ارائه روشی برای گسترش پرس‌وجو در بازیابی اطلاعات XML با استفاده از بازخورد کور

فروغ شهابیان[1]، احمد براانی[2]

[1] دانشجوی کارشناسی ارشد، گروه کامپیوتر دانشکده فنی و مهندسی، دانشگاه اصفهان ، اصفهان
shahabian@comp.ui.ac.ir

[2] استادیار، گروه کامپیوتر دانشکده فنی و مهندسی، دانشگاه اصفهان ، اصفهان
ahmadb@eng.ui.ac.ir

چکیده

در بازیابی اطلاعات XMLای که داده‌ها شمای ناهمگون، پیچیده و نامعلومی دارند، اغلب کاربران از پرس‌وجوهای مبتنی بر محتوا استفاده می‌کنند. بیشتر این پرس‌وجوها از اطلاعات مهمی که در ساختار XML نهفته است بهره نمی‌برند و نیاز کاربر را نیز به طور کامل بیان نمی‌کنند. از این رو و به منظور بالا بردن کیفیت بازیابی‌ها، می توان پرس وجوها را گسترش داد. برای گسترش پرس‌وجوها می‌توان از بازخورد رابطه کور استفاده کرد که اخیراً در بازیابی اطلاعات XML مورد توجه قرار گرفته است. در روش پیشنهادی این مقاله ابتدا بستر کلمات قابل افزودن به پرس‌وجو با استفاده از بازخورد کور شناسایی می‌شوند. سپس این بسترها برای یافتن کلمات مناسب مورد کاوش قرار می‌گیرند. بیشتر موتورهای بازخورد موجود، تنها از محتویات بازخورد کور بهره می‌برند در حالیکه روش ارائه شده در این رهاورد از اطلاعات ساختاری این بازخوردها نیز استفاده می‌کند. همچنین برای افزایش دقت بازیابی ها و کاهش ابهام پرس‌وجوها از معنای پرس‌وجو استفاده شده است. ارزیابی صورت گرفته، افزایش دقت بازیابی‌ها را با استفاده از روش پیشنهادی نشان می‌دهد.

کلمات کلیدی

بازیابی اطلاعات، XML ، گسترش پرس‌وجو، بازخورد رابطه کور،المان، روابط معنایی، شبکه واژگان

۱- مقدمه

امروزه XML به عنوان استانداردی جدید برای ذخیره سازی و مبادله اطلاعات در کتابخانه‌های دیجیتال، وب، اینترانت و غیره استفاده می‌شود. از این رو بازیابی اطلاعات روی داده‌های XML به شدت مورد توجه قرار گرفته‌است. از ویژگی‌های شاخص این نوع بازیابی‌ها، بازگرداندن بخشی از مستند به جای کل مستند است[1]. از تفاوت‌های این نوع سیستم‌های بازیابی با سیستم‌های بازیابی متنی رایج، در نوع اطلاعات مورد بازیابی و نوع پرس‌وجوها است. در بازیابی اطلاعات که روی داده‌های XML صورت می‌گیرد، داده‌ها به صورت XML کدگذاری شده‌اند. این ویژگی سبب می‌شود تا اطلاعات زمینه‌ی مستند در قالب مسیرهای XML ارائه شود. این امر برای پروسه بازیابی اهمیت بسیاری دارد. پرس‌وجوها در بازیابی اطلاعات XML به دو بخش عمده تقسیم می‌شوند: پرس‌وجوهای مبتنی بر محتوا و پرس‌وجوهای مبتنی بر محتوا و ساختار. با توجه به پیچیدگی زبان پرس‌وجوهای مبتنی بر محتوا و ساختار، ناهمگون بودن داده‌های XML و نیز فقدان دانش از شمای زیرین این داده‌ها برای کاربران غیر خبره، بسیاری از موتورهای جستجو از پرس‌وجوهای مبتنی بر محتوا استفاده می‌کنند. اما از آنجایی که اغلب این پرس‌وجوها نمی‌توانند نیاز کاربر را به طور کامل بیان کنند و از اطلاعات ساختاری که موجب بهبود

بازیابی‌ها می‌شود نیز استفاده نمی‌کنند، نیاز به غنی ساختن این پرس-وجوها احساس می‌شود. یکی از شیوه‌های مؤثر در غنی ساختن پرس‌و-جوها، گسترش دادن آنها است. برخی از روشهای گسترش پرس‌وجو مبتنی بر استفاده از بازخورد رابطه کور است. در این نوع بازخوردها فرض می‌شود تعدادی از المان‌ها که در مرحله‌ی اول بازیابی در بالای لیست مرتب نتیجه قرار دارند، مرتبط با پرس‌وجو هستند و می‌توان از این المان‌ها در غنی‌سازی پرس‌وجوها استفاده کرد.

در معدود کارهای انجام گرفته برای گسترش پرس‌وجو روی داده-های XMLای با استفاده از بازخورد، اغلب سعی بر آن بوده‌است که از محتویات این المان‌ها استفاده شود. اگر چه این کارها در بازیابی‌ها مؤثر بوده‌است، اما می‌توان گفت به دلیل بی‌توجهی به طبیعت نیم ساخت یافته متون XML ای، کاملاً از بازیابی اطلاعات متنی الهام گرفته شده‌است.

ایده اصلی این مقاله، گسترش پرس‌وجو با استفاده از اطلاعات ساختاری موجود در درخت XML و استفاده از معنای پرس‌وجو ها می‌باشد. تکنیک‌های مبتنی بر نزدیکی ساختاری المان‌ها از جمله تکنیک‌های مؤثر بازیابی اطلاعات روی داده‌های ساخت یافته هستند که منجر به بهبود دقت پرس‌وجو می‌شوند. این تکنیک‌ها روی داده-های نیم ساخت یافته کمتر بررسی شده‌اند.

۲- مروری برکارهای گذشته

از دیرباز، تکنیک‌های گوناگون بازخورد رابطه در بازیابی اطلاعات متنی رایج بوده است. این در حالی است که تاکنون تحقیقات چندانی روی بازخوردهای رابطه در بازیابی اطلاعات XML صورت نگرفته است. در ابتدا Rocchio تحقیقات روی گسترش پرس‌وجو را با استفاده از بازخوردهای رابطه کاربر در مدل فضا-بردار انجام داد [2]. وی کلماتی را به پرس‌وجو افزود که مستندات مرتبط‌تر را نشان می‌دهند و از افزودن کلماتی که مستندات غیرمرتبط را نشان می‌دهند اجتناب کرد. در [3] یک تحقیق کامل از تکنیک‌های بازخورد رابطه که روی مدل-های بازیابی مختلف صورت گرفته، ارائه شده است.

در بازیابی اطلاعات XML ای، اغلب موتورهای بازخورد تنها از محتویات المان‌هایی که مرتبط یا غیر مرتبط تشخیص داده شده‌اند استفاده می‌کنند، بدون آنکه به معنای ضمنی[1] که در ساختار این المان‌ها وجود دارد توجه کنند [4,5]. در[6]، چارچوبی برای پالایش پرس‌وجو بر اساس دانش آنتالوژی و بازخوردهای کاربر ارائه شده است. در این مقاله ابتدا وزن کلمات پرس‌وجو با استفاده از آنتالوژی مقداردهی و سپس با استفاده از بازخوردهای کاربر اصلاح می‌شوند. هر چند استفاده از بازخورد کاربر تا حدی باعث بهبود کارایی بازیابی شده است اما عدم استفاده از پتانسیلی که در ساختار XML نهفته است از نقاط ضعف این رهیافت محسوب می‌شود.

در معدود مقالاتی که در آنها به نوعی به ساختارXML توجه شده است، [7,8,9] گام مؤثری برای گسترش پرس‌وجوها برداشته‌اند. روش ارائه شده در این مقالات به گونه‌ای است که با استفاده از اطلاعات ساختاری (مسیر المان مرتبط، خصوصیات مستند شامل المان مرتبط) و محتوایی، نوع پرس‌وجو از پرس‌وجوی مبتنی بر محتوا به پرس‌وجوی مبتنی بر محتوا و ساختار تغییر داده می‌شود.

روش‌های ارائه شده در این مقالات باعث افزایش دقت در بازیابی‌ها می‌شوند. در[10] پرس‌وجوهای مبتنی بر محتوا با استفاده از بازخورد رابطه کور و اطلاعات ساختاری گسترش می‌یابند. این اطلاعات ساختاری از همسایگان المان‌هایی که شامل کلمات پرس‌وجو هستند، استخراج می‌شود. در [11] پیشوند مشترک المان‌های مرتبط به پرس-وجو افزوده شده است. ضمن اینکه در این مقاله گسترش هر دو نوع پرس‌وجو بحث شده است. متاسفانه هیچگونه ارزیابی برای روش پیشنهادی این مقاله صورت نگرفته است.

۳- نگاهی اجمالی به فرایند گسترش پرس‌وجو

یک روش برای گسترش پرس‌وجو، انتخاب کلماتی است که به مفهوم پرس‌وجو نزدیک هستند. برای یافتن کلمات مناسب قابل افزودن به پرس‌وجو باید بسترهای مناسبی مورد جستجو قرار گیرند. در این مقاله گسترش پرس‌وجو شامل دو فاز اصلی است. در فاز اول بسترهایی که شامل کلمات کاندید برای گسترش پرس و جو هستند انتخاب می شوند و در فاز دوم کلمات کاندید از این بسترها انتخاب می‌گردند.

شکل(۱) نگاهی اجمالی به فرایند گسترش پرس‌وجو در روش پیشنهادی این مقاله دارد. در ادامه، گام‌های این فرایند مرور خواهدشد.

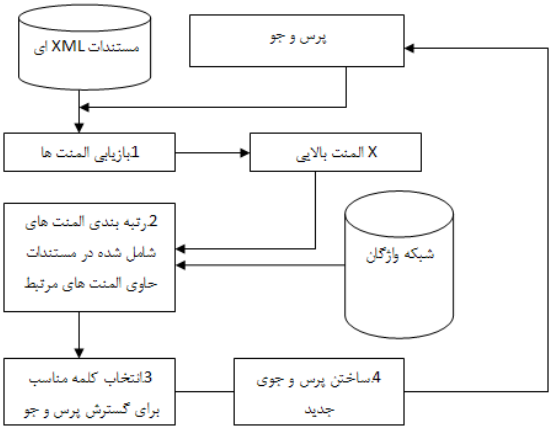
۱. در گام اول فرایند، المان‌ها طبق آنچه کاربر در پرس‌وجو آورده است بازیابی می‌شوند.

۲. در گام دوم، المان‌هایی که احتمال یافتن کلمات مناسب قابل افزودن به پرس‌وجو در آنها بیشتر است شناسایی می‌گردند. سپس المان‌ها طبق این معیار رتبه بندی خواهند شد.

۳. در گام سوم، کلمات کاندید وزن‌دهی و کلماتی با وزن بالاتر برای افزودن به پرس‌وجو انتخاب می‌شوند.

۴. در گام آخر کلمات منتخب به پرس‌وجو افزوده می‌شوند.

طبق آنچه در بازخورد رابطه کور مطرح است X المانی که در بالای لیست بازیابی قرار گرفته‌اند، به پرس‌وجو مرتبط‌تر هستند. بنابراین کافی است به جای آن که کل مستند XML برای یافتن کلمات مورد نظر جستجو شود، این المان‌ها و المان‌های شبیه به آنها مورد کاوش قرار گیرد.



شکل ۱.فرایند گسترش پرس وجو

۴- مفهوم شباهت در XML

معمولاً در داده‌های XML ای شباهت المان‌ها با استفاده از اطلاعات ساختاری برآورد می‌شود. منظور از اطلاعات ساختاری، اطلاعاتی است که از روی ساختار درخت XML بدست می آید. به عنوان مثال تعداد زیرشاخه های یک المان خاص، مسیر رسیدن به یک المان از ریشه، تعداد فرزندان و اجداد یک المان و اطلاعاتی از این دست اطلاعات ساختاریی هستند که می توان تا حدودی از آن‌ها برای تشخیص المان‌هایی که مضامین مرتبط و مشابهی دارند استفاده کرد.

اما وقتی سخن از بازیابی اطلاعات به میان می‌آید، انتظار می‌رود که برای تشخیص المان‌های مشابه و مرتبط علاوه بر اطلاعات ساختاری به محتویات المان‌ها نیز توجه شود. یعنی همانطور که از روی اطلاعات برگرفته از درخت XML می‌توان المان‌های هم‌زمینه که مضامین شبیه و مرتبطی دارند را تشخیص داد، از روی محتویات این المان‌ها نیز می‌توان دقت سنجش این تشابه را افزود.

از این رو و با توجه به توضیحات بالا تعریفی از نزدیکی المان ها ارائه خواهد شد که میزان نزدیکی المان‌ها را وابسته به دو عامل می‌داند. این دو عامل عبارتند از: نزدیکی‌های برگرفته از ساختار و نزدیکی برگرفته از محتوای المان‌ها.

فرمول ۱ همین تعریف را بیان می کند.

(۱)

$$N(e_i,e_j) = \gamma \times NS(e_i,e_j) + (1-\gamma) \times NC(e_i,e_j)$$

$NS(e_i,e_j)$ نزدیکی ساختاری، $NC(e_i,e_j)$ نزدیکی محتوایی و e_i المان i را نشان می‌دهد. در ادامه در این خصوص توضیحات بیشتری ارائه خواهد شد.

۴-۱- مفهوم نزدیکی بر گرفته از ساختار

آنچه که در این رهاورد نزدیکی برگرفته از ساختار المان‌ها نامیده می‌شود، نزدیکی مفهومی است که بین المان‌ها در درخت XML به دلیل چیدمان آن‌ها وجود دارد. به عبارت دیگر آرایش نودها در درخت XML به صورت اتفاقی نیست و محل قرار گرفتن المان‌ها در XML خود دارای معنا می‌باشد. با بررسی محل قرار گرفتن المان‌ها می‌توان دانش‌های مهمی را استخراج کرد.

از جمله دانش‌های مهمی که از چیدمان المان ها در درخت XML استنتاج می‌شود، میزان ارتباط معنایی محتوای المان‌هاست که به فاکتورهای گوناگونی بستگی دارد. یکی از این فاکتورها فاصله بین المان‌هاست که با شمارش تعداد یال‌های بین آن‌ها محاسبه می‌گردد. هر چه فاصله بین این المان‌ها در درخت XML بیشتر باشد گواه این مطلب خواهد بود که در مستند XML نیز محل قرار گرفتن این دو المان از هم فاصله دارد. از طرفی از آن جا که در مستندات، مطالب مرتبط که مضامین مشابه و مرتبطی با هم دارند، پیوسته و نزدیک به هم قرار خواهند گرفت این نتیجه گرفته خواهد شد که مفهوم این دو المان به هم کمتر شبیه خواهند بود. برعکس این مطلب نیز صحیح است یعنی اگر فاصله‌ی المان‌ها در درخت XML کم باشد نشان‌دهنده این مطلب خواهد بود که در مستند مذکور نیز این دو المان به‌هم نزدیک هستند در نتیجه مضمون این دو المان به هم مرتبط خواهد بود.

یکی دیگر از مهمترین معیارهای نزدیکی المان‌ها، کوچکترین جد مشترکشان[۲] است[11]. تحقیقات نشان داده است هر چه اختلاف ارتفاع کوچکترین جد مشترک دو المان با ارتفاع هر دو المان کمتر باشد، ارتباط معنایی آن دو بیشتر خواهد بود. عکس این مطلب نیز صحیح است یعنی هر چه اختلاف ارتفاع کوچکترین جد مشترک دو المان با ارتفاع المان‌های مورد نظر بیشتر باشد ارتباط معنایی آن دو المان کمتر است.

بر مبنای آنالیز صورت گرفته و در نظر گرفتن عوامل موثر بالا، رابطه‌ی زیر برای سنجیدن میزان نزدیکی المان ها از روی ساختار درخت XML به صورت زیر ارائه شده است:

(۲)

$$NS(e_i,e_j) = \propto \times \frac{depth\left(lca(e_i,e_j)\right)}{depth(e_i) + depth(e_j)}$$

ارتفاع المان e_i با $depth(e_i)$، ارتفاع المان e_j با $depth(e_j)$، کوچکترین جد مشترک با $lca(e_i,e_j)$ و فاصله بین دو المان با $distance(e_i,e_j)$ نشان داده می‌شود. $\propto$ نیز ضریبی است که توسط کاربر تعیین می‌شود.

۴-۲- مفهوم نزدیکی بر گرفته از محتوا

آنچه که در این رهاورد نزدیکی برگرفته از محتوا نامیده می‌شود، میزان شباهتی است که بین المان‌ها به دلیل محتوای آنها وجود دارد. در این تحقیق نزدیکی برگرفته از محتوای المان‌ها تلفیقی از شباهت‌های معنایی و لغوی المان‌ها در نظر گرفته شد. براورد این نوع شباهت از برآورد شباهت میان مستندات متنی الهام گرفته شده است

(۳)

$$NC(e_i, e_j) = Nsyn(e_i, e_j) \times Nsem(e_i, e_j)$$

۴-۲-۱- مفهوم شباهت لغوی

در این بخش میزان شباهت موجود بین محتویات المان‌ها از نظر لغوی محک خورده است. به عبارت دیگر در این بخش برای تعیین میزان شباهت محتوایی متنی المان‌ها تنها روی نحو کلمات تمرکز شده و توجهی به معنای کلمات نشده است.

شباهت کسینوسی معیار مناسبی برای سنجش شباهت بین دو بردار n بعدی است که با محاسبه زاویه بین دو بردار این مهم را انجام می دهد. در این روش ابتدا محتویات متنی مستندات به روش بردار ذخیره خواهد شد. اجزای بردارها، کلمات شاخص برگرفته از محتویات المان‌ها هستند که طبق فرمول های وزن‌دهی، وزن دهی خواهند شد. در این رهاورد، کلمات شاخص طبق دو معیار آماری (چگالی کلمه در متن و فرکانس کلمه) و با استفاده از فرمول OKAPI گسترش‌یافته [13]وزن دهی شده اند.

(۴)

$$Relevence(W_i, U_i) = \left(0.3 * \frac{tf}{\left(0.5 + 1.5 * \frac{L_i}{L}\right) + tf} * \log \frac{N - n_j + 0.5}{n_j + 0.9} \right)$$

$$\overrightarrow{U_i} = \left(Relevence(W_1, U_i), Relevence(W_2, U_i), \dots, Relevence(W_j, U_i)\right)$$

U_iمحتویات متنی المان‌ها، tf فرکانس کلمه W_i در U_i، L_i طول U_i، n_j تعداد u های شامل کلمه W_j، N تعداد کل Uها در مجموعه، $\bar{L} = \frac{\sum_{i=1}^{N} L_i}{N}$ میانگین طول U است.

بعد از محاسبه وزن‌ها و ساختن بردار هر متن، نوبت به محاسبه شباهت متن‌ها می‌رسد.

(۵)

$$Nsyn(e_i, e_j) = Sim_{Synthatic} = \frac{\overrightarrow{U_i} . \overrightarrow{U_h}}{\|U_i\| * \|U_h\|}$$

این روش اگر چه در حال حاضر یکی از ساده‌ترین و در عین حال پر-کاربردترین روشهای موجود برای ارزیابی شباهت مستندات متنی است. اما از آنجا که این روش تنها به نحو کلمات توجه می کند روش کاملی نیست. برای روشن‌تر شدن موضوع به مثال زیر توجه کنید:

در یک سیستم هوشمند انتظار می‌رود تا اگر محتویات متنی المان U_i شامل کلمات "clever men" و محتویات متنی المان U_j شامل کلمات "Bright person" است، شباهتی بین این دو المان در نظر گرفته شود. اما از آنجایی که بردار هر المان بر اساس حضور و یا عدم حضور کلمات شاخص ساخته می شود، شباهت بین این دو متن صفر در نظر گرفته می شود. به همین منظور و به سبب غنی ساختن برآورد شباهت‌های متنی دو المان، در این رهاورد شباهت معنایی جاری بین دو المان نیز بررسی می‌شود. در ادامه در این خصوص توضیحات بیشتری را ارائه خواهد شد.

۴-۲-۲- مفهوم شباهت معنایی

شباهت معنایی معیاری که میزان نزدیکی معنایی دو کلمه را در نشان می‌دهد. در این مقاله از شبکه واژگان برای استخراج روابط معنایی بین کلمات موجود در محتویات المان‌ها استفاده شده است. شبکه واژگان یک فرهنگ واژه‌ی الکترونیکی است که در آن معنای واژه‌ها از طریق روابط بین آنها بیان می‌شود. روابط معنایی که در این مقاله مورد بررسی قرار گرفته است رابطه ابرمعنایی، زیرمعنایی می‌باشد. کوتاهترین فاصله وزن‌دار کلمات المان در شبکه واژگان می‌تواند معیاری برای نزدیکی معنایی کلمات المان و در نتیجه خود المان‌ها باشد. در این مقاله وزن رابطه هم‌معنایی ۱ و وزن رابطه ابر معنایی و زیر معنایی ۰،۸۵ در نظر گرفته شده است.

(۶)

$$Nsem(W_i, W_j) = \frac{1}{distance(W_i, W_j)}$$

$distance(W_i, W_j)$ کوتاهترین فاصله وزن‌دار دو کلمه ی W_i وW_j در شبکه واژگان را نشان می‌دهد که باید از حد آستانه‌ای که توسط کاربر تعیین می‌گردد کمتر باشد

(در این مقاله حد آستانه پنج در نظر گرفته شده است).

(۷)

$$Nsem(e_i, e_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} Nsem(W_i, W_j)$$

m و n تعداد کلمات المان‌ها را نشان می‌دهد.

۵- انتخاب کلمات مناسب

پس از آنکه المان‌های مشابه با المان‌های مرتبط تشخیص داده شدند، نوبت به انتخاب کلمه از میان محتوای متنی آن‌ها می‌رسد. جستجو در محتویات این المان‌ها سبب می‌شود تا محتویات همه المان‌ها برای

انتخاب کلمات مناسب مورد کاوش قرار نگیرد که این کارسبب صرفه جویی در زمان و منابع می‌گردد.

به منظور انتخاب کلمه مناسب، تمام کلمات موجود در محتویات المان‌های مورد بررسی طبق رابطه (۱۰)وزن‌دهی شده و نهایتاً تعدادی از کلمات که وزن بالاتری دارند برای اضافه شدن به پرس‌وجو انتخاب می‌شوند. رابطه(۱۰) به سه عامل مهم وابسته می‌باشد:

(۱) اهمیت کلمه در محتوای متنی المان که طبق اطلاعات آماری محاسبه می‌گردد.

(۲) میزان ارتباط المان حاوی کلمه مورد نظر نسبت به پرس‌و-جوی اولیه

(۸)

$$N(e_j,q) = \frac{\sum_i^m Ns(e_j,e_i)}{m}$$

qپرس‌وجوی اولیه،m تعداد کلمات پرس‌وجوی اولیه بعد از حذف کلمات توقف، e_o المان شامل i امین کلمه پرس‌وجو و$N(e_j,q)$ درجه ارتباط المان e_jبا پرس‌وجوی q را نشان می‌دهد.

ذکر این نکته ضروری است که چنانچه چندین المان شامل i امین کلمه‌ی پرس‌وجو باشند از المانی به عنوان e_o در رابطه بالا استفاده می‌گردد که نزدیکترین فاصله را نسبت به المان مورد مقایسه داشته باشد.

(۳) کمیابی معنایی کلمه که در رهاورد با درجه چندمعنایی[۳] کلمه تعریف می‌شود . از آنجایی که افزودن کلماتی به پرس‌وجو که دارای چندین معنا هستند باعث انحراف در بازیابی‌ها می‌شود، این عامل مورد بررسی قرار گرفته است. رابطه زیر برای تعریف کمیابی معنایی آمده است

(۹)

$$rarity(w_i) = log\left(\frac{Max-Polysemy}{|senses(w_i)|}\right)$$

Max-Polysemy عددی ثابت است که مقدار آن بیشتر از تعداد معانی کلمه‌ای است که بیشترین معنی را در شبکه واژگان دارد. $|senses(w_i)|$ تعداد معانی کلمه w_i را نشان می‌دهد.

(۱۰)

$$weight(w_i) = Relevance(w_i,u_j) * rarity(w_i) * N(e_j,q)$$

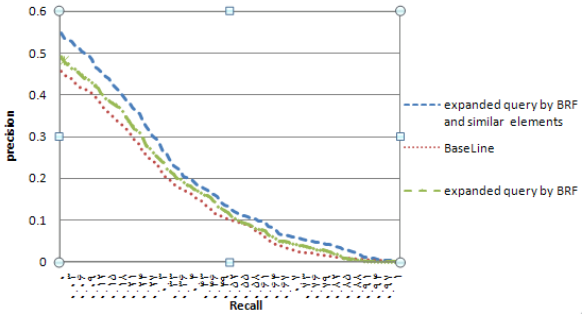
۶- پیاده سازی و ارزیابی

در این پژوهش، چارچوب پیشنهادی در محیط جاوا و با استفاده از پایگاه داده‌های Sql server روی ماشینی با ۵۱۲مگابایت RAM و ۱،۶ مگا هرتز Cpu پیاده سازی شده است. همچنین برای ارزیابی روش پیشنهادی از محک INEX ۲۰۰۹ [12]استفاده شده است که شامل مجموعه‌ای از مستندات XML و یک مجموعه از پرس‌وجوهای NEXI به همراه مجموعه ارزیابی نتایج برای هر پرس‌وجو می‌باشد.

برای بازیابی مستندات XML ای از موتور جستجوی Topx استفاده شده است. معیار ارزیابی روش پیشنهادی این مقاله بر مبنای دقت-فراخوانی می‌باشد. همانطور که در شکل۲ مشاهده می‌شود اگر تنها المنت‌های بازخورد کور به‌عنوان بستر کلمات گسترش مورد کاوش قرار گیرند دقت بازیابی‌ها نسبت به مبنا افزایش می‌یابد اما اگر از المنت-های مشابه المنت‌های بازخورد کور نیز به‌عنوان بستر کلمات گسترش استفاده شود دقت بازیابی‌ها نسبت به حالت اول و مبنا بهبود بیشتری می‌یابد.

۷- نتیجه گیری

این مقاله با استفاده‌ی مناسب از اطلاعات ساختاری موجود در درخت XML گام موثری در استفاده از بازخورد رابطه کور در گسترش پرس‌وجوها برداشته است. همچنین با کم‌اهمیت‌کردن کلمات چند معنا در فرایند وزن‌دهی کلمات کاندید گسترش موجب کاهش ابهام در پرس‌وجوهای جدید شده است. نتایج ارزیابی‌ها افزایش دقت بازیابی‌ها را نشان می‌دهد.



شکل ۲:نمودار دقت-فراخوانی

مراجع

[1] S. R. Joty and S.Sadid-AlHasan, "*Advances in Focused Retrieval: A General Review*", 10th international conference on Computer and information technology, 2008. iccit 2007.

[2] B.Ribeiro, R.Baeza and." *Modern Information Retieval*". NewTork: ACM Press, 1999,pp.117-120.

[3] I.Ruthven and M. Lalmas," *A survey on the use of relevance feedback for information access systems*", Knowledge Engineering Review, vol. 18, no.2, pp. 95 - 145 , june 2003.
[4] M.Mandelbrod, Y.Mass and. "*Relevance Feedback for XML Retrieval*." In INEX 2004 Workshop, 2004. pp.154-157.
[5] C.J.Crouch, A.Bellamkonda and A.Mahajan. "*Flexible XML retrieval based on the extended vector model*." INEx 2004 Workshop, 2004. pp.149-153.
[6] H.Pan, "*Relevance feedback in XML Retrieval*." , Ph.D. thesis, Max-Planck-Institut Informatik, Germany ,2003.
[7] M.Theobald, R.Schenkel and. "*Relevance Feedback for Structural Query Expansion," in Advances in XML Information Retrieval and Evaluation* , Vol. 3977/2006, Heidelberg: Springer, 2006, pp.344-357.
[8] M.Theobald, R.Schenkel and. "*Structural Feedback for Keyword-Based XML Retrieval*," in Advances in Information Retrieval. Vol. 3936/2006, Heidelberg: Springer, 2006,pp. 326-337.
[9] M.Theobald, R.Schenkel and,"*Feedback-driven structural query expansion for ranked retrieval of XML data*," In 10th International Conference on Extending Database Technologies (EDBT 2006), Munich, Germany, Mar. 2006.
[10] W. Hsu, M. L. Lee and X. Wu. "*Path-augmented keyword search for XML documents*," ICTAI 2004, 2004 , pp.526-530.
[11] M.Boughanem, L.Hlaoua and ,"*Towards Contextual and Structural Relevance Feedback in XML Retrieval*,"In Workshop on Open Source Web Information Retrieval(OSWIR). 2005.
[12] Otago university ,.inex., [Online]. vailable :http://www.inex.otago.ac.nz/ [access:octobr,20, 2008]
[13] Robertson, S., Walker, S., Jones, S., HancockBeaulieu,M. and Gatford, M. OKAPI at TREC3. 3rd Text Retrieval Conference, 1995

زیرنویس‌ها

[1] Implicit Semantic
[2] Lowest Common Ancestor
[3] Polysemy