نشریه علمی- ترویجی محاسبات نرم
شمارهٔ دوم/ پاییز و زمستان ۱۳۹۱/ صفحه ۳۶-۴۵

نقش ارتباطات معنایی در بهبود نتایج یک سیستم پیشنهاد استناد

فتانه زرین‌کلام[۱]، محسن کاهانی[۲]
دانشجوی کارشناسی ارشد، دانشگاه فردوسی مشهد، مشهد، ایران
zarrinkalam.fattane@stu-mail.um.ac.ir[۱]
استاد، دانشگاه فردوسی مشهد، مشهد، ایران
kahani@um.ac.ir[۲]

چکیده: حجم فراوان و رو به رشد اسناد علمی منتشر شده بر روی وب، فرآیند تصمیم‌گیری و انتخاب اسناد مرتبط با یک زمینهٔ تحقیقاتی را برای پژوهشگران دشوار کرده است. استفاده از کتابخانه‌های دیجیتال رایج با مشکلات مختلفی نظیر ناکارامدی جستجوی مبتنی بر کلمات کلیدی و ضعف معیارهای شباهت متنی مواجه است. راهکار دیگری که در سال‌های اخیر مورد توجه قرار گرفته، استفاده از سیستم‌های پیشنهاد استناد است که با دریافت یک متن، اسنادی را که باید توسط آن متن مورد استناد قرار گیرند، پیشنهاد می‌کند، و بدین ترتیب می‌تواند در یافتن اسناد مرتبط با یک موضوع به پژوهشگر کمک کند. در این مقاله، نقش ارتباطات معنایی اسناد در کنار ویژگی‌های متنی آن‌ها در بهبود نتایج یک سیستم پیشنهاد استناد مورد بررسی قرار گرفته است. نتایج ارزیابی‌ها نشان می‌دهد که در نظر گرفتن ارتباطات معنایی نقش به‌سزایی در تشخیص شباهت اسناد دارد و باعث بهبود کیفیت سیستم پیشنهاد استناد مورد نظر می‌شود.

واژه‌های کلیدی: پیشنهاد استناد، ارتباطات معنایی، شباهت متنی.

۱. مقدمه

هر پژوهشگری قبل از شروع کاری جدید در زمینهٔ مورد علاقهٔ خود باید از کارهای انجام شده دربارهٔ 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، کاملاً محسوس است [۱ و ۲].

امروزه، اغلب پژوهشگران برای یافتن کارهای مرتبط با یک موضوع، از روش‌های رایج، مثل جستجو در گوگل استفاده می‌کنند. ورودی این روش‌ها، اغلب تعدادی واژهٔ کلیدی، و خروجی آن‌ها اسنادی است که شامل این کلمات کلیدی هستند.

بدین ترتیب، اگر پژوهشگری در ارتباط با موضوع مورد علاقهٔ خود متنی داشته باشد و کارهای مرتبط با آن را بخواهد، ابتدا باید واژه‌های کلیدی موجود در متن را استخراج کند. استخراج این واژه‌ها برای پژوهشگری که به تازگی به تحقیق در یک زمینه پرداخته است، کار آسانی نیست. خروجی موتورهای جستجو برای این واژه‌های کلیدی، تعداد زیادی از اسنادی هستند که شامل این واژه‌های کلیدی می‌باشند، در نتیجه پژوهشگر باید زمان زیادی را صرف انتخاب گزینه‌های بهتر از بین خروجی‌ها کند [۳].

راه حل برطرف کردن چنین مشکلاتی، وجود یک سیستم پیشنهاد استناد[1] است که ورودی آن، یک قطعه متن و خروجی آن اسنادی است که باید در آن متن مورد استناد قرار بگیرند، به عبارتی اسناد مرتبط با آن متن است.

تکیهٔ تنها بر شباهت معنایی در سیستم‌های پیشنهاد استناد برای به دست آوردن شباهت بین دو سند، با توجه به مشکلات آن نظیر پیچیدگی و ابهام [٤] و همچنین با توجه به اینکه در یک سیستم پیشنهاد استناد، هدف به دست آوردن اسناد مرتبط است و نه صرفاً اسنادی که شباهت متنی زیادی با متن ورودی دارند، نمی‌تواند به خوبی در یافتن اسناد مرتبط با یک موضوع به پژوهشگر کمک کند.

در این مقاله، برای برطرف کردن این مشکلات، معیاری برای محاسبهٔ میزان شباهت بین دو سند ارائه شده است که در آن، از ارتباطات معنایی نظیر اشتراک در نویسندگان و مراجع و ارتباط معنایی بین عنوان دو سند استفاده شده است. همچنین با به کارگیری این معیار در کنار ویژگی‌های متنی در یک سیستم پیشنهاد استناد نشان داده شده است که استفاده از ارتباطات معنایی، نقش مؤثری در بهبود کیفیت سیستم‌های پیشنهاد استناد دارد و ضعف ناشی از تکیهٔ تنها به ویژگی‌های متنی را کاهش می‌دهد.

در بخش بعدی مقاله، رویکردهای موجود در سیستم‌های پیشنهاد استناد توضیح داده می‌شود. بخش سه به توصیف الگوریتم پیشنهادی، و بخش چهار به ارزیابی آن اختصاص دارد. بخش پنج نیز با بیان نتیجه‌گیری و کارهای آینده، مقاله را خاتمه می‌دهد.

۲. کارهای گذشته

کارهای انجام شده در زمینهٔ سیستم‌های پیشنهاد استناد، با توجه به هدف آن‌ها به سه دسته تقسیم می‌شوند که در ادامه به طور مختصر توضیح داده می‌شوند.

هدف دستهٔ اول، کامل کردن استنادهایی است که توسط پژوهشگر برای یک متن انتخاب شده‌اند؛ برای مثال، روش پیشنهادی در [۵] با استفاده از اطلاعات موجود در گراف اسناد- گرافی که گره‌های آن، اسناد و یال‌های آن، ارتباطات بین اسناد است- یک الگوریتم فیلتر همبستگی[2] [٦] ارائه کرده و اسنادی را که با استنادهای مشخص شده توسط کاربر بیشتر مورد استناد قرار گرفته‌اند، به کاربر پیشنهاد می‌دهد. در [۷] نیز با ترکیب الگوریتم‌های فیلتر همبستگی و فیلتر محتوایی نشان داده شده است که یک الگوریتم ترکیبی، در مقایسه با الگوریتمی که فقط از یکی از این دو تکنیک استفاده می‌کند، پیشنهادهای بهتری را تولید می‌کند.

1. Citation recommendation system

2. Collaborative Filtering

هدف سیستم‌های دستهٔ دوم، پیشنهاد اسناد مرتبط با یک متن می‌باشند، به عبارتی پیشنهاد اسنادی که آن متن می‌تواند به آن‌ها استناد کند؛ برای مثال، روش پیشنهادی [۸] برای ایجاد یک سیستم پیشنهاد استناد، ابتدا بر اساس روش‌های ابتکاری، از جمله داشتن نویسندهٔ مشترک با متن ورودی و مشابه بودن از نظر کلیدواژه‌های چکیده یا عنوان، یک مجموعه کاندید از مقالات انتخاب می‌کند، سپس با استفاده از یک مدل احتمالی شباهت متنی، اسناد موجود در مجموعه کاندید را بر اساس شباهت آن‌ها با متن ورودی مرتب می‌کند.

روش پیشنهادی [۲] یک سیستم پیشنهاد استناد مبتنی بر موضوع می‌باشد که در آن، از تعلیم یک ماشین بولتزمن سه لایه استفاده شده است. سه لایه این ماشین بولتزمن عبارت‌اند از: ۱. واژه‌های موجود در مجموعه اسناد، ۲. موضوعات موجود در اسناد و ۳. لیست مراجع تمامی اسناد. سیستم پیشنهادی پس از تعلیم ماشین بولتزمن، با دریافت یک متن ورودی قادر است موضوعات موجود در آن متن را استخراج کند و سپس مراجع هر موضوع را به عنوان استنادهای متن ورودی پیشنهاد دهد.

روش پیشنهادی [۱] بر اساس ویژگی‌هایی نظیر تاریخ انتشار، اعتبار نویسنده، متن، و موضوع، معیارهای مختلفی برای شباهت تعریف می‌کند و شباهت متن ورودی و هر سند را برابر میانگین وزن‌دار هر یک از این شباهت‌ها در نظر می‌گیرد، سپس اسناد با بیشترین مقدار شباهت را به عنوان استناد پیشنهاد می‌دهد. در روش پیشنهادی این مقاله، وزن هر ویژگی به کمک یک الگوریتم تکراری محاسبه شده است. در [۹] نیز یک معیار شباهت ارائه شده است که از ترکیب خطی ویژگی‌های متنی و ویژگی‌های قابل استخراج از گراف اسناد استفاده می‌کند. در [۱۰ و ۱۱]، با در نظر گرفتن این موضوع که کلمات به کار رفته در متن استناد ممکن است متفاوت با کلمات موجود در اسنادی باشد که در آن متن مورد استناد قرار می‌گیرند، یک سیستم پیشنهاد استناد مبتنی بر یک مدل ترجمه[1] ارائه شده است.

هدف کارهایی که در دستهٔ سوم قرار می‌گیرند، پیشنهاد استناد برای قسمت مشخصی از متن ورودی می‌باشد. در سیستم پیشنهادی [۸] مکان‌های مورد نظر نویسنده برای استناد با علامت "[?]" در متن ورودی مشخص می‌شوند و سیستم قادر است برای آن مکان‌های خاص، استنادهایی پیشنهاد کند. سیستم پیشنهادی [۱۲] که ادامهٔ کار سیستم ارائه شده در [۸] می‌باشد، **Error! Unknown switch argument.** قادر است مکان استنادها را نیز خود در متن ورودی مشخص کند. در [۲] پس از پیشنهاد استناد برای متن ورودی، از یک روش بازیابی اطلاعات برای تعیین مکان استنادها استفاده شده است، با استفاده از این روش، ارتباط بین استنادها و جمله‌های موجود در متن مشخص شده است.

۳. روش پیشنهادی

در این قسمت، پس از توضیح پیش‌زمینه و انگیزهٔ روش پیشنهادی، جزئیات روش، شرح داده خواهد شد.

۳-۱. پیش‌زمینه و انگیزه

اغلب روش‌های موجود برای به دست آوردن اسناد مرتبط با یک سند از بین مجموعه اسناد موجود در یک مجموعه داده، تنها مبتنی بر ویژگی‌های متنی اسناد می‌باشند. ضعف این روش‌ها در دو مورد زیر کاملاً محسوس است:

- دو سند مرتبط لزوماً، دو سند با شباهت متنی زیاد نمی‌باشند، مثلاً سندی که دربارهٔ زمانبندی پردازش‌ها به کمک الگوریتم ژنتیک بحث می‌کند، ممکن است به سندی که ایدهٔ کلی الگوریتم ژنتیک را توضیح داده است، شباهت متنی کمی داشته باشد، اما این دو سند کاملاً به هم مرتبط‌اند، زیرا سند دوم، پایهٔ علمی تکنیک مورد استفاده در سند اول را توضیح می‌دهد. در نتیجه استفادهٔ صرف از شباهت متنی باعث می‌شود مرتبط شناخته نشوند.
- موضوع دو سند ممکن است دقیقاً یکسان باشد، اما از آنجا که توسط دو نویسندهٔ مختلف نوشته شده‌اند و

1. Translation Model

کلمات مورد استفاده این دو نویسنده متفاوت است، شباهت متنی دو سند کم است و استفادهٔ صرف از شباهت متنی قادر به تشخیص ارتباط این اسناد نیست.

از آنجا که هر سند، علاوه بر ویژگی‌های متنی نظیر عنوان و چکیده، دارای ارتباطات معنایی با دیگر اسناد است، در محاسبهٔ شباهت دو سند باید از این ویژگی‌ها و ارتباطات معنایی در کنار هم استفاده شود.

دیدگاه مقالهٔ حاضر این است که مجموعه داده‌های مورد استفادهٔ سیستم، شامل اطلاعات N سند $P_i(1\leq i\leq N)$ به صورت زیر است:

$P_i= (Id_i, title_i, abs_i,, refList_i, citList_i, authList_i, venue_i, year_i)$

Id_i: شناسهٔ سند P_i که یک شماره منحصربه‌فرد است.

$title_i$: عنوان سند P_i

abs_i: چکیدهٔ سند P_i

$refList_i$: لیست مراجع P_i

$citList_i$: لیست اسنادی که به P_i استناد کرده‌اند.

$authList_i$: لیست نویسندگان P_i

$venue_i$: کنفرانس یا ژورنالی که P_i در آن ارائه شده است.

$year_i$: سال انتشار P_i

در روش پیشنهادی، ۷ رابطه با نام‌های $R_1, R_2, \ldots, R_7$ از سند P_i به سند P_j به شرح زیر تعریف شده است:

$R_1 : P_i \in citList_j$

$R_2 : P_i \in refList_j$

$R_3 : authList_i \bigcap authList_j \neq \varphi$

$R_4 : venue_i = venue_j$

$R_5 : refList_i \bigcap refList_j \neq \varphi$

$R_6 : citList_i \bigcap citList_j \neq \varphi$

$R_7 : title_i$ is similar to $title_j$

علت در نظر گرفتن هر یک از این روابط در ادامه توضیح داده شده است:

وقتی نویسندهٔ یک سند، در سند خود به دیگری استناد می‌کند، به معنی این است که این دو سند با هم مرتبط‌اند، این مفهوم در رابطهٔ R_1 و R_2 نشان داده شده است. به علاوه اغلب اسناد نوشته شده توسط یک نویسنده، و همچنین اسناد ارائه شده در یک کنفرانس یا ژورنال مربوط به یک زمینه‌اند و به نوعی به هم مرتبط‌اند. رابطه‌های R_3 و R_4 با همین دیدگاه در نظر گرفته شده‌اند.

رابطه‌های R_5 و R_6 نیز به ترتیب بر اساس دو مفهوم اصلی در زمینهٔ تحلیل استناد[1] [۱۳]، به نام‌های زوج‌های کتاب‌شناختی[2] [۱۴] و استناد مشترک[3] [۱۵] تعریف شده‌اند. زوج‌های کتاب‌شناختی مبتنی بر این ایده‌اند که سندهایی که در موضوع دارای شباهت هستند، مراجع مشترک دارند. همچنین مفهوم استناد مشترک این است که سندهایی که دارای شباهت‌اند، به احتمال زیاد توسط یک سند مشترک مورد استناد قرار می‌گیرند.

با توجه به اینکه عنوان یک سند شامل واژه‌های کلیدی موضوع آن سند است. اگر عنوان‌های دو سند شباهت معنایی بیشتری با یکدیگر داشته باشند، طبیعی است که آن دو سند مرتبط در نظر گرفته شوند. رابطه R_7 بر اساس این مفهوم تعریف شده است.

انگیزهٔ روش پیشنهادی این مقاله، استفاده از ارتباطات معنایی در کنار ویژگی‌های متنی، برای برطرف کردن ضعف‌های ویژگی‌های متنی در پیدا کردن اسناد مرتبط است.

در ادامه، ابتدا یک معیار جدید برای شباهت معنایی دو سند پیشنهاد شده است، و سپس یک الگوریتم پیشنهاد استناد ارائه شده است که از این معیار استفاده می‌کند.

۳-۲. شباهت معنایی

با توجه به ۷ نوع رابطهٔ تعریف شده، شباهت معنایی بین هر دو سند P_i و P_j به کمک فرمول (۱) تعریف می‌شود:

1. Citation analysis
2. Bibliographic coupling
3. Co-citation

$$similarity(P_i,P_j)=\frac{1}{7}\sum_{k=1}^{7}W_k F_k\left(P_i,P_j\right) \qquad (۱)$$

در فرمول بالا، F_k یک تابع است که میزان شباهت دو سند P_i و P_j را با توجه به رابطهٔ R_k به دست می‌آورد و W_k نشان‌دهندهٔ میزان تأثیر هر رابطه در شباهت معنایی دو سند است. تعریف این ۷ تابع در زیر آورده شده است.

مقدار بازگشتی $F_1(P_i , P_j)$، در صورت وجود رابطهٔ R_1 از سند P_i به سند P_j، برابر است با تعداد دفعاتی که در سند P_i به سند P_j استناد شده، تقسیم بر کل استنادهای موجود در سند P_i. در صورت عدم وجود رابطهٔ R_1 مقدار بازگشتی برابر صفر است.

مقدار بازگشتی $F_2(P_i , P_j)$، در صورت وجود رابطهٔ R_2 از سند P_i به سند P_j، برابر است با تعداد دفعاتی که در سند P_j به سند P_i استناد شده، تقسیم بر کل دفعاتی که سند P_i، توسط اسناد دیگر مورد استناد قرار گرفته است. در صورت عدم وجود رابطهٔ R_2 مقدار بازگشتی برابر صفر است.

مقدار بازگشتی $F_4(P_i , P_j)$، در صورت وجود رابطهٔ R_4 از سند P_i به سند P_j، برابر یک، و در غیر این صورت برابر صفر است.

مقدار بازگشتی توابع F_3، F_5، F_6 و F_7 به ترتیب به کمک فرمول‌های (۲)، (۳)، (٤) و (۵) به دست می‌آید:

$$F_3(P_i,P_j)=\frac{\left|authList_i \cap authList_j\right|}{\left|authList_i \cup authList_j\right|} \qquad (۲)$$

$$F_5(P_i,P_j)=\frac{\left|refList_i \cap refList_j\right|}{\left|refList_i \cup refList_j\right|} \qquad (۳)$$

$$F_6(P_i,P_j)=\frac{\left|citList_i \cap citList_j\right|}{\left|citList_i \cup citList_j\right|} \qquad (٤)$$

$$F_7(P_i,P_j)=semSim(title_i,title_j) \qquad (۵)$$

در هر یک از توابع بالا، مقدار بازگشتی صفر، به معنی ارتباط نداشتن و مقدار بازگشتی یک، نشان‌دهندهٔ یک ارتباط قوی از نظر رابطهٔ متناظر با آن است.

مقدار بازگشتی برای توابع F_1 و F_2 که به ترتیب به رابطه‌های R_1 و R_2 مرتبط‌اند، مبتنی بر ایدهٔ مشابهی است؛ برای مثال، در مورد R_1، هرچه در سند P_i به تعداد بیشتری به سند P_j استناد شده باشد، نشانهٔ آن است که سند P_i به سند P_j مرتبط‌تر است؛ مثلاً اگر سند P_i پنج مرتبه به سند P_j، و دو مرتبه به سند P_k استناد کرده باشد، منطقی است که ارتباط P_i به P_j را قوی‌تر از ارتباط P_i به P_k بدانیم. از طرفی، تعداد کل استنادهای موجود در سند P_i نیز مهم است؛ مثلاً اگر دو سند P_i و P_j هر یک پنج مرتبه به سند P_k استناد کرده باشند، اما در سند P_i در مجموع ۲۵ استناد، و در سند P_j در مجموع ۵۰ استناد موجود باشد، آنگاه می‌توان گفت که ارتباط P_i و P_k، قوی‌تر از ارتباط P_j و P_k می‌باشد. مقدار بازگشتی تابع F_1، بر اساس این دیدگاه محاسبه می‌شود.

مقدار بازگشتی برای توابع F_3، F_5 و F_6 نیز، مبتنی بر ایدهٔ مشابهی است؛ مثلاً در مورد F_3، هر چه تعداد نویسندگان مشترک دو سند بیشتر باشد، میزان ارتباط آن‌ها بیشتر در نظر گرفته می‌شود. البته نسبت نویسندگان مشترک به کل نویسندگان این دو سند نیز اهمیت دارد؛ برای مثال، دو نویسندهٔ مشترک از سه نویسنده، نسبت به دو نویسندهٔ مشترک از بین شش نویسنده، بیانگر رابطهٔ قوی‌تری می‌باشد؛ بنابراین، F_3 با تعداد نویسندگان مشترک، رابطهٔ مستقیم، و با تعداد کل نویسندگان، رابطهٔ عکس دارد.

مقدار بازگشتی برای تابع F_7 برابر شباهت معنایی عنوان سند P_i و عنوان سند P_j می‌باشد که از فرمول (٦) محاسبه می‌شود:

(٦)

$$semSim(title_i,title_j)=\frac{\sum_{x_m\in W_i}\sum_{y_n\in W_j} wordnetSim(x_m,y_n)}{|W_i|.|W_j|}$$

در این رابطه، W_i مجموعه واژگان موجود در عنوان سند P_i و W_j مجموعه واژگان موجود در عنوان سند P_j است. همچنین $wordnetSim(x, y)$ برابر شباهت معنایی دو واژه x و y بر اساس شبکه واژگان [١٦] $WordNet$[16] می‌باشد که بر اساس ترکیبی از الگوریتم‌های [١٧] Lin[17] و [١٨] Wu-$Palmer$[18] محاسبه می‌شود. در [١٩] نشان داده شده است

که استفاده از این شبکه واژگان می‌تواند باعث کاهش مشکل ابهام در شباهت متنی شود.

۳-۳. الگوریتم پیشنهاد استناد

از آنجا که ورودی الگوریتم پیشنهادی تنها از متن تشکیل شده و ویژگی‌های دیگری مثل لیست مراجع یا لیست نویسندگان ندارد، امکان استفادهٔ مستقیم از شباهت معنایی وجود ندارد. در نتیجه این الگوریتم ابتدا با به دست آوردن شباهت متنی هر سند موجود در مجموعه داده‌های محلی P_i و متن ورودی *input*، یعنی $txtSim(P_i, input)$، تعداد ثابتی، C، از اسنادی را که دارای بیشترین شباهت متنی می‌باشند، به عنوان مجموعه کاندید انتخاب می‌کند.

مجموعه کاندید که با توجه به ویژگی‌های متنی به دست آمده، شامل C سند از مجموعه داده‌های محلی است. هر یک از این اسناد علاوه بر ویژگی‌های متنی شامل ویژگی‌های دیگری از جمله لیست نویسندگان و لیست مراجع‌اند. در نتیجه برای پیشنهاد لیستی از اسناد موجود در مجموعه داده‌های محلی به عنوان استناد برای متن ورودی، میزان شباهت هر سند با متن ورودی به کمک فرمول (۷) به دست می‌آید.

$$score(P_i, input) = \sum_{j=1}^{C} w_j \times similarity(P_i, Q_j) \qquad (۷)$$

در فرمول بالا، Q_j نشان‌دهندهٔ j-امین سند موجود در مجموعه کاندید، C نشان‌دهندهٔ تعداد عناصر این مجموعه، و w_j که از طریق فرمول (۸) به دست می‌آید، به عنوان وزن شباهت معنایی P_i و Q_j در شباهت کلی سند P_i به متن ورودی می‌باشد.

$$w_j = \frac{txtSim(input, Q_j)}{\max_{(1 \le k \le C)} txtSim(input, Q_k)} \qquad (۸)$$

پس از به دست آوردن مقدار تابع *score* برای هر سند، با مرتب کردن آن‌ها به صورت نزولی، M سند که بیشترین مقدار را دارند، به عنوان استناد برای متن ورودی پیشنهاد داده می‌شوند.

٤. ارزیابی

سیستم پیشنهادی به زبان برنامه‌نویسی *Java* پیاده‌سازی شده و مورد ارزیابی عملی قرار گرفته است. در این قسمت، ابتدا مجموعه داده انتخابی، و سپس نحوهٔ ارزیابی، مقادیر پارامترهای آزمایش و نتایج آزمایش‌ها ارائه شده است.

٤-١. مجموعه داده

برای ایجاد مجموعه داده‌های لازم برای آزمایش سیستم پیشنهادی، یک پیمایشگر پیاده‌سازی شده است که با استفاده از سرویس *OAI*[1] ارائه شده توسط سیستم *CiteSeerX*، اطلاعات اسناد منتشر شده در این سیستم را جمع‌آوری می‌کند. بعد از پالایش داده‌های به دست آمده و حذف اسنادی که مقدار زیادی از اطلاعات آن‌ها غیر معتبر بود، اسنادی که سال انتشار آن‌ها بعد از سال ۲۰۰۷ بود، به عنوان داده‌های ورودی و بقیه اسناد به عنوان مجموعه دادهٔ محلی انتخاب شدند و در پایگاه داده *MySQL* ذخیره شدند. در نهایت، مجموعه داده‌ای شامل ۱۳۰۰۰ سند آماده شد که ٦۰۰ سند از این مجموعه به عنوان مجموعه دادهٔ ورودی برای آزمایش انتخاب شد.

٤-٢. روش و معیارهای ارزیابی

روش ارزیابی سیستم پیشنهادی یک ارزیابی خودکار است. در این روش، یک سند از مجموعه اسناد ورودی انتخاب شده و متن آن به عنوان دادهٔ ورودی به سیستم داده می‌شود تا برای آن استنادهایی پیشنهاد شود. همچنین لیست مراجع این سند به عنوان خروجی مورد انتظار در نظر گرفته می‌شود. بدیهی است هرچه لیست استنادهای پیشنهادی الگوریتم با لیست مراجع این سند مطابقت بیشتری داشته باشد، الگوریتم پیشنهاد موفق‌تر است.

معیارهای در نظر گرفته شده جهت سنجش کارآیی این سیستم در ادامه توضیح داده شده است.

فراخوانی:

1. http://citeseerx.ist.psu.edu/oai.html

برای هر سند ورودی، با مقایسهٔ لیست مراجع آن با لیست استنادهای پیشنهاد شده، فراخوانی الگوریتم برای آن سند محاسبه می‌شود. برای به دست آوردن فراخوانی کل سیستم، میانگین فراخوانی‌ها برای کلیهٔ اسناد موجود در مجموعه داده‌های ورودی محاسبه می‌شود.

احتمال استناد مشترک[1]:

برای هر سند ورودی، ممکن است برخی از استنادهای پیشنهاد داده شده توسط الگوریتم، در لیست مراجع آن سند وجود نداشته باشد. چنین پیشنهادهایی لزوماً نامناسب نیستند، بلکه ممکن است استنادهایی قابل قبول و یا حتی بهتر از لیست مراجع آن سند باشند. در اغلب کارهای مرتبط، برای ارزیابی چنین پیشنهادهایی، از ارزیابی مبتنی بر متخصص استفاده شده است، یعنی از تعدادی متخصص خواسته شده است که هر یک از چنین استنادهایی را بررسی و مشخص کنند آیا به عنوان یک پیشنهاد مناسب، قابل قبول است یا خیر.

در [۸ و ۱۲] یک معیار برای ارزیابی خودکار این پیشنهادها، به نام احتمال استناد مشترک پیشنهاد شده است. در این معیار به ازای هر سند از لیست استنادهای پیشنهاد شده که متعلق به لیست مراجع سند ورودی نیست، احتمال اینکه آن سند به همراه هر یک از اسناد لیست مراجع، به طور مشترک مورد استناد قرار بگیرد، محاسبه شده و سپس میانگین آن برای همهٔ اسناد لیست مراجع به دست می‌آید.

میزان احتمال استناد مشترک کل سیستم نیز برابر میانگین مقدار احتمال استناد مشترک برای همه اسناد موجود در مجموعه ورودی در نظر گرفته شده است.

NDCG[2]:

مفید بودن سیستم‌های پیشنهاددهنده نه تنها به عناصر پیشنهاد شده، بلکه به ترتیب این عناصر وابسته است. این وابستگی با معیارهایی مثل احتمال استناد مشترک و فراخوانی قابل ارزیابی نیست. مسلم است که چنانچه استنادهایی که بیشتر مرتبط‌اند، در اوایل لیست استنادهای پیشنهاد شده قرار بگیرند، بهتر است. معیار $NDCG$ که یک معیار شناخته شده در بازیابی اطلاعات است، از این منظر به اندازه‌گیری کیفیت لیست پیشنهادها می‌پردازد [۸ و ۱۲].

٤-۳. پارامترهای آزمایش

همان‌طور که در بخش ۳-۲ اشاره شد، برای به دست آوردن شباهت معنایی بین دو سند از فرمول (۱) استفاده می‌شود. در این فرمول، از متغیر W_k استفاده شده است که نشان‌دهندهٔ میزان تأثیر تابع متناظر با هر رابطه R_k در شباهت معنایی بین دو سند می‌باشد. مقادیر در نظر گرفته شده برای W_k در آزمایشات، در جدول (۱) آورده شده است.

جدول (۱): مقادیر در نظر گرفته شده برای W_k

مقدار	W_k
۰.۸	W_1
۰.۲	W_2
۰.٤	W_3
۰.۲	W_4
۰.۵	W_5
۰.۷	W_6
۰.٦	W_7

علت در نظر گرفتن هر یک از مقادیر W_k در ادامه توضیح داده شده است:

وزن F_1 از بقیه بیشتر است و مقدار نسبتاً بزرگی در نظر گرفته شده است، زیرا اسناد موجود در مجموعه کاندید با متن ورودی شباهت متنی زیادی دارند و در نتیجه اگر اکثر این اسناد به سند خاصی استناد کنند، به احتمال زیاد متن ورودی نیز باید به آن سند استناد کند.

وزن اختصاص داده شده به تابع F_2 مقدار کوچکی دارد. توجیه آن بدین ترتیب است که اگر سندی به بیشتر اسناد موجود در مجموعه کاندید، که همگی از نظر متنی شبیه متن ورودی هستند، استناد کند، بین آن سند و متن ورودی رابطه‌ای وجود دارد، اما لزوماً نمی‌توان گفت که متن ورودی باید به آن

1. Cocited-probability
2. Normalized Discounted Cumulative Gain

سند استناد کند. طبیعی است که وزن F_2 باید از وزن F_1 کمتر باشد.

وزن در نظر گرفته شده برای F_4 برابر ۰/۲ است. این مقدار مؤید این دیدگاه است که اگرچه معمولاً هر کنفرانس یا ژورنالی بر روی یک موضوع متمرکز است، معمولاً هر موضوعی شامل تعداد زیادی زیرموضوع مرتبط می‌باشد. در نتیجه دو سند که در یک کنفرانس یا ژورنال ارائه می‌شوند. اگرچه زمینهٔ کلی‌شان یکسان است، اما لزوماً زیر موضوع یکسانی ندارند. در نتیجه سهم R_4 در شباهت کلی بین دو سند کم می‌باشد.

در مورد وزن F_3 می‌توان گفت اگرچه اسناد نوشته شده توسط یک نویسنده اغلب در یک زمینه است، ولی این موضوع در همهٔ حالات صدق نمی‌کند و ممکن است یک نویسنده در زمینه‌های مختلفی کار کند؛ بنابراین، وزن این ارتباط در شباهت معنایی نباید زیاد باشد. البته منطقی است که وزن F_3 از وزن F_4 بیشتر باشد، زیرا احتمال شباهت دو سند یک نویسنده، به مراتب بیشتر از احتمال شباهت دو سند ارائه شده در یک کنفرانس یا ژورنال است.

وزن F_6 نسبتاً زیاد است، زیرا اگر یک سند P وجود داشته باشد که به دفعات زیادی، هم‌زمان با اسناد مجموعه کاندید مورد استناد قرار گرفته باشد، یعنی بیشتر اسنادی که به اسناد درون مجموعه کاندید اولیه استناد کرده‌اند، به سند P نیز استناد کرده باشند، آنگاه با توجه به اینکه اسناد درون مجموعه کاندید به متن ورودی شباهت متنی زیادی دارند، به احتمال زیاد می‌توان گفت که متن ورودی باید به سند P استناد کند.

وزن F_5 از وزن F_6 کمتر است و از مقدار متوسطی برخوردار است. دلیل این امر را می‌توان این‌گونه بیان کرد. اگر یک سند P وجود داشته باشد که با اسناد مجموعه کاندید، تعداد زیادی مرجع مشترک داشته باشد، این بدان معناست که سند P به اسناد مجموعه کاندید شباهت زیادی دارد. به واسطهٔ این شباهت تا حدی می‌توان گفت که متن ورودی باید به سند P استناد کند؛ البته واضح است که میزان این امر در مقایسه با F_6 کمتر است.

وزن F_7 با توجه به اهمیت عنوان یک سند که شامل واژه‌های کلیدی موجود در آن سند است و همچنین با توجه به مزیت استفاده از شباهت معنایی در مقایسه با شباهت متنی، در نظر گرفته شده است؛ البته واضح است که با توجه به محدود بودن تعداد واژه‌های یک عنوان، استفاده از این رابطه نمی‌تواند به اندازهٔ F_1 و F_6 اثرگذار باشد.

با توجه به بخش ۳ـ۳، در مرحلهٔ تولید مجموعه کاندید از شباهت متنی استفاده شده است و تعداد ثابت C سند به عنوان مجموعه کاندید انتخاب شده است. در آزمایش‌های انجام شده، شباهت متنی از طریق معیار $TFIDF$ محاسبه می‌شود و با توجه به بررسی انجام شده و برقراری تعادل بین زمان پیشنهادها و کارآیی آن‌ها، مقدار $C=25$ انتخاب شده است.

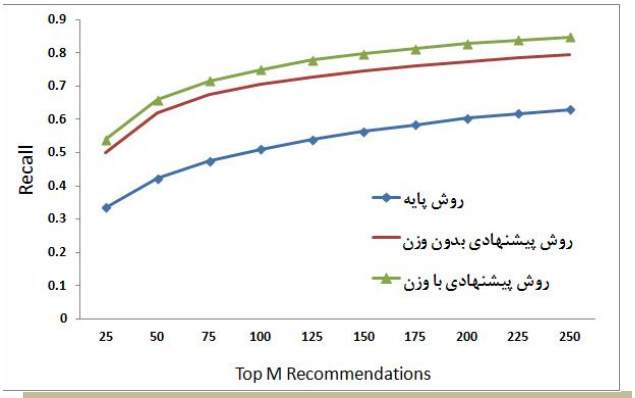
در مرحلهٔ انتخاب پیشنهادهای مرتب شده نیز مقدار M در آزمایش‌ها، در بازهٔ [۲۵, ۲۵۰] انتخاب شده است.

٤-٤. نتایج

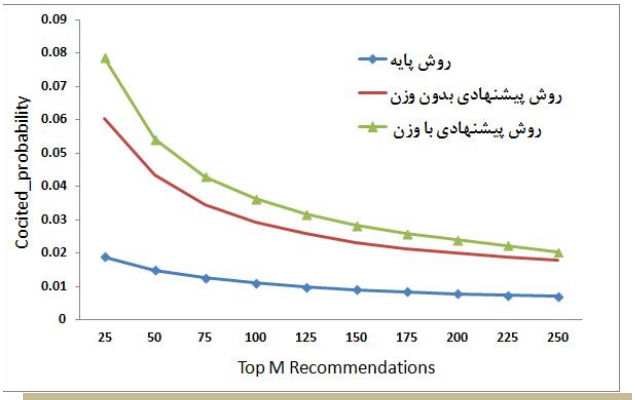
به منظور ارزیابی روش پیشنهادی و نشان دادن اهمیت استفاده از ارتباطات معنایی علاوه بر ویژگی‌های متنی در به دست آوردن اسناد مرتبط، یک روش پایه، برای مقایسه با روش پیشنهادی در نظر گرفته شده است. در این روش، تنها از ویژگی‌های متنی برای پیشنهاد اسناد مرتبط با یک متن استفاده شده است.

همچنین برای ارزیابی اهمیت در نظر گرفتن وزن‌های مختلف برای رابطه‌های استفاده شده در معیار شباهت معنایی، روش پیشنهادی در دو حالت (با وزن‌های نشان داده شده در جدول (۱) و بدون وزن یا به عبارتی با وزن یک برای تمام رابطه‌ها) اجرا شده است.

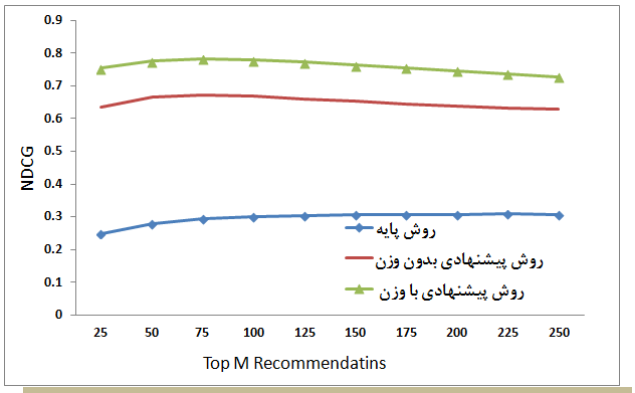
نتایج ارزیابی بر اساس ۳ معیار مورد نظر، در شکل (۱) تا شکل (۳) نمایش داده شده است.



شکل (۱): نتایج ارزیابی از نظرمعیار فراخوانی



شکل (۲): نتایج ارزیابی از نظر معیار احتمال استناد مشترک



شکل (۳): نتایج ارزیابی از نظر معیار NDCG

همان‌طور که نتایج ارزیابی نشان می‌دهد، استفاده از معیار ارائه شده برای شباهت معنایی، تأثیر قابل ملاحظه‌ای در بهبود نتایج از نظر هر یک از معیارهای ارزیابی داشته است. در نتیجه ثابت می‌شود که ایدهٔ استفاده از ارتباطات معنایی در کنار ویژگی‌های متنی، برای برطرف کردن ضعف‌های ویژگی‌های متنی در پیدا کردن اسناد مرتبط و بهبود سیستم‌های پیشنهاد استناد که انگیزهٔ راه حل پیشنهادی این مقاله نیز بوده است، کاملاً درست است.

همچنین با توجه به نتایج آزمایش‌ها، در نظر گرفتن وزن برای رابطه‌های استفاده شده در شباهت معنایی، باعث بهبود نتایج از نظر هر سه معیار ارزیابی شده است. در نتیجه، تحلیل-های در نظر گرفته شده در بخش **Error! Unknown switch argument.** برای مشخص کردن میزان تأثیر هر رابطه در شباهت معنایی دو سند، صحیح است.

۵. نتیجه‌گیری و کارهای آینده

در این مقاله، یک سیستم پیشنهاد استناد ارائه شده است که می‌تواند پژوهشگران را در پیدا کردن کارهای مرتبط با یک زمینهٔ تحقیقاتی کمک کند. این سیستم با دریافت متن ورودی، لیستی از اسنادی را که آن متن باید به آن‌ها استناد کند، پیشنهاد می‌کند. این پیشنهادها بر اساس یک معیار شباهت معنایی جدید که در این مقاله معرفی شده، تولید می‌شوند. آزمایش تجربی نشان می‌دهد که استفاده از این معیار در کنار معیارهای شباهت متنی، با بهبود کیفیت پیشنهادها، کارآیی سیستم را افزایش می‌دهد.

چند نمونه از کارهایی که می‌توان در آینده به منظور بهبود کیفیت پیشنهادها انجام داد، عبارت‌اند از: استفاده از روش‌های پردازش تکاملی مانند الگوریتم ژنتیک به منظور به دست آوردن مقادیر بهینه برای وزن‌های استفاده شده در شباهت معنایی، ارزیابی دستی روش ارائه شده با استفاده از نظر افراد خبره و افزودن قابلیت پیشنهاد استناد برای نقاط خاصی از متن ورودی.

مراجع

[1] Bethard, S., Dan Jurafsky, D.,"*Who should I cite? Learning literature search models from citation behavior*", ACM Conference on Information and Knowledge Management, pp. 609-618, 2010.

[2] Tang, J., Zhang, J.,"*A Discriminative Approach to Topic-Based Citation Recommendation*", Proceedings of PAKDD, pp. 572-579, 2009.

[3] Henzinger, M.R., Motwani, R. and Silverstein, C., "*Challenges in web search engines*", Proceedings of the

international joint conference on artificial intelligence, Vol. 18, pp. 1573-1579, 2003.

[4] Vallez, M. and Pedraza-Jimenez, R., "*Natural Language Processing in Textual Information Retrieval and Related Topics*", Available Online. www.hipertext.net, 2007.

[5] McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Rashid, A., Konstan, J., Ried, J., "*On the Recommending of Citations for Research Papers*", Proceedings of the 2002 ACM conference on Computer supported cooperative work New York, NY, USA, pp. 116-125, 2002.

[6] Adomavicius, G., Tuzhilin, A., "*Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*", Journal of IEEE Trans. Knowl. Data Eng. Vol. 17, no 6, pp. 734-749, 2005.

[7] Torres, R., McNee, S. M., Abel, M., Konstan, J.A., Riedl, J.,"*Enhancing digital libraries with TechLens*", Proceedings of IEEE/ACM Joint Conference on Digital Libraries (ACM/IEEE JCDL'2004), Tuscon, AZ, USA, pp. 228-236, 2004.

[8] He, Q., Pei, J., Kifer, D., Mitra, P., Giles, C.L., "*Context-aware Citation Recommendation*", Proceedings of the International World Wide Web Conference (WWW), Vol. 19, pp. 421-430, 2010.

[9] Strohman, T., Croft, W. B., Jensen, D., "*Recommending citations for academic papers*", Proceedings of Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Vol. 30, pp. 705-706, 2007.

[10] He, J., J-Y. Nie, Y. Lu, W. Xin Zhao, "*Position-Aligned Translation Model for Citation Recommendation*". Proceedings ofSPIRE 2012, pp. 251-263, 2012.

[11] Lu, Y., He, J., Shan, D., Yan, H., "*Recommending citations with translation model",* Proceedings of CIKM 2011, pp. 2017–2020, 2011.

[12] He, Q., Kifer, D., Pei, J., Mitra,P., Giles, C.L., "*Citation recommendation without author supervision*", Proceedings of WSDM'11, pp. 755-764, 2011.

[13] Smith, L.C., "*Citation analysis*", journal of Library Trends, Vol. 30, No. 1, pp. 83-106, 1981.

[14] Kessler, M., "*Bibliographic coupling between scientific papers*", Journal of American Documentation, Vol. 14, No. 1, pp. 10-25, 1963.

[15] Small, H.,"*Co-citation in the scientific literature: A new measurement of the relationship between two documents*", Journal of the American Society of Information Science, Vol. 24, No. 4, pp. 265-269,1973.

[16] Miller, G., "*Wordnet: A Lexical Database for English*", Commun. ACM 38, pp. 39-41, 1995.

[17] Lin, D., "*An information-theoretic definition of similarity*", Proceedings of Fifteenth International Conference on Machine Learning, pp. 296–304, 1998.

[18] Zhibiao, W., Palmer, M., "*Verb semantics and lexical selection*", Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133–138, 1994.

[19] Voorhees, E., "*Using WordNet to Disambiguate Word Senses for Text Retrieval*", Proceedings of SIGIR, pp. 171-180, 1993.