مناسب برای ارزیابی است. روش های پیشنهادی معمولاً از مجموعه ای از مجموعه داده های از پیش تعریف شده استفاده می کنند که باعث سوگیری می شوند، در حالی که مجموعه داده های دنیای واقعی به ما امکان می دهد ارزیابی واقع بینانه تری داشته باشیم. مسئله دیگر تعریف اقدامات جدید با توجه به EFS علمی است که با سایر محیطهای شبکه اجتماعی متفاوت است.IEEE Xplore دسترسی آزاد به داده های مورد نیاز را فراهم می کند و گزینه مناسبی برای این کار است. به منظور پاسخگویی به برخی از این خواسته ها، ما قصد داریم یک روش موثر با یک روش وزن سبک برای توصیه متخصص ارائه دهیم. این سیستم می تواند با تجزیه و تحلیل فراداده های پایگاه داده علمی IEEE، لیستی از متخصصان تولید کند. در ادامه، با تعریف اقدامات مناسب در محدوده تحقیق، با موضوع دوم نیز روبرو خواهیم شد

3. مدل پیشنهادی

طبق این مطالعه، ایده اصلی این است که با استفاده از اطلاعات کتابشناختی نشریات کاربر، محل انتشار و مقالات منتشر شده، می توان متخصصان را در زمینه خاصی پیدا کرد و برای همکاری به کاربر ارائه داد. نمای کلی مدل ارائه شده در شکل 2 نشان داده شده است.

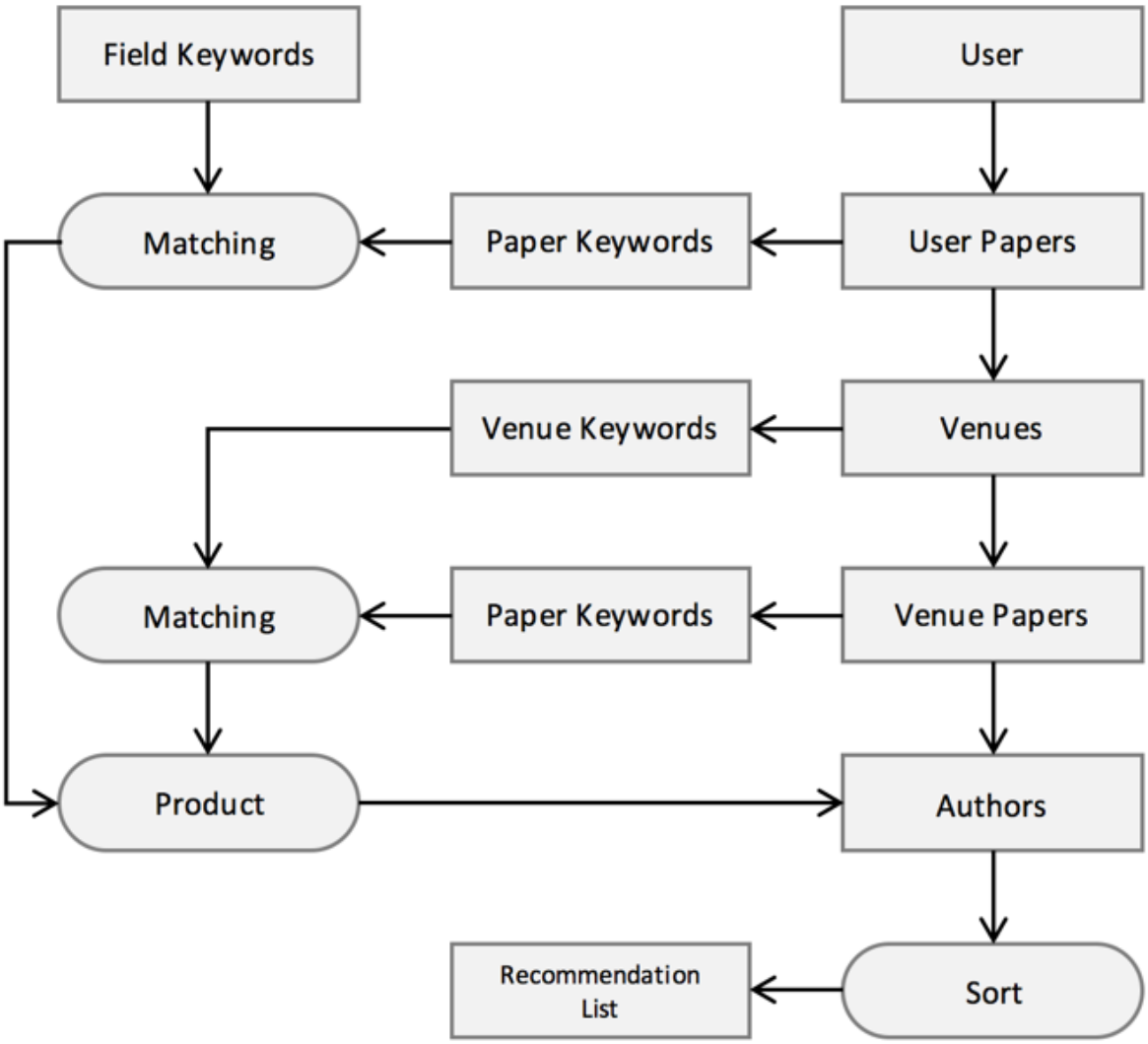
کاربر سیستم شخصی است که برای کار فعلی خود در یک زمینه خاص به دنبال یک متخصص است. ورودی سیستم شامل یک نام کاربری و کلمات کلیدی زمینه است. نام کاربری هویتی است که شخص با آن در جامعه علمی شناخته می شود و از کلمات کلیدی زمینه برای تعیین

حوزه علمی استفاده می شود که توسط کاربر صریحاً مشخص شده است. خروجی سیستم تعداد مشخصی از توصیه های رتبه بندی شده خواهد بود. هر توصیه نام یک متخصص است که سیستم نشان می دهد مناسب برای همکاری است. در روش پیشنهادی، سیستم نیاز به دسترسی به یک مجموعه داده دارد که شامل اطلاعات کتابشناسی مقالات علمی منتشر شده در سالهای اخیر است.



شکل 2. بررسی اجمالی مدل پیشنهادی.

داده های سال انتشار، محل انتشار، نام نویسندگان و کلمات کلیدی برای هر مقاله در مجموعه داده ضروری است. جدول 1 ویژگی های اصلی و زیر ویژگی ها را در پایگاه داده کتابشناسی نشان می دهد. روند سیستم با جزئیات بیشتر در شکل 3 نشان داده شده است.



شکل 3. فرآیند سیستم.

در ابتدا، با استفاده از نام نویسندگان، کاربر مقالات منتشر شده را در مدت زمان محدودی (یک تاریخ مشخص به تاریخ فعلی) از مجموعه داده بازیابی می شود. به اوراق مقاله های کاربر گفته می شود. همچنین مدت زمان مشخص شده را مدت کارنامه کاربر می نامند. دلیل این محدودیت زمانی این است که علایق تحقیق کاربر ممکن است با گذشت زمان تغییر کند. بعد، به هر یک از مقالات کاربر وزنی داده می شود که از طریق اندازه گیری شباهت بین کلمات کلیدی مقاله و کلمات کلیدی یک زمینه بدست می آید. ما در بخش بعدی در مورد این جزئیات بیشتر بحث خواهیم کرد. در واقع، وزن مشخص می کند که هر یک از مقالات کاربر که کارهای قبلی وی هستند، تا چه اندازه با کارهای فعلی وی ارتباط دارد.

سپس مقالات کاربر بر اساس محل انتشار طبقه بندی می شوند و وزن کلی که برای هر مقاله به این دسته داده می شود به محل انتشار مربوطه اختصاص می یابد. در نتیجه، مکان انتشار با وزن بیشتر به عنوان بیشتر مربوط به کارهای فعلی کاربر شناخته خواهد شد. محل انتشار

به دست آمده در این مرحله را مکان انتشار کاربر می نامند. سپس، برای هر مکان انتشار کاربر، کلمات کلیدی مکان، که مجموعه ای از کلمات کلیدی مقالات منتشر شده در مدت زمان محدود (یک تاریخ مشخص به تاریخ فعلی) است، بدست می آید. مدت زمان این مدت زمان مدت زمان کلمات کلیدی محل برگزاری است. دلیل این محدودیت زمانی بازیابی کلمات کلیدی به روزرسانی است که احتمالاً مربوط به نیاز کاربر است. در مرحله بعدی، با استفاده از محل انتشارات کاربر، مقالاتی که در محل برگزار می شوند، در یک بازه زمانی خاص به همین دلیل بازیابی می شوند. به این اوراق مقاله محل برگزاری گفته می شود. همچنین مدت زمان مشخص شده، مدت زمان مجلات محل برگزاری است. سپس به هر کاغذ سالن که از محصول وزن سالن کاغذ که قبلاً بدست آمده و وزن حاصل از اندازه گیری شباهت بین کلمات کلیدی کاغذ و کلمات کلیدی محل، وزنی داده می شود.

$$W_{venue\ paper} = W_{paper venue} \cdot S(paper\ keywords, venue\ keywords) \quad (1)$$

در این معادله، W مخفف وزنه است و S نشانگر محاسبه شباهت (یا فرآیند تطبیق) است که در بخش بعدی توضیح داده خواهد شد.

Table 1. Main features and sub-features.

Main Feature	Sub-feature(s)
User	Paper(s)
Paper	Author(s), Keywords, Publication Venue
Venue	Papers, Keywords

در نتیجه، محصول دو وزن اندازه گیری تعیین رابطه بین مقاله و کار فعلی کاربر را فراهم می کند. سپس، مقالات محل انتشار بر اساس نویسنده طبقه بندی می شوند و مجموع وزن های ارائه شده به هر مقاله مقوله به نویسنده مربوطه اختصاص می یابد. همانطور که قبلاً توضیح داده شد، می توانیم بگوییم که وزن نشان دهنده رابطه بین تخصص نویسنده و کار فعلی کاربر است، بنابراین از آن به عنوان معیار توصیه متخصص استفاده می شود. در آخرین مرحله، نویسندگان بر اساس وزن داده شده طبقه بندی می شوند و تعداد مشخصی با بیشترین وزن به عنوان لیست متخصصان ارائه می شود. به تعداد مشخصی، لیست پیشنهادی گفته می شود.

3.1 اندازه گیری شباهت ساده ترین معیار تشابه بین دو کلمه کلیدی برابری است که بر این اساس شباهت بین دو کلمه کلیدی مشابه 1 است و در غیر این صورت 0 است. با این حال، این اندازه گیری کارآمد نیست زیرا، به عنوان مثال، کلمه کلیدی رمزگشایی تکراری بیشتر است مربوط به برنامه نویسی کانال است تا رایانش ابری اما اندازه گیری برای هر دو آنها مقدار 0 را به ارمغان می آورد. از آنجا که در روش پیشنهادی، وزنی که از اندازه گیری شباهت بین کلمات کلیدی بدست می آید، نقش تعیین کننده ای در توصیه های نهایی دارد و کارایی اندازه گیری شباهت برای سیستم بسیار مهم است، در مبادله هزینه و کارایی، کارآیی مورد توجه قرار گرفت در روش پیشنهادی، برای اندازه گیری شباهت بین دو کلمه کلیدی، ما دو اندازه گیری متفاوت همسایه مشترک را ارائه می دهیم. اندازه گیری وقوع دو کلمه کلیدی به معنای تعداد مقاله هایی است که هر دو کلمه کلیدی در آنها نشان داده شده است. معیار همسایگان مشترک برای دو کلمه کلیدی نشانگر تعداد کلمات کلیدی همسایه است، یعنی حداقل هر یک از آنها در مقاله ظاهر شده اند. ترتیب شکل ظاهری یک کلمه کلیدی در یک مقاله به وضوح وجود آن در کلمات کلیدی مقاله است. برای دستیابی به هر یک از این اقدامات، مقالاتی را که در مدت زمان محدودی منتشر شده است بررسی می کنیم (یک تاریخ مشخص به تاریخ فعلی). مدت زمان این دوره از زمان را مدت اندازه گیری شباهت می نامند. جدول 2 مقادیر اندازه گیری شباهت را با مدت زمان دو سال برای چند جفت کلید واژه نشان می دهد.

Table 2. Values of the similarity measures with a duration of two years for a few keyword pairs.

First keyword	Second keyword	Co-occurrence	Shared neighbors
iterative decoding	channel coding	105	365
	computational complexity	37	388
	equalisers	15	289
	multi-access systems	14	265
	galois fields	7	141
	underwater acoustic communication	5	236
	delays	3	362
	radiocommunication	3	297
	statistical analysis	2	373
	multiprocessing systems	2	262
	polynomial approximation	1	173
	amplitude modulation	1	173
	newton method	1	170
	multi-threading	1	142
	cloud computing	0	271
	hardware-software codesign	0	135
	mathematical morphology	0	63
	bipolar logic circuits	0	5
	passive solar buildings	0	0
	strontium alloys	0	0

مقادیر اندازه گیری های تعریف شده طبیعی نیست. معادله (2) برای عادی سازی مقادیر استفاده می شود، جایی که S شباهت اندازه گیری عملکرد نرمال سازی است، M اندازه گیری شباهت است و k1 و k2 کلمات کلیدی هستند.

$$S(k_1,k_2)=\begin{cases}1 & ,k_1=k_2\\ 1-\dfrac{1}{M(k_1,k_2)+1} & ,k_1\neq k_2\end{cases} \quad (2)$$

به منظور اندازه گیری شباهت دو مجموعه از کلمات کلیدی، دو معیار تمایل مرکزی نیز در نظر گرفته شده است یکی میانگین مشابهت دوتایی کلمات کلیدی مجموعه ها، و دیگری دیگر، به طور مشابه، میانه کلمات کلیدی مجموعه تشابه جفتی. از آنجا که دو اندازه گیری متفاوت برای اندازه گیری تشابه دو کلمه کلیدی (همسایگی مشترک و همسایگان مشترک) ارائه شده است، و دو اندازه گیری متفاوت برای اندازه گیری شباهت دو مجموعه کلمات کلیدی (میانگین و متوسط) در نظر گرفته شده است، در مجموع چهار روش وجود دارد برای به دست آوردن شباهت دو مجموعه از کلمات کلیدی. همانطور که بحث خواهد شد، پس از ارزیابی چهار روش، بهترین نتیجه با استفاده از معیارهای اندازه گیری وقوع و میانگین بدست می آید

. 4- نتایج تجربی

در این بخش، ما ابتدا به ارائه مجموعه داده مورد نیاز و پردازش قبل از آن و همچنین مشخصات مجموعه داده می پردازیم. سپس برخی از چالش ها بررسی می شوند.

4.1 جدول پیکربندی

3 پارامترها و مقادیر مختلف آنها را نشان می دهد. با توجه به این داده ها، در مجموع، هشت حالت مختلف ارزیابی وجود دارد. به هر حالت، شماره پیکربندی داده شده است، همانطور که در جدول 4 مشاهده می شود. مقادیر پارامترهای دیگر سیستم که در ارزیابی همه موارد آزمایشی ثابت نگه داشته شده است، جدول 5 است.

همانطور که بعداً بحث خواهیم کرد، پیچیدگی زمانی الگوریتم توصیه عملکرد غیر مستقیم پارامترهای توصیف شده در جدول 5 است. از یک طرف، انتخاب مقدار زیادی برای این پارامترها باعث کندی سیستم می شود. از طرف دیگر، مقدار بسیار کم به نتایج ضعیفی منجر خواهد شد. مقادیر نشان داده شده در جدول 5 به صورت آزمایشی انتخاب می شوند تا به تعادل خوبی بین تأخیر و کیفیت توصیه ها برسند.

Table 3. Assessment configuration parameters.

Parameter	Value
Similarity measure	Co-occurrence
	Shared neighbors
Central tendency measure	Mean
	Median
Keywords type	Controlled index terms
	Thesaurus terms

Table 4. Different configurations of the assessment.

No.	Keywords	Measure of measuring similarity	Central tendency measure
0	Controlled index terms	Co-occurrence	Mean
1	Controlled index terms	Co-occurrence	Median
2	Controlled index terms	Shared neighbors	Mean
3	Controlled index terms	Shared neighbors	Median
4	Thesaurus terms	Co-occurrence	Mean
5	Thesaurus terms	Co-occurrence	Median
6	Thesaurus terms	Shared neighbors	Mean
7	Thesaurus terms	Shared neighbors	Median

Table 5. Constant parameters in the assessment.

Parameter	Value
Duration of a user papers	2 years
Duration of keywords of the publication venue	1 years
Duration of the publication venue papers	2 years
Duration of the similarity measures	2 years

4.2 Dataset برای ارزیابی قابل اعتماد روش پیشنهادی، ما به مجموعه بزرگی از داده های دنیای واقعی نیاز داشتیم که شامل اطلاعات کتابشناسی مقالات علمی منتشر شده در سالهای اخیر باشد. یکی از گزینه های موجود برای مجموعه داده ها، کتابخانه کتابخانه دیجیتال (DBLP) بود. DBLP یک پایگاه داده کتابشناسی علوم کامپیوتر است که دانشگاه تریر در آلمان میزبان آن است. از آنجا که تمام داده های DBLP در یک فایل XML ذخیره می شوند، دسترسی به آنها ساده است. این مجموعه داده حاوی داده های سال انتشار، محل انتشار و نام نویسندگان است، اما داده کلمات کلیدی ندارد [25]. به همین دلیل، ما از آن استفاده نکردیم. ما کتابخانه دیجیتال IEEE Xplore را برای کار خود انتخاب کردیم این کتابخانه دیجیتالی شامل مقالات علوم کامپیوتر، مهندسی برق و الکترونیک است که توسط IEEE (موسسه مهندسان برق و الکترونیک) و دیگر ناشران شریک منتشر شده است. IEEE Xplore دسترسی به وب را به بیش از 3 میلیون سند علمی و فنی فراهم می کند و ماهانه حدود بیست هزار سند جدید به آن اضافه می شود. محتوای IEEE Xplore شامل موارد زیر است [26]

- بیش از 170 مجله
- بیش از 1400 مقاله کنفرانسی
- بیش از 5،100 استاندارد فنی

- حدود 2000 کتاب
- بیش از 400 دوره آموزشی

مهمترین داده های موجود برای اسناد IEEE Xplore بصورت رایگان در زیر ذکر شده است.

- موضوع
- نویسنده (ها)
- وابستگی نویسنده (ها) words کلمات کلیدی ven محل انتشار تنوع سند (کنفرانس ها، مجلات، کتاب ها، دسترسی زودرس، استانداردها یا دوره های آموزشی)
- ناشران (IEEE، AIP، IET، AVS یاIBM)
- سال انتشار
- چکیده
- ISBN
- ISSN شناسه دیجیتال شی (DOI)

از آنجا که IEEE Xplore دسترسی آزاد به داده های مورد نیاز بسیاری از مقالات موجود در کتابخانه دیجیتال در سال انتشار، محل انتشار، نام نویسنده (ها) و کلمات کلیدی را فراهم می کند، گزینه مناسبی برای کار ما است. اگرچه IEEE Xplore برخلاف DBLP داده های مورد نیاز را فراهم می کند، بارگیری مجموعه داده به عنوان پرونده امکان پذیر نیست. بنابراین لازم بود که داده های کتابخانه دیجیتال و ذخیره شده در یک پایگاه داده محلی به نوعی واکشی شود. دروازه جستجوی IEEE Xplore یک رابط برنامه نویسی برنامه (API) را برای جستجوی پایگاه داده کتابخانه دیجیتال فراهم می کند [27]. اگرچه برای ارزیابی این مطالعه، فقط چند سال انتشار کافی است، ما تصمیم گرفتیم داده های IEEE Xplore را در صورت نیاز برای استفاده در کارهای بعدی، واکشی و ذخیره کنیم

پاسخ های دروازه جستجوی IEEE Xplore به صورت XML است که هم برای انسان و هم برای ماشین قابل خواندن است. این یک استاندارد باز است. لازم به ذکر است که در داده های واکشی شده از درگاه جستجوی IEEE Xplore، دو نوع کلمه کلیدی \"اصطلاحات اصطلاحنامه\" و \"اصطلاحات نمایه کنترل شده\" وجود دارد. ما برای ارزیابی از هر دو نوع کلمات کلیدی استفاده کردیم.

همانطور که در آینده بحث خواهد شد، پس از ارزیابی هر دو نوع کلمات کلیدی به صورت جداگانه، متوجه شدیم که با استفاده از اصطلاحات نمایه کنترل شده، بهترین نتیجه حاصل شده است.

4.3 پیش پردازش

در پایگاه داده IEEE Xplore، برای یک کنفرانس در سال های مختلف، یک مکان متمایز در نظر گرفته شده است. مثال زیر از داده های واقعی پایگاه داده استخراج شده است .

- انفورماتیک زیست پزشکی و بهداشت ((BHI، IEEE EMBS 2012 کنفرانس بین المللی (شماره انتشار 6204368)
- انفورماتیک زیست پزشکی و بهداشت (BHI)، کنفرانس بین المللی IEEE EMBS 2014 (شماره انتشار 6853543)

همانطور که مشاهده می شود، برای یک کنفرانس در دو سال مختلف، دو شماره انتشار متفاوت در نظر گرفته شده است. با این حال، از آنجا که دامنه کنفرانسی که در سالهای مختلف برگزار می شود، ثابت است، ما مایل هستیم همه آنها را به عنوان یک مکان منحصر به فرد در نظر بگیریم. بنابراین، قبل از استفاده از مجموعه داده، باید این مشکل را برطرف کنیم. با توجه به تعداد زیاد اسناد موجود، بررسی و تصحیح دستی امکان پذیر نیست. بنابراین با بررسی تعداد قابل توجهی از اسناد، الگوی خاصی را شناسایی کردیم. در تمام موارد بررسی شده، همانند مثال قبلی، سال برگزاری کنفرانس در بخشی از موضوع کنفرانس ظاهر می شود که با یک ویرگول جدا می شود. با دانستن این الگو، تصحیح می تواند به صورت خودکار انجام شود. اگر ویرگول در موضوع انتشار ظاهر نشده باشد، دیگر نیازی به بررسی نیست. در غیر این صورت، ما موضوع را به بخشهایی جدا می کنیم و به دنبال یک سال می گردیم (به طور خاص، یک عدد صحیح بین سالهای 1800 و 2099). اگر در بخشی یک سال وجود داشته باشد، قسمت از موضوع خارج می شود. همچنین، اگر چندین موضوع مختلف از نشریات پس از تصحیح یکسان شوند، تعداد انتشار همه آنها نیز یکسان است. به عنوان نمونه، هر دو عنوان قبلی به Biomedical و Health Informatics (BHI) تغییر یافته اند. پس از استفاده از پردازش قبل، تعداد مکانهای منحصر به فرد از 27102 به 17252 کاهش یافت، که ارزیابی را دقیق تر کرد.

4.4 چالش ها

یک چالش اساسی در اجرا، پیچیدگی زیاد محاسبه شباهت کلمات کلیدی بود. با توجه به زمانبر بودن محاسبه شباهت بین دو کلمه کلیدی و تعداد زیادی کلمات کلیدی که باید در فرایند سیستم برای هر بار اندازه گیری شود، محاسبه آنلاین عملی و مقرون به صرفه نیست. بنابراین شباهت زوجی کلیه کلمات کلیدی مقالاتی که در مدت زمان تعیین شده منتشر شده اند محاسبه و ذخیره شده است. برای اینکه بتوانیم این کار را در یک زمان مناسب انجام دهیم، از یک روش پردازش چندگانه برای نوشتن برنامه استفاده کردیم در این مطالعه، برای مدت زمان اندازه گیری شباهت دو سال (2012 و 2013)، دو معیار همسانی مشترک و همسایگان مشترک برای 38،531،031 جفت کلید واژه موجود محاسبه و ذخیره شد. 4.5 ارزیابی ما سال 2014 را به عنوان سال ارزیابی در نظر گرفتیم، به این معنی که دسترسی سیستم به داده های مربوط به مقالات منتشر شده تا پایان سال 2013 محدود است. برای انتخاب یک مورد آزمایشی، ابتدا از بین همه نویسندگانی که مقاله ای را در سال ارزیابی منتشر کرده اند، حداقل به صورت تصادفی یکی انتخاب می شود اگر در ده سال منتهی به سال ارزیابی (به استثنای سال ارزیابی) چهار مقاله منتشر شده است.

Figure 4. Dataset Segmentation.

سپس در میان تمام مقالات منتشر شده نویسنده در طول سال ارزیابی، اگر هر دو نوع اصطلاح نامه و اصطلاحات کنترل شده برای مقاله در دسترس باشد، یکی به طور تصادفی انتخاب می شود. برای ارزیابی، نام نویسنده انتخاب شده و کلمات کلیدی انتخاب شده به عنوان ورودی سیستم نام کاربر و کلمات کلیدی زمینه و خروجی سیستم نام کارشناسان توصیه شده برای هر مورد آزمون ذخیره می شود یک چالش اساسی در ارزیابی روش پیشنهادی فقدان تدبیر مشخص برای قضاوت در مورد کیفیت یک پیشنهاد است. متأسفانه، رویکردهای موجود از پارامترهای مختلفی برای کار توصیه متخصص استفاده می کنند و روش های آنها به راحتی قابل تکرار نیستند. علاوه بر این، مجموعه داده استفاده شده آنها در دسترس عموم نیست. در نتیجه، ما نتوانستیم روش خود را در مجموعه داده های آنها اعمال کنیم تا مقایسه ای منصفانه داشته باشیم

برای انجام این کار، ما از سه متخصص خواسته ایم که داده ها را به صورت دستی برچسب گذاری کنند. برای ارزیابی هر مورد آزمایشی، با جستجوی نام کاربری و نام هر متخصص توصیه شده در وب، علایق و زمینه کارهای اخیر پیدا شد و سپس با استفاده از عقل سلیم، کیفیت توصیه ها مورد قضاوت قرار گرفت و هر مورد برچسب خوب یا بد داشت.

به منظور ارزیابی عملکرد سیستم در تنظیمات مختلف، ما سیستم را برای بیست مورد آزمایشی با هر یک از هشت پیکربندی ممکن پیاده سازی کردیم. به عنوان مثال یکی از موارد آزمایشی در زیر ذکر شده است:

- نام کاربری Marques، E
- اصطلاحات شاخص کنترل شده
- سنتز سطح بالا
- زبانc
- آرایه های قابل برنامه ریزی گیت میدانی یا زبانهای توصیف سخت افزار
- اصطلاحات اصطلاحنامه
- ساعت
- آشکارساز تشعشع
- سخت افزار
- تست معیار پردازش خط لوله
- آرایه های دروازه قابل برنامه ریزی

توصیه های سیستم برای مورد آزمایشی فوق، با پیکربندی 0، همراه با نتایج ارزیابی، در زیر ارائه شده است:

- Luk, W. – Good
- Kumar, A. – Good
- Amano, H. – Good
- Maruyama, T. – Not Good
- Cheung, P.Y.K. – Good
- Stroobandt, D. – Good
- Bruneel, K. – Good
- Betz, V. – Good
- Benkrid, K. – Good
- Chow, P. – Good

نتایج ارزیابی در جدول 6 نشان داده شده است.

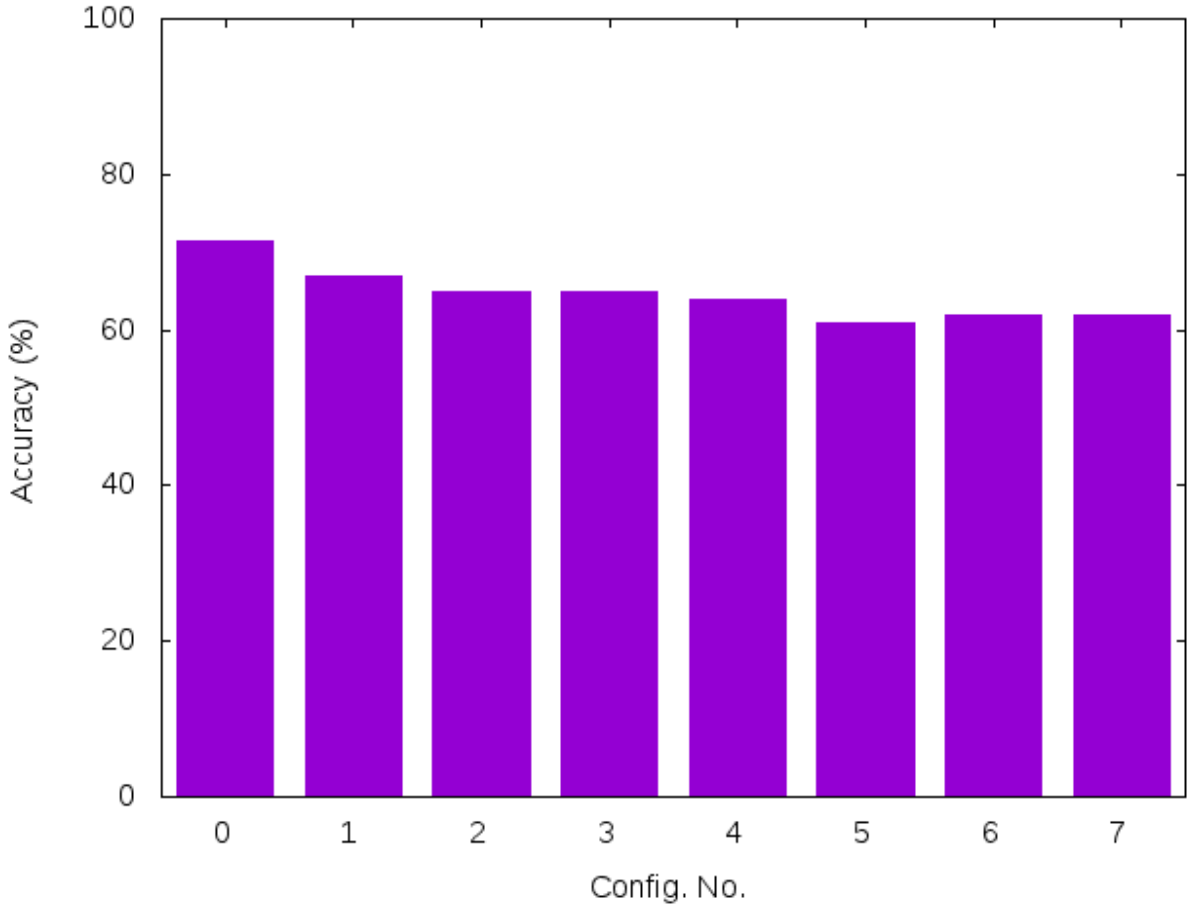
Table 6. Assessment results.

Configuration number	Accuracy (%)
0	71.50
1	67.00
2	65.00
3	65.00
4	64.00
5	61.00
6	62.00
7	62.00
Average	**64.60**

دقت سیستم در هر پیکربندی، درصد توصیه های دارای برچسب خوب در آن پیکربندی، برای بیست مورد آزمایشی است. همانطور که مشاهده می شود، پیکربندی شماره صفر بهترین نتیجه را با دقت 71.50٪ ارائه داده است.

4.6 تجزیه و تحلیل

مرحله اول الگوریتم نویسندگان را استخراج می کند، یعنی توصیه های بالقوه با وزن مرتبط آنها. پیچیدگی زمانی این مرحله$O(n)$، جایی است کهn تعداد کل آن است مقالات بازیابی شده به عبارت دیگر، برابر با تعداد مقالات کاربر (منتشر شده در مدت زمان مقاله کاربر) به علاوه تعداد مقاله های محل برگزاری (منتشر شده در مدت زمان مقاله محل برگزاری). این مرحله را می توان با پیچیدگی فضا$O(1)$ انجام داد زیرا ما فقط باید همزمان با یک مقاله کار کنیم.



شکل 5. دقت در تنظیمات مختلف.

عملیات اصلی در مرحله اول مطابقت است، یعنی محاسبه شباهت دو مجموعه از کلمات کلیدی. پیچیدگی زمانیمطابقت به خودی خود، اندازه O(m,n), m وn اندازه این دو مجموعه کلمات کلیدی است. برای مرحله اول این مرحله، این دو مجموعه عبارت User Keywords و Field Keywords هستند. برای مرحله دوم، اینها کلمات کلیدی Venue Paper و Venue Keywords هستند. انتظار می رود که کلمات کلیدی Field یک مجموعه کوچک باشد، زیرا کلمات کلیدی توسط کاربر وارد می شوند. این سیستم همچنین می تواند حد معقولی را اعمال کند. اندازه کلمات کلیدی User / Venue Paper تعداد کلمات کلیدی است که در مقاله ظاهر می شود، که معمولاً تعداد کمی است Venue Keywords مجموعه نسبتاً بزرگی است که اندازه آن توسط پارامتر Venue Keywords کنترل می شود. عملیات اصلی تطبیق محاسبه شباهت دو کلمه کلیدی است. از آنجا که شباهت تمام جفت کلمات کلیدی از قبل محاسبه و ذخیره شده است، در عمل مطابقت بسیار سریع است پیچیدگی فضایی این عملیات(1)O نیز است. هر دو معیار شباهتی که ما به کار برده ایم، یعنی همسایگی مشترک و همسایگان مشترک، دارای پیچیدگی زمانی(n)O هستند، که در آن n تعداد کل مقالاتی که در مدت زمان اندازه گیری شباهت منتشر شده است، است. از آنجایی که n می تواند بسیار بزرگ باشد، محاسبه شباهت در عمل هزینه بر است. با این حال، به محض محاسبه شباهت دو کلمه کلیدی، می توان از آن برای مدت طولانی استفاده مجدد کرد، زیرا شباهت نسبی کلمات کلیدی مورد استفاده در ادبیات اغلب تغییر نمی کند. به عنوان مثال،رمزگشایی تکراری به کدگذاری کانال نزدیکتر است تا رایانش ابری این اکنون درست است و احتمالاً در سال آینده نیز صادق خواهد بود. همچنین، کلمات کلیدی جدید فقط گاهی اوقات ساخته می شوند. همچنین، کلمات کلیدی جدید فقط گاهی اوقات ساخته

می شوند. ما از این خصوصیات کلمات کلیدی سو استفاده می کنیم تا یک حافظه پنهان از شباهت های کلمات کلیدی ایجاد کنیم که شامل تمام کلمات کلیدی است که در مقالات منتشر شده در مدت زمان اندازه گیری شباهت وجود دارد.

ایجاد چنین حافظه پنهانی هزینه بر است اما فقط باید به ندرت به روز شود، به عنوان مثال سالانه این حافظه پنهان به طور قابل توجهی سرعت سیستم توصیه ما را افزایش می دهد. پیچیدگی فضایی روش هم وقوع، روشی که در نهایت انتخاب شده O 1 است. در مورد روش همسایگان مشترک، پیچیدگی فضا O(m) در جایی که m، تعداد کلمات کلیدی متمایز در مقالاتی که در مدت زمان اندازه گیری شباهت منتشر شده است.

مرحله دوم الگوریتم توصیه های نتیجه نهایی k یا نویسندگان برتر k را برمی گرداند: این یک مشکل مرتب سازی جزئی است. با استفاده از یک راه حل مبتنی بر heap، می توان این کار را انجام داد، O(n.log(k)) جایی که n تعداد کل نویسندگان بازیابی شده است و k تعداد توصیه ها است. از آنجا که k یک ثابت از پیش تعیین شده است، این برابر است O(n) .