یک سیستم جدید پیشنهاد استناد مبتنی بر ویژگی‌های رابطه‌ای اسناد

فتانه زرین کلام[1]، محسن کاهانی[2]

آزمایشگاه فناوری وب، گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد

zarrinkalam.fattane@stu-mail.um.ac.ir[1]

kahani@um.ac.ir[2]

چکیده

حجم فراوان و روبه رشد اسناد علمی منتشر شده بر روی وب، فرآیند تصمیم‌گیری و انتخاب اسناد مرتبط با یک زمینه تحقیقاتی را برای پژوهشگران دشوار کرده است. یک سیستم پیشنهاد استناد، با دریافت متن ورودی، اسنادی که باید توسط آن متن مورد استناد قرار گیرند را پیشنهاد می‌کند، و بدین ترتیب می‌تواند در یافتن اسناد مرتبط با یک موضوع به پژوهشگر کمک کند. در این مقاله، یک معیار جدید برای محاسبه شباهت دو سند ارائه شده است که مبتنی بر ویژگی‌های رابطه‌ای اسناد می‌باشد. همچنین یک سیستم پیشنهاد استناد ارائه شده است که از معیار فوق در کنار ویژگی‌های متنی اسناد استفاده می‌کند. ارزیابی انجام شده نشان می‌دهد که معیار مورد نظر، در تشخیص شباهت اسناد موفق است و باعث بهبود کیفیت سیستم پیشنهاد استناد مورد نظر می‌شود.

کلمات کلیدی

پیشنهاد استناد، شباهت رابطه‌ای، شباهت متنی.

۱- مقدمه

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 [1,2].

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. خروجی موتورهای جستجو برای این کلمات کلیدی تعداد زیادی از اسنادی هستند که شامل این کلمات کلیدی می‌باشند، در نتیجه پژوهشگر باید زمان زیادی را صرف انتخاب گزینه‌های بهتر از بین خروجی‌ها کند. ضمنا از آنجایی‌که هدف بدست آوردن اسناد مرتبط است و نه صرفا اسنادی که شباهت متنی زیادی با متن ورودی دارند، ممکن است تعدادی از کارهای مرتبط در بین جواب‌های جستجو ظاهر نشوند [3].

بعنوان مثال اگر موضوع متن ورودی زمانبندی پردازش‌ها به کمک الگوریتم ژنتیک باشد، واضح است که سندی که به معرفی الگوریتم ژنتیک پرداخته است با این متن مرتبط است اما لزوماً شباهت متنی زیادی با این متن نخواهد داشت، استفاده از روش‌هایی که تنها با استفاده از شباهت‌های متنی به پیدا کردن اسناد مرتبط می‌پردازند در پیدا کردن چنین اسنادی دارای ضعف هستند.

راه‌حل 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 [1-7].

کارهایی که در زمینه سیستم‌های پیشنهاد استناد انجام شده است با توجه به هدفشان می‌توانند در سه دسته قرار گیرند، هدف دسته اول کامل کردن لیست استنادهایی است که توسط پژوهشگر برای یک متن انتخاب شده است. دسته دوم، با گرفتن یک متن، اسنادی که باید در آن متن مورد استناد قرار گیرند را پیشنهاد می-

دهند و هدف دسته سوم پیشنهاد استناد برای مکان خاصی از متن ورودی می‌باشد.

روش پیشنهادی این مقاله که در دسته دوم قرار می‌گیرد، ابتدا با استفاده از ویژگی‌های رابطه‌ای یک سند روشی برای بدست آوردن شباهت رابطه‌ای بین دو سند ارائه می‌دهد و سپس با ارائه الگوریتمی مبتنی بر ترکیب آن با ویژگی‌های متنی، نشان می‌دهد که استفاده از شباهت رابطه‌ای، نقش موثری در بهبود کیفیت سیستم‌های پیشنهاد استناد دارد و ضعف ناشی از تکیه تنها به ویژگی‌های متن 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، مقاله را خاتمه می‌دهد.

۲- کارهای گذشته

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

هدف دسته اول کامل کردن استنادهایی است که توسط پژوهشگر برای یک متن انتخاب شده‌اند. برای مثال روش پیشنهادی در [5] با استفاده از اطلاعات موجود در گراف اسناد- گرافی که گره‌های آن، اسناد و یال‌های آن، ارتباطات بین اسناد می‌باشد- یک الگوریتم فیلتر همبستگی[۲][1] ارائه کرده و اسنادی که با استنادهای مشخص شده توسط کاربر بیشتر مورد استناد قرار گرفته‌اند را به کاربر پیشنهاد می‌دهد. در [6] 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.

هدف سیستم‌های دسته دوم، پیشنهاد اسناد مرتبط با یک متن می‌باشند، به عبارتی پیشنهاد اسنادی که آن متن می‌تواند به آن‌ها استناد کند. برای مثال، روش پیشنهادی [7] برای ایجاد یک سیستم پیشنهاد استناد، ابتدا بر اساس روش‌های ابتکاری، از جمله داشتن نویسنده مشترک با متن ورودی، 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 [2] 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 عبارتند از: (۱) کلمات 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 کرده و سپس مراجع هر موضوع را بعنوان استنادهای متن ورودی پیشنهاد دهد.

روش پیشنهادی [1] بر اساس ویژگی‌هایی نظیر تاریخ انتشار، اعتبار نویسنده، متن، و موضوع، معیارهای مختلفی برای شباهت تعریف کرده 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 بعنوان استناد پیشنهاد می‌دهد. در روش پیشنهادی این مقاله، وزن هر ویژگی به کمک یک الگوریتم تکراری محاسبه شده است. در [4] نیز یک معیار شباهت ارائه شده است که از ترکیب خطی ویژگی‌های متنی و ویژگی‌های قابل استخراج از گراف اسناد استفاده می‌کند.

هدف کارهایی که در دسته سوم قرار می‌گیرند پیشنهاد استناد برای قسمت مشخصی از متن ورودی می‌باشد. در سیستم پیشنهادی [7] 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 [8] که ادامه کار سیستم ارائه شده در [7] می‌باشد، قادر است مکان استنادها را نیز خود در متن ورودی مشخص کند. در [2] 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 انگیزه روش پیشنهادی، جزئیات روش، شرح داده خواهد شد.

۳-۱- پیش‌زمینه و انگیزه

اغلب روش‌های موجود برای بدست 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 کاملا محسوس است:

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

استفاده صرف از شباهت متنی قادر به تشخیص ارتباط این اسناد نیست.

دیدگاه مقاله حاضر این است که مجموعه داده‌های مورد استفاده سیستم، شامل اطلاعات N سند P_i $(1 \le i \le N)$ به صورت زیر می‌باشد:

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

Id_i: شناسه سند P_i، که یک شماره منحصربفرد است

$text_i$: متن P_i، که شامل عنوان و چکیده می‌باشد

$refList_i$: 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

از آنجایی‌که هر سند علاوه بر ویژگی‌های متنی نظیر عنوان و چکیده، دارای ویژگی‌های غیرمتنی، که به آنها ویژگی‌های رابطه‌ای می‌گوییم، نیز می‌باشد، می‌توان از آن ویژگی‌ها برای پیدا کردن اسناد مرتبط استفاده کرد.

در روش پیشنهادی، ۶ رابطه با نام‌های $R_1, R_2, \ldots, R_6$ از سند P_i به سند P_j بشرح زیر تعریف شده است:

$$R_1 : P_i \in citList_j$$
$$R_2 : P_i \in refList_j$$
$$R_3 : authList_i \cap authList_j \neq \phi$$
$$R_4 : venue_i = venue_j$$
$$R_5 : refList_i \cap refList_j \neq \phi$$
$$R_6 : citList_i \cap citList_j \neq \phi$$

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

وقتی نویسنده یک سند، در سند خود به دیگری استناد می‌کند به معنی این است که این دو سند با هم مرتبط هستند، این مفهوم در رابطه 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 زمینه می‌باشند و بنوعی بهم مرتبط هستند. 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 زمینه تحلیل استناد[۳][9] ، به نام‌های زوج‌های کتاب‌شناختی[۴][10] و استناد مشترک[۵] [11] تعریف شده‌اند. زوج‌های کتاب‌شناختی مبتنی بر این ایده هستند که سندهایی که در موضوع دارای شباهت هستند مراجع مشترک دارند. همچنین مفهوم استناد مشترک این است که سندهایی که دارای شباهت هستند به احتمال زیاد توسط یک سند مشترک مورد استناد قرار می‌گیرند.

انگیزه روش پیشنهادی این مقاله، استفاده از ویژگی‌های رابطه‌ای در کنار ویژگی‌های متنی، برای برطرف کردن کاستی‌های ویژگی‌های متنی در پیدا کردن اسناد مرتبط می‌باشد.

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

۳-۲- شباهت رابطه‌ای

با توجه به ۶ نوع رابطه تعریف شده، شباهت رابطه‌ای بین هر دو سند P_i و P_j به کمک فرمول (۱) تعریف می‌شود:

$$relSim(P_i, P_j) = \sum_{k=1}^{6} F_k(P_i, P_j) \qquad (۱)$$

در فرمول بالا، F_k یک تابع است که میزان شباهت هر دو سند P_i و P_j را با توجه به رابطه متناظر با آن یعنی R_k بدست می‌آورد. تعریف این ۶ تابع در زیر آورده شده است.

مقدار بازگشتی $F_1(P_i , P_j)$، در صورت وجود رابطه R_1 از سند P_i به سند P_j، برابر ۱، و در غیر اینصورت برابر ۰ می‌باشد.

مقدار بازگشتی $F_2(P_i , P_j)$، در صورت وجود رابطه R_2 از سند P_i به سند P_j، برابر ۱، و در غیر اینصورت برابر ۰ می‌باشد.

مقدار بازگشتی $F_4(P_i , P_j)$، در صورت وجود رابطه R_4 از سند P_i به سند P_j، برابر ۱، و در غیر اینصورت برابر ۰ می‌باشد.

مقدار بازگشتی توابع F_3، F_5 و F_6 به ترتیب به کمک فرمول‌های (۲)، (۳) و (۴) بدست می‌آید:

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

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

$$F_6(P_i, P_j) = \frac{|citList_i \cap citList_j|}{|citList_i \cup citList_j|} \qquad (۴)$$

در هر یک از توابع بالا، مقدار بازگشتی ۰، به معنی ارتباط نداشتن و مقدار بازگشتی ۱، نشان‌دهنده یک ارتباط قوی از نظر رابطه متناظر با آن می‌باشد.

مقدار بازگشتی برای توابع F_3، F_5 و F_6 که به ترتیب به رابطه‌های R_3، R_5 و R_6 مرتبط هستند، مبتنی بر ایده مشابهی هستند. مثلا در مورد F_3، هر چه تعداد نویسندگان مشترک دو سند بیشتر باشد، میزان ارتباط آن‌ها بیشتر در نظر گرفته شده است. البته نسبت نویسندگان مشترک به کل نویسندگان این دو سند نیز اهمیت دارد. برای مثال، *دو نویسنده مشترک از سه نویسنده*، نسبت به *دو نویسنده مشترک از بین شش نویسنده*، بیانگر رابطه قوی‌تری می‌باشد. بنابراین F_3 با تعداد نویسندگان مشترک، رابطه مستقیم، و با تعداد کل نویسندگان، رابطه عکس دارد.

۳-۳- الگوریتم پیشنهاد استناد

از آنجایی که ورودی الگوریتم پیشنهادی تنها از متن تشکیل شده و ویژگی‌های دیگری مثل لیست مراجع یا لیست نویسندگان ندارد، امکان استفاده مستقیم از شباهت رابطه‌ای وجود ندارد. در نتیجه این الگوریتم ابتدا با بدست آوردن شباهت متنی هر سند موجود در مجموعه داده‌های محلی P_i و متن ورودی $input$، یعنی $textSim(P_i, input)$، تعداد ثابتی، C، از اسنادی که دارای بیشترین شباهت متنی می‌باشند را بعنوان مجموعه کاندید انتخاب می‌کند.

مجموعه کاندید که با توجه به ویژگی‌های متنی بدست آمده است شامل C سند از مجموعه داده‌های محلی می‌باشد، هر یک از این اسناد علاوه بر ویژگی‌های متنی شامل ویژگی‌های دیگری از جمله لیست نویسندگان و لیست مراجع می‌باشند. در نتیجه برای پیشنهاد لیستی از اسناد موجود در مجموعه داده‌های محلی بعنوان استناد برای متن ورودی، میزان شباهت هر سند با متن ورودی به کمک فرمول (۵) بدست می‌آید.

$$score(P_i, input) = \sum_{j=1}^{C} w_j \times relSim(P_i, Q_j) \qquad (۵)$$

در فرمول بالا، Q_j نشان‌دهنده j-امین سند موجود در مجموعه کاندید، C نشان‌دهنده تعداد عناصر این مجموعه، و w_j که از طریق فرمول (۶) بدست می‌آید، بعنوان وزن شباهت رابطه‌ای P_i و Q_j در شباهت کلی سند P_i به متن ورودی می‌باشد.

$$w_j = \frac{txtSim(input, Q_j)}{\max_{(1 \le k \le C)} txtSim(input, Q_k)} \qquad (۶)$$

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

۴- 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 نحوه ارزیابی و نتایج آزمایش‌ها ارائه شده است.

۴-۱- مجموعه داده

برای ایجاد مجموعه داده‌های لازم برای آزمایش سیستم پیشنهادی، یک پیمایشگر پیاده‌سازی شده است که با استفاده از سرویس *OAI*[۶] ارائه شده توسط سیستم *CiteSeerX*، اطلاعات اسناد منتشر شده در این سیستم را جمع‌آوری می‌کند. بعد از پالایش داده‌های بدست آمده و حذف اسنادی که مقدار زیادی از اطلاعات آن‌ها غیر معتبر بود، اسنادی که سال انتشار آن‌ها بعد از سال ۲۰۰۷ بود بعنوان داده‌های ورودی و بقیه اسناد بعنوان مجموعه داده محلی انتخاب شدند و در پایگاه داده *MySQL* ذخیره شدند. در نهایت مجموعه داده‌ای شامل ۱۳۰۰۰ سند آماده شد که ۶۰۰ سند از این مجموعه بعنوان مجموعه داده ورودی برای آزمایش انتخاب شد.

۴-۲- 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 بعنوان داده ورودی به سیستم داده می‌شود تا برای آن استنادهایی پیشنهاد شود. همچنین لیست مراجع این سند بعنوان 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 کارایی این سیستم در ادامه توضیح داده شده است.

فراخوانی:

برای هر سند ورودی، با مقایسه لیست مراجع آن با لیست استنادهای پیشنهاد شده، فراخوانی الگوریتم برای آن سند محاسبه می‌شود. برای بدست آوردن فراخوانی کل سیستم، میانگین فراخوانی‌ها برای کلیه اسناد موجود در مجموعه داده‌های ورودی محاسبه می‌شود.

احتمال استناد مشترک[۷]:

برای هر سند ورودی، ممکن است برخی از استنادهای پیشنهاد داده شده توسط الگوریتم، در لیست مراجع آن سند وجود نداشته باشد. چنین پیشنهادهایی لزوما 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 بعنوان یک پیشنهاد مناسب، قابل قبول است یا خیر.

در [7,8] 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 نمی‌باشد، احتمال اینکه آن سند بهمراه هر یک از اسناد لیست مراجع، بطور مشترک مورد استناد قرار بگیرد محاسبه شده و سپس میانگین آن برای همه اسناد لیست مراجع بدست 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[۸]:

مفید بودن سیستم‌های پیشنهاددهند نه تنها به عناصر پیشنهاد شده، بلکه به ترتیب این عناصر وابستگی دارد. این وابستگی با معیارهایی مثل احتمال استناد مشترک و فراخوانی قابل ارزیابی نیست. مسلم است که چنانچه استنادهایی که بیشتر مرتبط هستند در اوایل لیست استنادهای پیشنهاد شده قرار داشته باشند، بهتر است. معیار

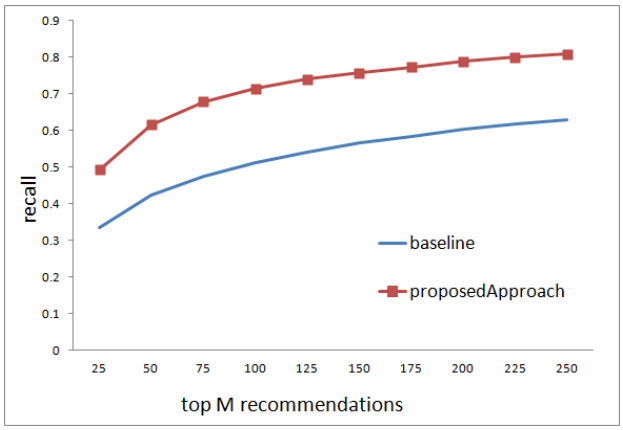
NDCG که یک معیار شناخته شده در بازیابی اطلاعات است، از این منظر به اندازه‌گیری کیفیت لیست پیشنهادها می‌پردازد [7,8].

۴-۳- نتایج

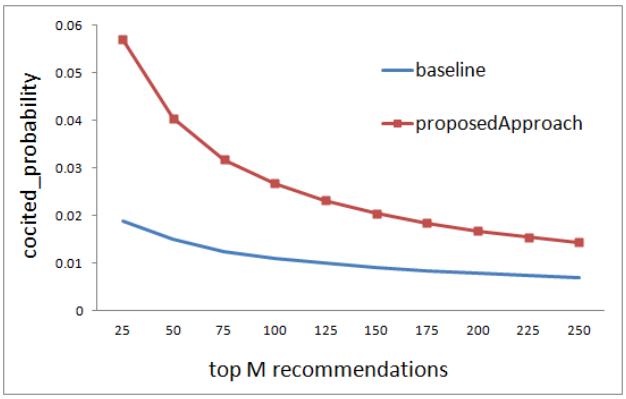
همانطور که در بخش ۳-۳ اشاره شد، در مرحله تولید مجموعه کاندید از شباهت متنی استفاده شده است و تعداد ثابت C سند بعنوان مجموعه کاندید انتخاب شده است. در آزمایشات انجام شده، شباهت متنی از طریق معیار *TFIDF* محاسبه می‌شود و با توجه به بررسی انجام شده و برقراری تعادل بین زمان پیشنهادها و کارایی آن‌ها، مقدار $C=25$ انتخاب شده است.

در مرحله انتخاب پیشنهادهای مرتب شده نیز مقدار M در آزمایش‌ها، در بازه [25, 250] انتخاب شده است.

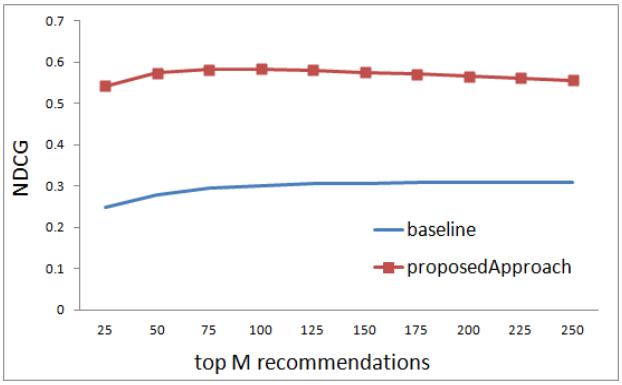
به منظور ارزیابی روش پیشنهادی و نشان دادن اهمیت استفاده از ویژگی‌های رابطه‌ای علاوه بر ویژگی‌های متنی در بدست آوردن اسناد مرتبط، یک روش پایه، برای مقایسه با روش پیشنهادی در نظر گرفته شده است. در این روش تنها از ویژگی‌های متنی برای پیشنهاد اسناد مرتبط با یک متن استفاده شده است. نتایج ارزیابی بر اساس ۳ معیار مورد نظر، در شکل (۱) تا شکل (۳) نمایش داده شده است.



شکل (۱) : نتایج ارزیابی از نظر معیار 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 رابطه‌ای، تاثیر قابل ملاحظه‌ای در بهبود نتایج از نظر هر یک از معیارهای ارزیابی داشته است. در نتیجه ثابت می‌شود که ایده استفاده از ویژگی‌های رابطه‌ای در کنار ویژگی‌های متنی، برای برطرف کردن کاستی‌های ویژگی‌های متنی در پیدا کردن اسناد مرتبط و بهبود سیستم‌های پیشنهاد استناد که انگیزه راه حل پیشنهادی این مقاله نیز بوده است، کاملا درست است.

۵- نتیجه‌گیری

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

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] 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.

[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] 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.

[7] 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.

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

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

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

[11] 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.

---

[1] Citation recommendation system
[2] Collaborative Filtering
[3] Citation analysis
[4] Bibliographic coupling
[5] Co-citation
[6] http://citeseerx.ist.psu.edu/oai.html
[7] Cocited-probability
[8] Normalized Discounted Cumulative Gain