خوشه‌بندی فراابتکاری اسناد فارسی اِکس‌اِم‌اِل مبتنی بر شباهت ساختاری و محتوایی

علی مرادی لالمی، اسد ا... شاه‌بهرامی، رضا ابراهیمی‌آتانی* و مهران علیدوست‌نیا
گروه مهندسی کامپیوتر، دانشکدهٔ فنی، دانشگاه گیلان، رشت، ایران

چکیده

با توجه به رشد فزایندهٔ تعداد اسناد XML، سازماندهی مؤثر این اسناد به‌منظور بازیابی اطلاعات مفید از آنها ضروری است. یک راه حل امکان‌پذیر، انجام خوشه‌بندی برروی اسناد XML به‌منظور کشف دانش است. مسئلهٔ کلیدی در خوشه‌بندی اسناد XML این است که چگونه می‌توان شباهت بین اسناد XML را اندازه‌گیری کرد. استفاده از روش‌های متداول خوشه‌بندی اسناد متنی که اطلاعات محتوایی را برای اندازه‌گیری شباهت سند به کار می‌گیرند، باعث نادیده‌گرفتن اطلاعات ساختاری موجود در اسناد XML می‌شود. در این مقاله، مدل جدیدی با نام مدل فضای ماتریسی برای بازنمایی هر دو ویژگی ساختاری و محتوایی داده‌ها در اسناد XML، پیشنهاد می‌شود. بر اساس این مدل، معیار شباهت جاکارد تعریف و در نهایت از الگوریتم رقابت استعماری برای خوشه‌بندی اسناد XML استفاده می‌شود. نتایج تجربی نشان می‌دهد که مدل پیشنهادی و تابع نزدیکی معرفی‌شده در شناسایی اسناد مشابه که دارای اطلاعات ساختاری و محتوایی یکسان هستند، مؤثر است. این روش می‌تواند به منظور بهبود دقت خوشه‌بندی و افزایش بهره‌وری در بازیابی اطلاعات XML مورد استفاده قرار گیرد.

واژگان کلیدی: خوشه‌بندی، زبان فارسی، الگوریتم رقابت استعماری، پردازش زبان طبیعی و بازیابی اطّلاعات.

۱- مقدّمه

امروزه XML یا زبان نشانه‌گذاری توسعه‌پذیر[1]، به‌دلیل انعطاف‌پذیری بالا و ذات خودتوصیفی[2] آن، به‌عنوان یک استاندارد برای نمایش و تبادل اطلاعات در وب پدید آمده است. درنتیجه، مقدار زیادی از اطلاعات در XML نشان داده شده و راه‌کارهای متعددی برای ارائه، ذخیره‌سازی، ادغام و پرس وجوی اسناد XML توسعه داده شده‌اند. بنابراین توسعه راه‌بردهایی با کارایی بالا برای مدیریت و تجزیه و تحلیل کارآمد مجموعهٔ بسیار زیادی از اسناد XML، اجتناب‌ناپذیر است. یکی از روش‌هایی که بسیاری از پژوهش‌گران بر روی آن متمرکز هستند، خوشه‌بندی اسناد XML در گروه‌هایی است که از نظر محتوا و ساختار مشابه‌اند. فرآیند خوشه‌بندی اسناد XML نقش حیاتی در بسیاری از حوزه‌ها مانند بازیابی اطلاعات[3]، بهبود پردازش پرس‌وجو[4]، یکپارچه‌سازی داده‌ها[5]، وب سرویس‌ها[6] مانند خوشه‌بندی سرویس‌های وب به‌منظور بهبود فرآیند کشف سرویس، وب‌کاوی[7] مانند خوشه‌بندی نتایج جستجو در موتورهای جستجو[8] و بیوانفورماتیک[9] دارد (الجلواری و همکاران، ۲۰۱۱).

خوشه‌بندی، به‌طورکلی، یک راه‌برد مفید برای گروه‌بندی یا دسته‌بندی اشیای داده‌ای در داخل یک گروه یا خوشه است که در ویژگی‌های مشابه، مشترک هستند (زو و وانش، ۲۰۰۵). خوشه‌بندی اسناد XML یک فرآیند پیچیده است و آن را به میزان قابل‌توجهی با خوشه‌بندی داده‌های

[1] Extensible Markup Language
[2] Self-Describing

[3] Information Retrival
[4] Query Processing
[5] Data Integrity
[6] Web Service
[7] Web Mining
[8] Search Engine
[9] Bioanformatic

مسطح و اسناد متنی متفاوت می‌کند. مشکلات خوشه بندی اسناد XML عبارتند از: اوّل‌این‌که، الگوریتم های خوشه‌بندی نیاز به محاسبهٔ شباهت بین مجموعه‌های مختلف اسناد XML دارند، به‌دلیل ناهم‌گونی در اسناد XML، چالش‌های بسیاری برای شناسایی تابع تشابه ایده‌آل وجود دارد. دوم‌این‌که، سازماندهی ساختاری و محتوایی اسناد XML، ابعاد ضمنی را که یک الگوریتم خوشه‌بندی برای اداره کردن آن نیاز دارد، افزایش می دهد که در نهایت به خوشه های بی‌معنا منجر می‌شود. به این مشکلات، چالش‌های ذاتی خط و زبان فارسی، نظیر نبود نکات گرامری تعریف‌شده، وجود لغات ترکیبی چندجزئی، نبود اطلاعات آوایی، فقدان یک واژگان کامل زبان فارسی و منابع زبان‌شناختی یا هستان‌شناسی[1] معتبر را نیز باید اضافه کرد.

اسناد XML دارای ویژگی‌های مختلفی از جمله ساختاری، محتوایی و معنایی[2] هستند، بنابراین خوشه‌بندی اسناد XML بر اساس یک ویژگی درحالی‌که ویژگی‌های دیگر نادیده گرفته می‌شود، رسیدن به خوشه‌بندی دقیق را با شکست مواجه می‌کند. با توجه به‌این‌که بسیاری از پژوهش‌های موجود در مورد خوشه‌بندی، فقط بر روی یکی از ویژگی‌های اسناد XML و آن هم در محدودهٔ اسناد XML با محتوای انگلیسی متمرکز هستند، در این مقاله یک مدل جدید، به نام مدل فضای ماتریسی[3] برای شناسایی شباهت ساختاری و محتوایی اسناد فارسی XML پیشنهاد می شود. این روش براساس ایدهٔ نمایش ساختار و محتوای یک سند XML به‌صورت یک ماتریس است. به این معنی که هر مسیر[4] از برچسب ریشه به برگ در ساختار درختی سند XML به‌همراه مقدار برگ وابسته به آن که نشان‌دهندهٔ محتوای سند است، به یک عنصر در ماتریس نگاشت می‌شود؛ سپس از معیار شباهت جاکارد[5]، برای بهبود سنجش مشابهت بین اسناد بهره‌گیری و درنهایت عمل خوشه‌بندی با استفاده از الگوریتم رقابت استعماری[6] انجام می‌شود.

با توجه به مطالب گفته شده، سهم علمی مقاله به شرح زیر است:

- معرفی چارچوبی برای خوشه‌بندی اسناد فارسی XML.
- ارائهٔ مدل جدیدی برای بازنمایی هردو ویژگی ساختاری و محتوایی اسناد XML.
- معرفی الگوریتم خوشه‌بندی کارآمد با استفاده از رویکرد فراابتکاری[7].
- توانایی خوشه‌بندی مجموعه‌ای از اسناد XML همگن و ناهمگن.

ادامهٔ مقاله به شرح زیر سازماندهی شده است، در بخش دوم کارهای مشابه انجام‌شده در زمینهٔ خوشه‌بندی اسناد XML شرح داده و در بخش سوم تعاریف اولیهٔ مورد نیاز بیان می‌شود؛ روش پیشنهادی در بخش چهارم معرفی و ارزیابی و مقایسهٔ روش پیشنهادی در بخش پنجم انجام و در نهایت در بخش ششم نتایج پژوهش ارائه می‌شود.

۲- کارهای مشابه

الگوریتم‌های مختلفی برای خوشه‌بندی اسناد XML پیشنهاد شده است؛ در برخی از این روش‌ها، اسناد XML به‌صورت درختان برچسب خوردهٔ ریشه‌دار مدل می‌شوند، در این روش میزان شباهت بین اسناد با فاصلهٔ ویرایشی[8] بین دو درخت محاسبه می‌شود. منظور از فاصلهٔ ویرایشی حداقل تغییرات برای تبدیل یک درخت به درخت دیگر است که شامل درج، حذف و برچسب‌زدن یک گره است. در الگوریتم‌های دیگر، خوشه‌بندی بر اساس ساختار اسناد و روابط سلسله‌مراتبی موجود بین عناصر سند، انجام می شود. در این دسته الگوریتم ها شباهت بین اسناد بر اساس عناصر مشترک و روابط ساختاری در نظر گرفته می شود. در برخی الگوریتم‌ها از روش‌های معنایی، مانند بررسی این‌که آیا شباهت معنایی میان نام های دو عنصر وجود دارد یا نه، استفاده شده است. در ادامه به معرفی برخی کارهای انجام‌شده در این زمینه،می‌پردازیم.

الگوریتم (نایاک ۲۰۰۹) یکی از روش‌های خوشه‌بندی اسناد XML است. این الگوریتم با استنتاج ساختاری از سند، شروع به کار می‌کند. در ابتدا سندها به‌صورت درختان برچسب‌دار مرتب‌شده، تبدیل می‌شوند. هر برچسب یا نام عنصر استفاده شده در سند بر اساس ترتیب ظاهرشده در سند با یک عدد صحیح مشخص می‌شود. ساختار سطحی نشان‌گر سطوح و برچسب عناصر در هر سطح از ساختار است. ساختار سطحی شامل اطلاعاتی از قبیل نام عناصر، رویداد و سطوح آن‌ها به‌صورت

[1] Ontology
[2] Semantic
[3] Matrix Space Model
[4] Path
[5] Jaccard
[6] Imperialist Competitve Algorithm
[7] Metaheuristic
[8] Edit Distance

سلسله‌مراتبی است. در این روش از تابع مشابهت سطحی[1] برای اندازه‌گیری میزان شباهت ساختاری بین دو شیء XML مورد استفاده قرار می‌گیرد. در این روش محتوای اسناد نادیده گرفته شده است.

در روش قبلی ارتباطات ساختاری مانند رابطهٔ والد-فرزندی در نظر گرفته نمی‌شود و اطلاعات ساختاری که در ساختار سطحی وجود دارد؛ این است که کدام عناصر در کدام سطح از سند XML وجود دارند. درنظرگرفتن رابطهٔ بین عناصر در سطوح مختلف می‌تواند منجر به ساخت یک ساختار سطحی یکسان برای دو سند با دو ساختار متفاوت شود. در روش پیشنهادی (آنتونلیس و همکاران، ۲۰۰۸) ساختار جدیدی به نام ساختار لبه‌ای‌سطحی[2] جهت نمایش اسناد XML ارائه شده است. ساختار لبه‌ای‌سطحی، لبه‌های مجزا و مشخص از هر سند را دسته‌بندی می‌کند و به‌صورت برداری از لبه‌ها سازماندهی و در هر سطح، فهرستی از لبه‌ها قرار می‌گیرد. هر لبهٔ مشخص به‌انحصار توسط دو گرهٔ مشخص از نمایش درختی سند XML تعریف می‌شود. در ابتدا لبه‌ها با اعداد صحیح برچسب‌گذاری می‌شوند و این اعداد برای نمایش ساختار لبه‌ای سطحی مورد استفاده قرار می‌گیرد. مزیت این روش حفظ ارتباطات ساختاری بین گره‌های سطوح متوالی از سند XML است.

روش ارائه‌شده در (چویی و همکاران ۲۰۰۷)، از شباهت مسیر بین گره‌های داده برای خوشه‌بندی اسناد استفاده می‌کند. این روش نیز اسناد XML را به‌صورت درختان برچسب‌خوردهٔ ریشه‌دار مدل می‌کند؛ سپس تمام مسیرهای مطلق موجود در درخت سند را استخراج و از آن به‌عنوان اساس محاسبهٔ شباهت استفاده می‌کند. یک مسیر، دنبالهٔ اسامی عناصر از برچسب ریشه به برچسب برگ موجود در هر سند است. برای بررسی شباهت بین دو مسیر نیز از فاصلهٔ ویرایشی بین دو مسیر استفاده می‌کند. یعنی حداقل عملیات مورد نیاز شامل درج، حذف و اضافه برای تبدیل یک مسیر به مسیر دیگر است؛ سپس با ایجاد ماتریس مشابهت مسیر عملیات خوشه‌بندی صورت می‌گیرد.

اگرچه در روش (آگروال و همکاران ۲۰۰۷) با استخراج ساختارهای الگوهای تکراری و سپس وزن‌دهی به ساختارهای متداول سعی داشته خوشه‌بندی دقیق و کارا را نتیجه دهد، ولی با نادیده‌گرفتن رابطهٔ بین گره‌های هم‌زاد یا دارای والد مشترک، به خاطر توجه به ساختارهای سلسله‌مراتبی مسیر عنصرهای XML و گره‌های صفت و تنها درنظرگرفتن ساختار اسناد، اطلاعات ارزشمندی را در خوشه‌بندی لحاظ نکرده است. به نظر می‌رسد اگر در یافتن شباهت و درنهایت خوشه‌بندی از ترکیب خصوصیات محتوایی، ساختاری و سلسله‌مراتبی به‌شکل مناسب بهره گرفته می‌شد، نتیجهٔ به‌دست‌آمده کارآیی و کیفیت بالاتری داشت.

در مقالهٔ (کیم و همکاران ۲۰۰۸) و نیز در بسیاری از مقالات معرفی‌شده در (تاگارلی و همکاران ۲۰۱۰) یکی از نقاط چالش‌برانگیز، وجود پارامترهایی است که با تغییر اندک آنها خوشه‌بندی دست‌خوش نتایج بسیار متفاوتی می‌شود؛ نظیر تعیین و تعریف حد آستانه[3] برای فاصلهٔ معنایی مورداستفاده در خوشه‌بندی است، که می‌تواند بر تعداد و اندازهٔ خوشه‌ها تأثیرگذار باشد. اگر بتوان به طریقی امکان تعیین خودکار این مقدار توسط الگوریتم داشته باشیم، این مشکل مرتفع خواهد شد. از سوی دیگر در این مقاله به‌شکل مناسبی از ساختارهای سلسله‌مراتبی و رابطه والد-فرزند و نیز برادر-برادر یا هم‌زادی[4] بهره گرفته شده است؛ اما جای استفاده و شرکت‌دادن محتوای اسناد و درضمن تمایزدهی گره‌های صفت با سایر گره‌ها، خالی مانده است. مسئلهٔ مورد توجه دیگر محاسبه و یافتن فرکانس تمامی گره‌ها برای تخصیص درجهٔ اهمیت به آن‌ها با توجه به موقعیت گره و فرکانس رویداد هر گره است که به نظر می‌رسد به‌دلیل نیاز به بررسی تمامی سند XML پیچیدگی زمانی را بالا می‌برد.

روش پیشنهادی در (هوانگ و ریو ۲۰۱۰) نیاز به بررسی برچسب تمامی مسیرها و سپس محاسبهٔ وزن هریک دارد؛ پس زمان اجرای زیاد و پیچیدگی بالایی نیز دارد. از سوی دیگر، اگر شباهت را بین عنصرهای[5] موجود در شِمای[6] مطرح می‌کرد و کل ساختار درختی شِمای را مدنظر قرار می‌داد و از بررسی جزئیات موجود در سند جلوگیری به عمل می‌آمد و درنتیجه زمان اجرا بهبود و خوشه‌بندی با کیفیت‌تری را نتیجه می‌داد، همچنین اگر از تلفیق این روش با روش‌های تطبیق با محتوای اسناد استفاده می‌شد، دقت خوشه‌بندی افزایش پیدا می‌کرد.

الگوریتم پیشنهادی (نایاک و همکاران ۲۰۰۷) با درنظرگرفتن ساختار موجود در شِمای یعنی عنصرها و نام برچسب‌ها و سطح هر عنصر، استفادهٔ مناسبی از اطلاعات

[1] Level Similarity
[2] Level Edge
[3] Threshold
[4] Sibling
[5] Elements
[6] Schema

موجود در اسناد XML می‌نماید ولی اگر محتوای اسناد و صفات آن‌ها را نیز نادیده نمی‌گرفت، قاعدتاً نتایج بهتری را به‌دست‌می‌آورد. اگر نویسندهٔ مقاله می‌توانست به‌شکل مناسبی از الگوریتم خوشه‌بندی دوبه‌دویی[1] که از روش فاصلهٔ ویرایشی یعنی شباهت میان اسناد از روی هزینهٔ تبدیل یک سند به دیگری، استفاده می کرد، به نظر می‌رسد شاهد کیفیت بالاتر خوشه‌بندی بودیم. درضمن در این الگوریتم حد آستانهٔ تعیین خوشه‌ها که توسط کاربر تعیین می‌شود، نیز روی کیفیت و زمان اجرای الگوریتم مؤثر بوده و همان‌طور که پیش‌تر عنوان شد اگر بتوان به‌صورت خودکار و با توجه به اسناد در الگوریتم تولید شود، قطعاً شاهد نتایج چشم‌گیری خواهیم بود.

در (تا و همکاران ۲۰۰۷) به نظر می‌رسد، به‌دست‌آوردن خلاصهٔ ساختاری، یک روش جدید و مناسب برای افزایش سرعت و سهولت محاسبهٔ شباهت بین اسناد است، ولی اگر شباهت محتوا و صفات در اسناد را نیز به‌گونه‌ای در یافتن خلاصهٔ ساختاری دخیل می کرد، امیدِ حاصل‌شدن نتیجهٔ بهتر یا سریع‌تر وجود داشت. همچنین می‌توانست از ترکیب خوشه‌بندی سلسله‌مراتبی و غیرسلسله‌مراتبی با تولید خودکار حد آستانه در الگوریتم، برای خوشه‌بندی استفاده نماید تا از مزایای هر دو روش برای بیشینه‌کردن کیفیت و دقت خوشه‌بندی و با کمینهٔ زمان اجرا، بهره گیرد.

با توجه به‌این‌که روش‌های معرفی‌شدهٔ قبلی برای مجموعه‌هایی از اسناد که به لحاظ ساختار و محتوا ایستا می‌باشد، کارآیی خواهند داشت؛ بنابراین نیاز است، شرایطی برای اسناد XML که به‌صورت پویا تغییر می‌کنند نیز لحاظ شود. با توجه به مطالب عنوان‌شده، مقالهٔ (روسو و همکاران ۲۰۰۸) به‌دلیل توجه به ماهیت پویای اسناد XML و تفکری و ایده‌ای که برای آن در نظر گرفته نسبت به سایر مقالات پیشرو است. در این مقاله نیز حد آستانه‌ای برای تولید خوشه ها وجود دارد که باز هم توسط کاربر تعیین می‌شود و همان‌طورکه بارها در این نوشتار شرح داده شد، چالش‌برانگیز است. به‌طوری‌که برای انجام آزمایش‌ها مقدار آن ثابت در نظر گرفته شده است. درضمن در این مقاله نیز محتوا و صفات اسناد XML نادیده گرفته شده است.

در (لی و همکاران ۲۰۰۲) برای اسنادی که از یک تعریف نوع سند[2] یا DTD تشکیل شده‌اند، روش خوشه‌بندی را معرفی می‌کند و برای اسنادی که از چند DTD متفاوت ساخته شده‌اند، ایده‌ای ندارد. به عبارت دیگر این روش فقط برای مجموعه اسناد همگن قابل استفاده است. بنابراین با وجود گستردگی منابع تولید اسناد XML، الگوریتم پیشنهادی این مقاله کارآیی چندانی نخواهد داشت. درضمن اگر می‌توانست از شکل خلاصه‌شدهٔ اسناد یا درجهٔ اهمیت هریک از اجزای سند پیش از اندازه‌گیری شباهت میان اسناد بهره‌گیرد، سرعت اجرای بیشتر و کارآیی بهتر الگوریتم حاصل می‌شد.

در مقالهٔ (تران و همکاران ۲۰۰۸) از ترکیب هر دو ویژگی ساختاری و محتوایی اسناد XML برای خوشه‌بندی استفاده شده است. محتوا و ساختار اسناد به‌صورت مجزا و با استفادهٔ مدل فضای برداری بازنمایی می‌شود؛ سپس شباهت بین محتوا و ساختار دو سند، جداگانه محاسبه شده و درنهایت معیار شباهت کل، از ترکیب وزنی این دو مقدار به‌دست‌می‌آید. در این روش به‌دلیل استفادهٔ محتوای اسناد، نتایج خوشه‌بندی دارای دقت بیشتر است؛ اما ترکیب وزنی دو مقدار شباهت به‌دست آمده، برای هر مجموعه اسناد ورودی نیاز به تنظیم مجدد دارد. از طرفی، استفاده از روش خوشه‌بندی دوبه‌دویی دارای پیچیدگی زمانی بالایی است.

با مطالعه و بررسی هرچه بیشتر مقالات، این تصور به وجود می‌آید که به‌درستی کدام روش بهتر عمل می‌کند و آیا نادیده‌گرفتن برخی از خصوصیات اسناد نظیر محتوا، نام برچسب‌ها، ساختار سلسله‌مراتبی، رابطهٔ والد-فرزند یا برادر-برادر، فرکانس رویداد هر کلمه در محتوای سند، ساختارهای تکراری یا تودرتو و یا درنظرگرفتن یک DTD برای کل اسناد، باعث می‌شود ما منحرف شده و نتوانیم به‌درستی یا در کمترین زمان ممکن اسناد را خوشه‌بندی کنیم. با این وجود چند مسئله روشن است؛ نخست استفادهٔ تنها از یک ویژگی اسناد XML درحالی‌که ویژگی‌های دیگر نادیده گرفته می‌شود، مناسب نیست. دوم این‌که یافتن مسیرهای مهم در ساختار درختی اسناد، به‌دست‌آوردن خلاصهٔ ساختاری با کاستن تکرارها و تودرتویی‌ها و یا یافتن گره‌هایی با نام مشابه که به یک موجودیت تعلق دارند، باعث بهبود عمل تطبیق و اندازه‌گیری مشابهت و موجب می‌شود نتایج رضایت‌بخشی در کیفیت و دقت خوشه‌بندی به‌دست آید.

[1] Pairwise
[2] Document Type Definition

۳- تعاریف اولیه

در ادامه برخی تعاریف اولیهٔ مورد نیاز برای روش پیشنهادی این مقاله، مطرح شده است.

۳-۱- پارامتر TF-IDF

یکی از پرکاربردترین روابط در حوزهٔ بازیابی اطلاعات، پارامتر TF-IDF است (مانینگ و همکاران ۲۰۰۸)، که از حاصل‌ضرب فراوانی کلمه در فراوانی معکوس سند[1] به‌دست می‌آید. این روش، یک روش مبتنی بر چندسندی است که در آن منظور از فراوانی کلمه، تعداد تکرار کلمه در یک سند خاص است. همچنین منظور از فراوانی معکوس سند، تعداد اسنادی است که یک کلمهٔ خاص در آنها موجود بوده است. دلیل مقبولیت این روش را نسبت به سایر روش‌ها می‌توان با توجه به سهولت در استفاده از این روش، محاسبات کم و نتایج قابل قبول دانست. فراوانی معکوس سند از رابطهٔ (۱) و TF-IDF طبق رابطهٔ (۲) محاسبه می‌شود که n مجموع تعداد اسناد مورد بررسی است.

$$IDF_i = log\frac{n}{DocFrequency_i} + 1 \qquad (۱)$$

که در آن $DocFrequency_i$ تعداد اسنادی است که عبارت iام در آن آمده است.

$$Weight_{ij} = TF_{ij} \bullet IDF_{ij} \qquad (۲)$$

که در آن $Weight_{ij}$ وزن عبارت i ام در سند j ام، TF_{ij} فراوانی عبارت i ام در سند jام، و IDF_{ij} فراوانی معکوس سند عبارت i ام در سند jام است.

۳-۲- نرمال‌سازی Pivoted

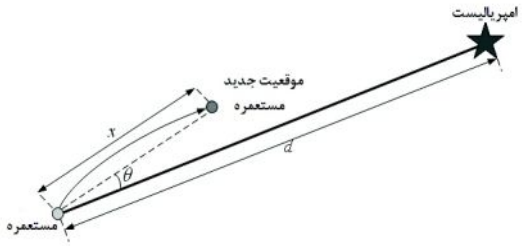
با توجه به این‌که اسناد دارای طول‌های متفاوتی هستند، انتخاب یک حد آستانه برای حذف کلمات با وزن پایین، مشکل است، زیرا هر چه حد آستانه را افزایش دهیم، کارآیی برای اسناد با طول کوتاه، کمتر می‌شود و برعکس هر چه حد آستانه را کاهش دهیم، کارآیی برای اسناد با طول زیاد، کمتر می‌شود. درنتیجه با استفاده از روش نرمال‌سازی طول سند، نتایج بهتری به‌دست می‌آید. یکی از روش‌های نرمال‌سازی Pivoted است (مانینگ و همکاران ۲۰۰۸). وزن عبارت سند با استفاده از پارامتر TF-IDF و نرمال‌سازی Pivoted از رابطهٔ (۳) به‌دست‌می‌آید.

$$Weight_{ij} = \frac{TF_{ij} \bullet IDF_{ij}}{(1-b) + b.\frac{DocLenght_j}{AvgLenght}} \qquad (۳)$$

که در $Weight_{ij}$ وزن عبارت i ام در سند j ام، TF_{ij} فراوانی عبارت i ام در سند jام و IDF_{ij} فراوانی معکوس سند عبارت i ام در سند jام، $DocLenght_j$ طول سند j و AvgLenght میانگین طول اسناد مجموعه است. در این فرمول b پارامتر نرمال‌سازی است که مقدار آن به طورمعمول ۰.۲ قرار داده می‌شود.

۳-۳- الگوریتم رقابت استعماری

الگوریتم رقابت استعماری یک الگوریتم جدید در زمینهٔ محاسبات تکاملی[2] است که نخستین‌بار توسط «آتش‌پز و لوکاس» معرفی شده است (آتش پز و لوکاس ۲۰۰۷). روندنمای این الگوریتم را در شکل (۲) نشان داده است. همانند دیگر الگوریتم‌های تکاملی، این الگوریتم نیز با تعدادی جمعیت اولیهٔ تصادفی که هرکدام از آن‌ها یک کشور[3] نامیده می‌شوند، شروع می‌شود. تعدادی از بهترین عناصر جمعیت، به‌عنوان امپریالیست[4] انتخاب و باقیمانده نیز به‌عنوان مستعمره[5]، در نظر گرفته می‌شوند. برای شروع الگوریتم، تعداد $N_{Country}$ کشور اولیه را ایجاد می کند. N_{imp} تا از بهترین اعضای این جمعیت، یعنی کشورهای دارای کمترین مقدار تابع هزینه را به‌عنوان امپریالیست انتخاب می‌کنیم. باقیمانده N_{col} تا از کشورها، مستعمره‌هایی را تشکیل می دهند که هر کدام به یک امپراطوری[6] تعلق دارند. کشورهای استعمارگر با اعمال سیاست جذب در راستای محورهای مختلف بهینه‌سازی، کشورهای مستعمره را به سمت خود می‌کشند. شکل (۱)، شمای کلی این حرکت را نشان می‌دهد.



(شکل- ۱): حرکت مستعمرات به سمت امپریالیست

کشور مستعمره، به اندازه‌ی x واحد در جهت خط واصل مستعمره به استعمارگر، حرکت کرده و به موقعیت جدید کشانده می‌شود. فاصلهٔ میان استعمارگر و مستعمره با

[1] Inverse Document Frequency

[2] Evolutionary Computation
[3] Country
[4] Imperialist
[5] Colony
[6] Empire

d نشان داده شده و x نیز عددی تصادفی با توزیع یکنواخت (۴) و یا هر توزیع مناسب دیگر، است. یعنی مقدار x تقریبا برابر است با:

$$x \approx U(0, \beta \times d) \qquad (۴)$$

که در آن β عددی بزرگ‌تر از یک و نزدیک به دو است. وجود ضریب β>۱ باعث می‌شود تا کشور استعمارگر، از جهت‌های مختلف به آن نزدیک شود. در الگوریتم رقابت استعماری با یک انحراف احتمالی، مستعمره در مسیر جذب استعمارگر پیش می‌رود. این انحراف با زاویهٔ θ نشان داده شده است. که به‌صورت تصادفی و با توزیع یکنواخت انتخاب می‌شود، که این زاویهٔ حرکت به‌صورت توزیع یکنواخت (۵) در نظر گرفته شده است.

$$\theta \approx U(-\gamma, \gamma) \qquad (۵)$$

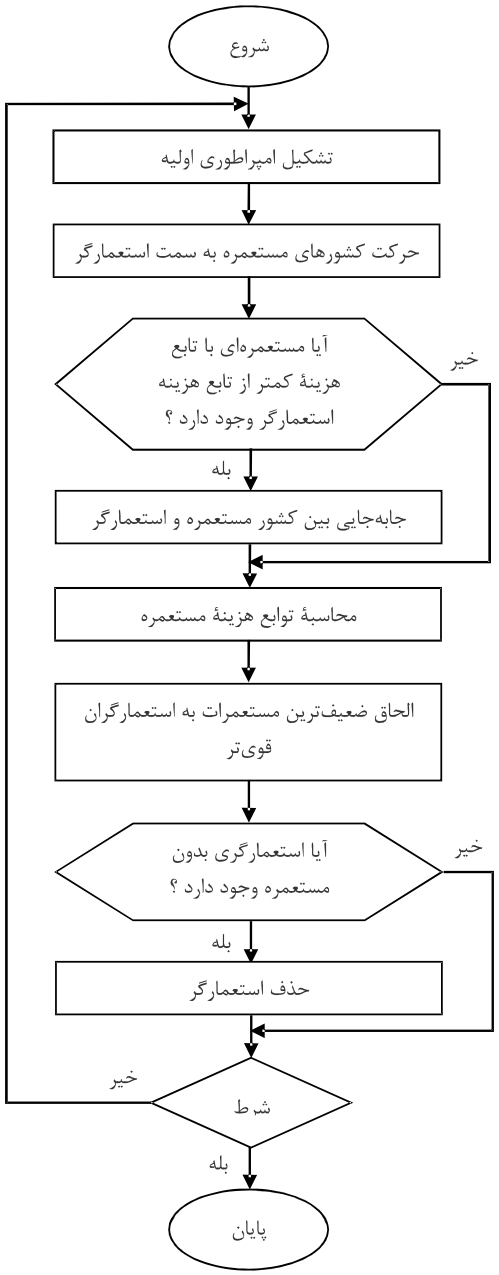
در حین حرکت مستعمرات به سمت کشور استعمارگر، ممکن است بعضی از این مستعمرات به موقعیتی بهتر از استعمارگر برسند؛ در این حالت کشور استعمارگر و کشور مستعمره جای خود را با هم عوض می‌کنند. قدرت کل هر امپراطوری، با محاسبهٔ قدرت هر دو بخش تشکیل‌دهندهٔ آن یعنی قدرت کشور استعمارگر، به اضافهٔ درصدی از میانگین قدرت مستعمرات آن، تعیین می‌شود.

رقابت استعماری، بخش مهم دیگری از این الگوریتم را تشکیل می‌دهد. هر امپراطوری که نتواند بر قدرت خود بیفزاید و قدرت رقابت خود را از دست بدهد، در جریان رقابت‌های امپریالیستی، حذف خواهد شد. این حذف‌شدن، به‌صورت تدریجی صورت می‌پذیرد. بدین معنی که به‌مرور زمان، امپراطوری‌های ضعیف، مستعمره‌های خود را از دست داده و امپراطوری های قوی‌تر، این مستعمرات را تصاحب کرده و بر قدرت خویش می‌افزایند. الگوریتم موردنظر تا برآورده‌شدن یک شرط هم‌گرایی و یا تا اتمام تعداد کل تکرارها، ادامه می‌یابد. پس از مدتی، همهٔ امپراطوری‌ها، سقوط کرده و تنها یک امپراطوری باقی می‌ماند و بقیهٔ کشورها تحت کنترل این امپراطوری واحد، قرار می‌گیرند.

۴- روش پیشنهادی

با الهام از گام‌های خوشه بندی داده‌ها، روش پیشنهادی را می‌توان به چهار مرحلهٔ اساسی پیش‌پردازش[1]، بازنمایی داده[2]، محاسبهٔ شباهت[3] و خوشه‌بندی تقسیم کرد. شکل (۳) نمای کلی از روش خوشه‌بندی پیشنهادی را نشان می‌دهد.

[1] Pre Proccessing
[2] Data Representation



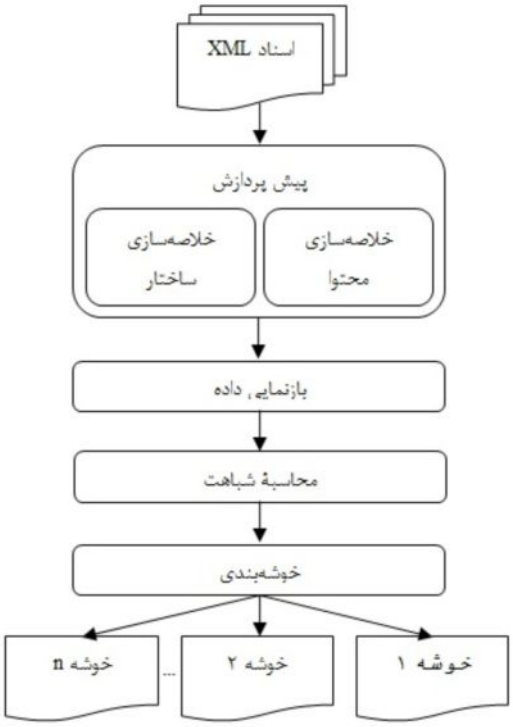
(شکل- ۲): روندنمای الگوریتم رقابت استعماری

۴-۱- پیش‌پردازش

ابتدا با استفاده از ساختار درختی اسناد XML، داده‌های ساختاری و محتوایی موجود در هریک از آن‌ها استخراج و در سه جدول با نام‌های سند، مسیر و برگ ذخیره می‌شوند که در شکل (۴) نشان داده شده است. جدول سند نشانی فیزیکی تمامی سندها را نگه‌داری می‌کند. جدول مسیر شامل دنبالهٔ اسامی عناصر، از برچسب ریشه به برچسب برگ

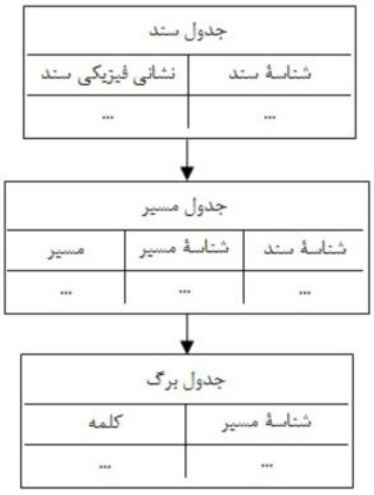
[3] Similarity Computation

موجود در هر سند است. در جدول برگ نیز کلمه‌های موجود در محتوای متنی هر سند XML که وابسته به یک مسیر است، نگه‌داری می‌شود. این کار موجب سرعت‌یافتن عمل تطبیق و اندازه‌گیری مشابهت در مراحل بعد می‌شود. در ادامه با استفاده از این جدول‌ها، اطلاعات ساختاری و محتوایی اسناد خلاصه‌سازی می‌شود. هدف از این‌کار، کاهش حجم پردازش و ابعاد مدل فضای ماتریسی است که در مرحلهٔ دوم ایجاد می شود. خلاصه‌سازی ساختار که ساده‌تر است، شامل حذف مسیرهای تکراری از ساختار اسناد است؛ اما خلاصه سازی محتوا که پیچیده تر است، شامل استخراج کلمات کلیدی از محتوای متنی اسناد است.



(شکل- ۳): فرآیند خوشه‌بندی اسناد XML در روش پیشنهادی

مراحل خلاصه سازی محتوای اسناد XML در شکل (۵) نشان داده شده است. در گام نخست کلمه‌های عمومی[1] حذف می‌شوند، این عمل با یک مقایسهٔ سادهٔ کلمه‌ها با یک فهرست از قبل آماده انجام می‌گیرد. بعضی از کلمات مثل ضمایر، قیود، حروف اضافه و ربط و بعضی از افعال پرتکرار در همهٔ متون با فراوانی زیاد وجود دارند که ارزش محتوایی ندارند. به این کلمات، کلمات عمومی گفته می‌شود. با حذف کلمات عمومی فارسی از میزان محاسبات کم شده و کارآیی روش‌ها نیز بیشتر می‌شود.



(شکل- ۴): ساختار ذخیره‌سازی هر سند XML در مرحلهٔ پیش‌پردازش

گام دوم ریشه‌یابی[2] کلمه‌ها است، یکی از مهم‌ترین کارها در استخراج واژگان کلیدی از متون فارسی، ریشه‌یابی کلمه‌ها است. هدف از ریشه‌یابی، حذف اضافات از کلمه و رسیدن به ریشهٔ اصلی کلمه است. بدین‌منظور در این مقاله از روش مبتنی بر قاعده و بر پایهٔ حذف طولانی‌ترین پسوندها و پیشوندهای ممکن از کلمه‌ها، استفاده شده است. برای جلوگیری از ایجاد ریشه‌های نادرست نیز، از کلمات موجود در پیکرهٔ فارسی بی‌جن‌خان استفاده شده است (بی جن خان ۲۰۰۶). در گام سوم، وزن‌دهی به کلمات بر اساس اهمیت آن‌ها در محتوای سند انجام می‌گیرد که برای این کار از پارامتر TF-IDF به همراه نرمال‌سازی Pivoted استفاده شده است. در گام آخر با اعمال حدّ آستانه روی وزن کلمات، فهرست کلمه‌های کلیدی استخراج می‌شود.

۴-۲- بازنمایی داده

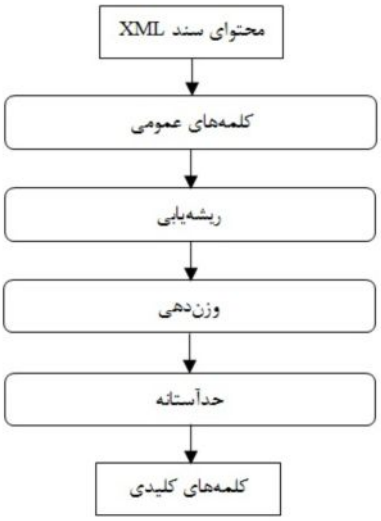
همان‌طورکه گفته شد، بازنمایی اسناد XML تنها با استفاده از ویژگی‌های ساختاری برای مدل‌سازی مؤثر آنها کافی نیست. برای این کار باید، هر دو ویژگی محتوا و ساختار در نظر گرفته شود. بدین‌منظور مدل جدیدی به نام مدل فضای ماتریسی را تعریف می‌کنیم که تصویری از اطلاعات محتوایی و ساختاری اسناد XML است. مدل فضای ماتریسی، توسعه یافتهٔ مدل فضای برداری[3] است که به‌طور گسترده

[1] Stop Words

[2] Steamming

[3] Vector Space Model

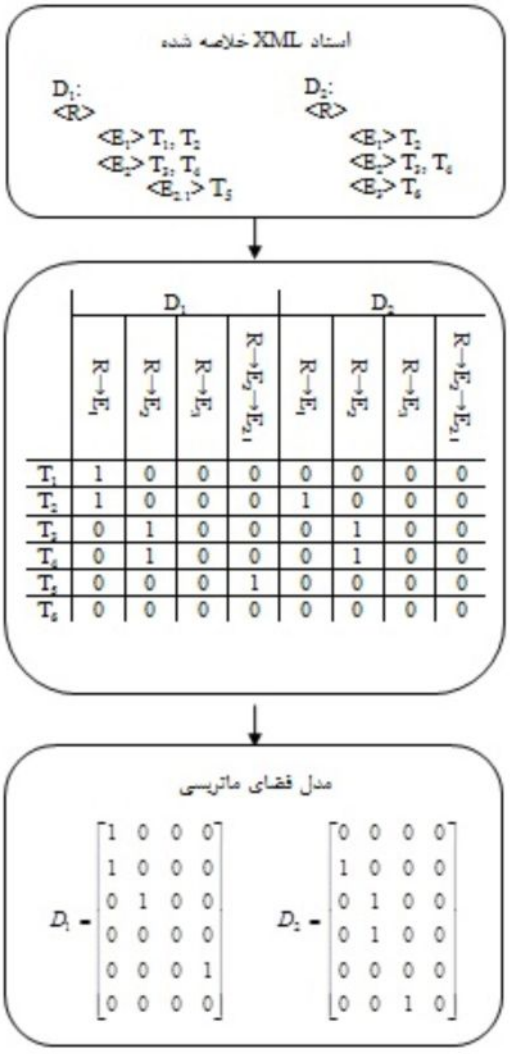
برای بازنمایی داده در اسناد متنی استفاده می‌شود (مانینگ و همکاران ۲۰۰۸).



(شکل- ۵): فرآیند خلاصه‌سازی محتوای سند XML در مرحله پیش پردازش

در مدل فضای برداری، هر سند به‌صورت برداری از کلمه ها نشان داده می‌شود. هر بعد در این بردار، منطبق بر یکی از کلمه‌های متن است. تعداد ابعاد بردار متن مبتنی بر تعداد تمام کلمه‌های مجزای موجود در کل مجموعه یک متن است. به ازای هر کلمه، اگر کلمهٔ مورد نظر در آن متن وجود نداشته باشد، در موقعیت منطبق بر آن در بردار متن، مقدار صفر و یا غیر صفر قرار داده می‌شود. این مقدار بسته به کاربرد می‌تواند یک باشد، که نشان دهندهٔ وجود کلمهٔ موردنظر در متن است و یا می‌تواند مقداری باشد که نشان‌دهندهٔ وزن آن کلمه در متن مورد نظر است.

شکل (۶) مثالی از مدل فضای ماتریسی را نشان می‌دهد. در مدل هر سند به‌صورت ماتریسی از مسیرها و کلمه‌های وابسته به آن، نشان داده می‌شود. هر ستون در این ماتریس منطبق بر یکی از مسیرهای موجود در ساختار درختی اسناد (R) است و هر سطر آن منطبق بر یکی از کلمه‌های کلیدی موجود در محتوای اسناد (T) است. به ازای هر کلمه، اگر کلمه مورد نظر در محتوای برگ وابسته به یک مسیر وجود داشته باشد در موقعیت منطبق بر آن در ماتریس سند، مقدار یک و در غیر این صورت مقدار صفر قرار داده می‌شود. درنهایت هر سند به یک ماتریس با مقادیر ۰ و ۱ تبدیل می‌شود تا در مرحلهٔ بعد به‌عنوان ورودی به الگوریتم خوشه‌بندی داده شود.



(شکل- ۶): مثالی از مدل فضای ماتریسی

۴-۳- محاسبهٔ شباهت

در این مرحله توابع نزدیکی برای اندازه گیری مشابهت بین جفت‌های اشیای داده تعیین می‌شود؛ سپس اسناد XML با توجه به شباهت در ویژگی‌های استخراج شده، گروه‌بندی می‌شوند. کارآیی راه‌حل خوشه‌بندی به‌طور عمده بر معیار سنجش مشابهت به‌کارگرفته‌شده، بستگی دارد. بر اساس نوع مدل داده که برای نمایش داده های XML مورد استفاده قرار می گیرد، انواع مختلفی از معیارهای شباهت وجود دارد که در اینجا از معیار شباهت جاکارد[1] استفاده شده است (مانینگ و همکاران ۲۰۰۸) که به‌صورت رابطهٔ (۶) تعریف می‌شود. از این رابطه معیار فاصلهٔ جاکارد به‌دست‌می‌آید، که به‌صورت رابطهٔ (۷) تعریف می‌شود.

[1] Jaccard

$$Jaccard\ Similarity = \frac{|x \cap y|}{|x \cup y|} = \frac{a}{a+b+c} \quad (۶)$$

$$Jaccard\ \text{Distance} = 1 - Jaccard\ Similarity \quad (۷)$$

که در آن $x=(x_1,x_2,\ldots,x_n)$ و $y=(y_1,y_2,\ldots,y_n)$ دو بردار می‌باشد و a تعداد دفعاتی است که $x_i=1$ و $y_i=1$، b تعداد دفعاتی است که $x_i=0$ و $y_i=1$، c تعداد دفعاتی است که $x_i=1$ و $y_i=0$ است، مقدار تولیدشده توسط شباهت جاکارد عددی بین صفر و یک است که هر چه به یک نزدیک‌تر شویم، شباهت افزایش و فاصله کاهش و برعکس هرچه به صفر نزدیک‌تر شویم، شباهت کاهش و فاصله افزایش پیدا می‌کند. به‌عنوان مثال برای دو ماتریس (شکل ۶)، شباهت ماتریسی برابر با ۰.۳۳ است.

۴-۴- خوشه‌بندی

در این مرحله اسناد XML که با یکدیگر مشابه هستند بر اساس یک تابع نزدیکی و با استفاده از یک الگوریتم خوشه‌بندی مناسب گروه‌بندی می‌شوند. الگوریتم‌های خوشه‌بندی به‌طور ضمنی یا صریح به معیارهای شباهت به کار گرفته شده، مرتبط هستند. الگوریتم های خوشه بندی اسناد XML را می‌توان به دو گروه افزایشی و دوبه‌دویی[1] تقسیم‌بندی کرد.

در الگوریتم های مبتنی بر روش‌های دوبه‌دو که متداول‌ترند، ابتدا یک ماتریس شباهت به ازای کل اسناد XML ایجاد و سپس توسط مقیاسی برای اندازه‌گیری شباهت بین هر دو سند، پر می‌شود. درنهایت پس از تکمیل ماتریس، می‌توان یکی از روش‌های متداول خوشه بندی را مورد استفاده قرار داد و سندها را بر اساس میزان شباهت‌شان در خوشه‌ها جای داد؛ بر عکس در روش‌های افزایشی به‌ازای وارد شدن هر سند XML میزان شباهت سند با خوشه های موجود محاسبه می شود؛ درصورتی‌که میزان شباهت از یک حدّ آستانهٔ تعریف‌شده توسط کاربر بیشتر باشد، سند ورودی در خوشهٔ مورد نظر جای می‌گیرد درغیراین‌صورت یک خوشهٔ جدیدی ایجاد می شود و در آن قرار می‌گیرد.

به‌دلیل مشکلات الگوریتم های دوبه‌دویی، یعنی پیچیدگی زمانی بالا به‌دلیل استفاده از ماتریس شباهت اسناد و الگوریتم‌های افزایشی، یعنی دقت خوشه‌بندی پایین به‌دلیل حساسیت به‌ترتیب اسناد ورودی و حدّ آستانه تعریف شده، لذا در این مقاله برای نخستین بار از رویکرد فراابتکاری برای خوشه‌بندی استفاده می‌شود. به‌طوراصولی الگوریتم‌های فراابتکاری نوعی از الگوریتم های تصادفی هستند که با رویکرد یافتن پاسخ بهینه مورد استفاده قرار می‌گیرند. از بین الگوریتم‌های بهینه‌سازی فراابتکاری، الگوریتم رقابت استعماری به‌دلیل عملکرد مناسب خود در خوشه‌بندی داده‌ها انتخاب شد *(Niknama et al., 2011)*؛ اما در ابتدا نیاز است خوشه بندی به‌عنوان یک مسئلهٔ بهینه سازی مدل شود. هدف در خوشه‌بندی، کمینه کردن فاصلهٔ داده‌ها از مراکز خوشه‌ها است، درنتیجه تابع هزینه رابطهٔ (۸) تعریف می‌شود.

$$CostFunc = \sum_{i=1}^{n}\sum_{j=1}^{k} Jaccard\ \text{Distance}(M_i, C_j) \quad (۸)$$

که در آن M_i ماتریس‌های ورودی است که متناظر با یک سند XML است و C_j مراکز خوشه‌ها است، که توسط الگوریتم رقابت استعماری به‌عنوان جمعیت اولیه و به‌صورت تصادفی تولید می‌شود. همچنین n تعداد ماتریس‌ها یا اسناد ورودی و k تعداد خوشه‌ها است. با کمینه‌کردن این تابع هزینه، درعمل به هدف نهایی خود یعنی خوشه‌بندی اسناد نایل می‌شویم. در این روش همانند روش خوشه‌بندی دوبه‌دویی، دیگر نیاز به تشکیل ماتریس مشابهت اسناد نیست که موجب افزایش پیچیدگی می‌شود. از طرف دیگر مانند روش افزایشی، به حدّ آستانه برای ایجاد خوشه‌ها نیاز ندارد و نسبت به ترتیب ورودی‌ها نیز حساس نیست.

۵- ارزیابی و تحلیل نتایج

این بخش شامل جزئیات داده‌های آزمایش‌شده، روش‌های ارزیابی، نتایج به‌دست‌آمده از پیاده سازی و بحث و بررسی دربارهٔ این نتایج است.

۵-۱- مجموعه داده

مجموعهٔ اسناد همشهری با خزش وب‌سایت همشهری و چندین مرحلهٔ پیش‌پردازش و برچسب‌گذاری حاصل آمده است (آلاحمد و همکاران ۲۰۰۹). نسخه ۱ این مجموعه نمونه‌ای است که در همایش‌های CLEF در سال‌های ۲۰۰۸ و ۲۰۰۹ مورد استفاده قرار گرفته است. نسخهٔ دو، آخرین نسخهٔ مجموعه است که نسبت به نسخهٔ یک جامع‌تر می‌باشد. این مجموعه داده دارای نه طبقهٔ اصلی، شامل «ادب و هنر»، «اجتماعی»، «علمی فرهنگی»، «اقتصاد»، «گوناگون»، «سیاسی»، «ورزش»، «محیط زیست» و

[1] Pairwise

«گردشگری» است. با توجه به این‌که این دو نسخه به لحاظ ساختاری با یکدیگر متفاوتند بنابراین برای افزایش ناهم‌گونی در ساختار و محتوای مجموعه داده، مجموعهٔ مقاله‌های خبری دو سال مختلف، متعلق به هرکدام از نسخه ها، به‌صورت تصادفی انتخاب شد.

۵-۲- معیار ارزیابی

برای ارزیابی نتایج از پارامتر $F_{measure}$ استفاده شده است. این پارامتر درحقیقت میانگین هارمونیک پارامترهای دقت و بازخوانی است (مانینگ و همکاران ۲۰۰۸). شیوهٔ محاسبهٔ پارامترهای دقت، بازخوانی و $F_{measure}$ به‌ترتیب در روابط (۹)، (۱۰) و (۱۱) نشان داده شده است.

$$Precision = \frac{TP}{TP + FP} \quad (۹)$$

$$Recall = \frac{TP}{TP + FN} \quad (۱۰)$$

$$F_{mesure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (۱۱)$$

با توجه به خوشهٔ C_i، TP تعداد اسناد خوشهٔ C_i است که مشابه هستند؛ یعنی خوشه‌های صحیح، FP تعداد اسناد خوشهٔ C_i است که شبیه نیستند؛ یعنی خوشه‌های اشتباه، FN نیز تعدادی اسنادی است که در C_i نیستند؛ اما باید باشند. واضح است که هرچه $F_{measure}$ به یک نزدیک‌تر باشد، خوشه‌بندی بهتری صورت گرفته است.

۵-۳- ویژگی‌های آزمایش

همان‌طورکه بیان شد، از مجموعه همشهری، دوسری پرس‌وجوی استاندارد در سال‌های ۲۰۰۸ و ۲۰۰۹ تهیه شد. هر سری از پرس‌وجوها شامل پنجاه موضوع به دو زبان انگلیسی و فارسی بوده که توسط ۲۵ کاربر، ساخته و پرداخته شده است و معیاری برای استفاده از کلمات کلیدی در آزمایش به‌شمار می‌رود.

خوشه‌های انتخابی در آزمایش با اعداد ثابت دو، نه و هجده مورد استفاده قرار گرفته‌اند. علت انتخاب خوشه‌ها به این دلیل است که از دو نسخهٔ همشهری یک و دو معادل دو خوشه استفاده کرده‌ایم. هر خوشه نیز از نظر موضوعی به نه دسته تقسیم‌بندی می‌شود. از دیدگاه محتوایی و فارغ از انتخاب نسخه‌های همشهری نیز به‌طور کلی با هجده موضوع متمایز در دو نسخه مواجه هستیم. این عوامل، انگیزهٔ اصلی در انتخاب خوشه‌ها هستند.

از نقطه‌نظر کمی، تعداد اسناد مورد استفاده در نسخهٔ یک همشهری بالغ بر ۱۶۰ هزار سند و در نسخهٔ دو همشهری ۳۱۸ هزار سند متنی بوده که مجموعه تصاویر آن از کل متن جدا شده است. حجم یونی‌کدهای مورد استفاده در قالب داده‌های نسخهٔ یک، ۷۰۰ مگابایت و در نسخهٔ دو در حدود ۱۴۰۰ مگابایت است. بازهٔ زمانی اسناد نسخهٔ یک از سال ۱۳۷۵ الی ۱۳۸۱ و در نسخهٔ دو به سال‌های بین ۱۳۷۵ تا ۱۳۸۶ مربوط می‌شود.

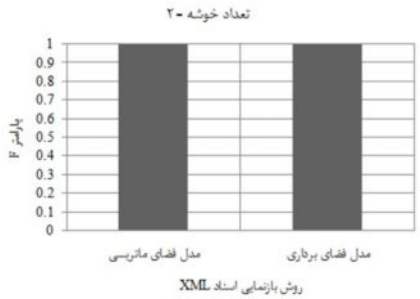
این اسناد همگی در قالب فایل‌های XML و دارای طبقه‌بندی مشخص هستند. در نسخهٔ دوم، پیوند به صفحات وب و تصاویر نیز افزوده شده است. در هر دو نسخه، قابلیت پرس‌وجو و داوری ارتباط برای مخاطبان فراهم خواهد بود.

در روش خوشه‌بندی، ترکیبی از الگوریتم رقابت استعماری ICA برای استخراج مراکز اولیه در الگوریتم K-Means مورد استفاده قرار می‌گیرد. در ابتدا، با تعداد تکرارهای محدود به جوابی نزدیک به جواب بهینه، هم‌گرا می‌شویم. بعد از به‌دست‌آمدن مراکز توسط ICA، الگوریتم K-Means از آن‌ها برای شروع فرآیند خوشه‌بندی بهینه استفاده خواهدکرد.

مراحل پیاده‌سازی نیز به‌طور خلاصه شامل: تولید جمعیت اولیه، ارزیابی تابع هدف، مرتب‌سازی جمعیت اولیه بر اساس تابع هدف، شکل‌گیری امپراطوری‌ها، تقسیم‌بندی مستعمرات، اجرای الگوریتم K-means روی هر امپراطوری، حرکت مستعمره به سمت استعمارگر، اجرای عملکرد جهش، مقایسهٔ تابع هدف مستعمرات، مقایسهٔ مقدار تابع هدف، رقابت بین امپراطوری‌ها، نابودی ضعیف‌ترین امپراطوری و آزمایش تعداد امپراطوری‌ها است.

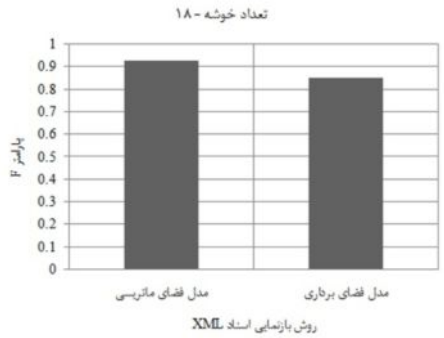
۵-۴- نتایج تجربی

نتایج خوشه‌بندی روش پیشنهادی در شکل (۷)، شکل (۸) و شکل (۹) نشان داده شده است. در این نمودارها محور افقی روش بازنمایی مورد استفاده و محور عمودی مقدار پارامتر $F_{measure}$ را نشان می‌دهد. این مقایسه‌ها به‌خوبی بیان‌گر این واقعیت است که میزان کارآیی روش پیشنهادی به‌مراتب بهتر از روش‌هایی است که تنها از یک ویژگی اسناد XML برای خوشه‌بندی استفاده می‌کنند.



(شکل- ۷): مقایسهٔ کارآیی روش خوشه‌بندی پیشنهادی و خوشه‌بندی ساختاری

(شکل- ۸): مقایسهٔ کارآیی روش خوشه‌بندی پیشنهادی و خوشه‌بندی محتوایی



(شکل- ۹): مقایسهٔ کارآیی روش خوشه‌بندی پیشنهادی و خوشه‌بندی ترکیبی ساختاری و محتوایی

در تمامی نتایج به‌دست‌آمده، از دو روش پیاده‌سازی استفاده شده است. یکی روش SCMSM است که مبتنی بر شباهت ساختاری و محتوایی با استفاده از مدل فضای ماتریسی ارائه‌شده در این مقاله است. روش دوم، روش SCVSM مبتنی بر ترکیب شباهت ساختاری و محتوایی با استفاده از مدل فضای برداری و تنها استفاده از K-Means است که در (ژانگ و همکاران ۲۰۰۸) آمده است.

در شکل (۷) مقایسهٔ دقت روش پیشنهادی و روش خوشه‌بندی ساختاری نشان داده شده است. از آنجاکه اسناد به‌لحاظ ساختاری، یا متعلق به نسخهٔ یک و یا متعلق به نسخهٔ دو مجموعهٔ همشهری است، بنابراین مجموعه اسناد همشهری به‌لحاظ ساختاری به دو خوشه تقسیم می‌شود. هنگامی‌که تنها از شباهت ساختاری یعنی مسیرهای مشابه در ساختار درختی اسناد، مبتنی بر مدل فضای برداری برای خوشه‌بندی استفاده می‌شود، پارامتر $F_{measure}$ در هر دو روش برابر یک می‌شود که بهترین حالت است.

در شکل (۸) مقایسه دقت روش پیشنهادی و روش خوشه‌بندی محتوایی نشان داده شده است. از آنجاکه مجموعهٔ اسناد همشهری از نظر محتوا شامل نه طبقه می‌باشد، بنابراین مجموعهٔ اسناد همشهری به‌لحاظ محتوایی به نه خوشه تقسیم می‌شوند. هنگامی‌که تنها از شباهت محتوایی مبتنی بر مدل فضای برداری برای خوشه‌بندی استفاده می‌شود، پارامتر $F_{measure}$ آن نسبت به روش پیشنهادی کمتر است؛ زیرا در روش پیشنهادی، استخراج واژگان کلیدی از محتوای اسناد که در مرحلهٔ پیش‌پردازش صورت می‌گیرد، باعث بهبود مرحله محاسبه شباهت و شناسایی اسناد مشابه در مرحله خوشه‌بندی می‌شود. با افزایش دقت ریشه‌یابی در مرحلهٔ پیش‌پردازش می‌توان روش پیشنهادی را بهبود بخشید.

در شکل (۹) مقایسه دقت روش پیشنهادی و روش خوشه‌بندی ترکیبی ساختاری و محتوایی نشان داده شده است. از آنجاکه مجموعه اسناد همشهری به‌لحاظ ساختاری به دو خوشه و محتوایی به نه خوشه تقسیم می‌شود، بنابراین مجموعه اسناد همشهری به‌لحاظ ساختاری و محتوایی به هجده خوشه تقسیم می‌شود. هنگامی که تنها از ترکیب وزنی شباهت ساختاری و شباهت محتوایی، مبتنی بر مدل فضای برداری برای خوشه‌بندی استفاده می‌شود، پارامتر $F_{measure}$ کمتر است؛ زیرا به‌کارگیری ترکیب وزنی شباهت، نیاز به تنظیم پارامترهای وزن‌دهی دارد که شناسایی این پارامترها خود یک چالش است و پیچیدگی را افزایش می‌دهد و برای هر مجموعه از اسناد این ضرایب وزنی ممکن است متفاوت باشد. همچنین به‌دلیل تفکیک ساختار از محتوای، وابستگی بین محتوا و ساختار اسناد در نظر گرفته نمی‌شود؛ اما در روش پیشنهادی که از مدل فضای ماتریسی مبتنی بر شباهت ساختاری و محتوایی اسناد XML استفاده

شده، هنگامی که تعداد خوشه هجده است، مقدار پارامتر $F_{measure}$ به مقدار مطلوب یا همان یک بسیار نزدیک شده است که نشان از دقت خوشه‌بندی این روش دارد.

۶- نتیجه‌گیری و ادامهٔ کار

در این مقاله مدل جدیدی به نام مدل فضای ماتریسی که توسعه‌یافتهٔ مدل فضای برداری است، برای بازنمایی اسناد XML پیشنهاد شد. این مدل مبتنی بر هر دو ویژگی ساختاری و محتوایی اسناد XML است؛ سپس از ترکیب معیار شباهت جاکارد با رویکرد فرابتکاری در فرآیند خوشه‌بندی استفاده شد. نتایج تجربی نشان می‌دهد، استفادهٔ هم‌زمان از اطلاعات ساختاری و محتوایی اسناد XML در افزایش دقت خوشه بندی مؤثر است. در آینده می‌توان از راه‌بردهای معنایی و معیارهای شباهت مناسب برای بهبود روش پیشنهادی استفاده کرد. همچنین با توجه به‌این‌که در این مقاله فرض شده است که مجموعه اسناد XML ایستا هستند، می‌توان چارچوبی برای خوشه‌بندی اسناد XML پویا ارائه کرد. اسناد XML پویا، اسنادی است که در طول زمان دچار تغییر و به‌هنگام‌سازی می‌شوند و با توجه به این تغییرات، خوشه‌های موجود نیز باید به‌روزرسانی شوند.

۷- مراجع

Aggarwal, C. C., Ta, N., Wang, J., Feng, J., and Zaki, M. J. 2007. "Xproj: a framework for projected structural clustering of XML documents", In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07), pp.46–55.

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F., 2009, "Hamshahri: A Standard Persian Text Collection", Journal of Knowledge-Based Systems, Vol.22 No.5, pp.382-387.

Algergawy, A., Mesiti, M., Nayak, R., Saake, G., 2011. "XML Data Clustering: An Overview", ACM Computing Surveys Journal, Vol. 43, No. 4.

Antonellis, P., Makris, C., and Tsirakis, N. 2008. "XEdge: Clustering homogeneous and heterogeneous XML documents using edge summaries", In Proceedings of the 2008 ACM Symposium on Applied Computing (SAC). Brazil, 1081–1088.

Atashpaz, E., Lucas, C., 2007, "Imperialist Competitive Algorithm: An Algorithm for Optimization Inspired by Imperialist Competition", IEEE Congress on Evolutionary Computation, pp.4661-4667.

Bijankhan, M., 2006. "Naghshe Peykarehaye Zabani dar Neveshtane Dasture Zaban: Mo'arrefiye yek Narmafzare Rayane'i [theRole of Corpus in generating grammar: Presenting a computational software and Corpus] ", Iranian Linguistic Journal, Vol. 19 pp. 48-67.

Choi, I., Moon, B., and Kim, H.-J. 2007. "A clustering method based on path similarities of XML data". Data & Knowledge Engineering, Vol.60, pp.361–376.

Hwang, K.H., Ryu, K.H., 2010, "A weighted common structure based clustering technique for XML documents", Journal of Systems and Software, Vol.83 No.7, pp.1267-1274.

Kim, T. S., Lee, J. H., and Song, S. W., 2008. "Semantic Structural Similarity for Clustering XML Documents". Convergence and Hybrid Information Technology, 2008. ICHIT '08. International Conference on, Vol.60, pp. 552-557.

Lee, M.L. , Yang, L.H. , Hsu, W., Yang, X., 2007. "XClust: Clustering XML Schemas for Effective Integration", International Symposium on telecommunications.

Manning, C. D., Raghavan, F., and Schuetze, H., 2008 و "Introduction to Information Retrieval", Cambridge University Press.

Nayak, R. 2008. "Fast and effective clustering of XML data using structural information", Knowledge and Information Systems, Vol. 14, No.2, pp.197–215.

Nayak, R. and Iryadi, W., 2006. "XML schema clustering with semantic and hierarchical similarity measures", Knowledge-Based Systems, Vol.20, No.4, pp.336–349.

Niknama, T.,Taherian Fardb, E., Pourjafarianb, N., Roustaa, A., 2011, "An Efficient Hybrid Algorithm Based on Modified Imperialist Competitive Algorithm and K-means for Data Clustering", Journal Engineering Applications of Artificial Intelligence, Vol.24 No.2, pp.306-317.

Rusu, L.I., Rahayu, W., Taniar, D., 2008, "Intelligent Dynamic XML Documents Clustering", 22nd International Conference on Advanced Information Networking and Applications, Vol. 14 No. 3, pp. 449- 456.

Ta, N., Wang, J., Feng, J., Zaki, M., 2007. "Xproj: a framework for projected structural clustering of xml documents", KDD '07 Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 46-55.

Tagarelli, A. , Greco, S., 2010, "Semantic Clustering of XML Documents", ACM Transactions on Information Systems, Vol.28, No.1, pp.1-56.

Tran, T., Nayak, R., Bruza, P., 2008, "Combining Structure and Content Similarities for XML Document Clustering", 7th Australasian Data Mining Conference, Vol.87, pp.35–41.

Xu, R., Wunsch, D., 2005, "Survey of clustering algorithms", Neural Networks IEEE Transactions on, Vol.16, No.3, pp.645-678.

Zhang, L., Li, Z., Chen, Q., Li, N., 2010, "Structure and Content Similarity for Clustering Xml Documents", Web-Age Information Management, Vol. 6185, No.1, pp.116-124.

علی مرادی لالمی مدرک کارشناسی مهندسی کامپیوتر را در سال ۱۳۸۹ از جهاد دانشگاهی بندرانزلی اخذ کرده و از سال ۱۳۹۰ دانشجوی کارشناسی ارشد رشتهٔ مهندسی نرم‌افزار دانشگاه

گیلان است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از: پردازش زبان طبیعی، شبکه‌های اجتماعی، وب معنایی، متن کاوی و داده کاوی.
نشانی رایانامهٔ ایشان عبارت است از:
amoradii@webmail.guilan.ac.ir

اسد ا... شاه‌بهرامی مدرک کارشناسی خود را از دانشگاه علم و صنعت در سال ۱۳۷۲ دریافت کرده است. همچنین کارشناسی ارشد خود را از دانشگاه شیراز در سال ۱۳۷۵ دریافت و درجهٔ دکترای خود را نیز از دانشگاه صنعتی دلفت هلند در سال ۱۳۸۷ اخذ کرده و در حال حاضر، ایشان عضو هیئت علمی و دانشیار در دانشکده فنی و مهندسی دانشگاه گیلان است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از: معماری پیشرفته کامپیوتر، برنامه‌نویسی SIMD، و پردازش ویدئو و تصلویر دیجیتال و معماری قابل پیکربندی.
نشانی رایانامهٔ ایشان عبارت است از:
shahbahrami@guilan.ac.ir

رضا ابراهیمی آتانی مدرک کارشناسی خود را از دانشگاه گیلان در سال ۱۳۸۱ دریافت و همچنین درجهٔ کارشناسی ارشد و دکترای خود را از دانشگاه علم و صنعت در سال ۱۳۸۳ و ۱۳۸۹ اخذ کرده است. در حال حاضر، ایشان عضو هیئت علمی و استادیار در دانشکده فنی و مهندسی دانشگاه گیلان است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از: رمزنگاری، امنیت کامپیوتر، امنیت شبکه، پنهان کردن اطلاعات و طراحی VLSI.
نشانی رایانامهٔ ایشان عبارت است از:
rebrahimi@guilan.ac.ir

مهران علیدوست‌نیا مدرک کارشناسی خود را از دانشگاه گیلان در سال ۱۳۹۰ دریافت و همچنین درجهٔ کارشناسی ارشد خود را نیز از دانشگاه گیلان در سال ۱۳۹۳ اخذ کرده است. زمینه‌های پژوهشی مورد علاقهٔ ایشان عبارتند از: امنیت سیستم‌های نرم‌افزاری، امنیت شبکه، شبکه‌های گمنام، بهینه‌سازی سیستم‌های نرم‌افزاری و زمینه‌های تئوری مرتبط با زبان‌های برنامه‌سازی.
نشانی رایانامهٔ ایشان عبارت است از:
alidoost@msc.guilan.ac.ir