کارگاه‌های آموزشی مرکز اطلاعات علمی

آموزش مهارت های کاربردی در تدوین و چاپ مقاله

اصول تنظیم قراردادها

مقاله نویسی علوم انسانی

International Conference on
INNOVATION in Science & Technology
Barcelona
Barcelona SPAIN 14th. October 2016
istconf
کنفرانس بین المللی
نوآوری در علوم و تکنولوژی
بارسلون – اسپانیا ۲۳ مهرماه ۹۵

ارزیابی روشهای خوشه بندی اسناد XML

داود شفیعی
[۱] دانشجوی کارشناسی ارشد ، گروه کامپیوتر ، واحد میبد ، دانشگاه آزاد اسلامی ، میبد ، ایران
davoud.shafiei@yahoo.com

محمد جواد کارگر
[۲] عضو هیئت علمی گروه کامپیوتر ، واحد میبد ، دانشگاه آزاد اسلامی ، میبد ، ایران
showcaran@gmail.com

چکیده

زبانXML به دلیل ساختار نیمه ساخت‌یافته و خاصیت خود توصیف بودن، به ابزاری مناسب جهت ذخیره و تبادل داده‌ها تبدیل گشته و امکان مدل کردن انواع مختلفی از داده‌ها را فراهم کرده است. با توجه به گسترش روزافزون استفاده از اسناد XML و اهمیت سازماندهی آنها، مطالعه و بررسی بهبود روش‌های خوشه‌بندی اسناد XML بسیار ضروری است. یکی از مهمترین چالش‌های موجود در این زمینه، کاوش حجم عظیمی از اسناد ناهمگن XML با در نظر گرفتن معنای ساختاری در کنار ساختار و محتوای اسناد می‌باشد. یکی از مشکلات اصلی اسناد XML، استخراج ویژگی‌های مهم، مدل کردن و ترکیب ساختار و محتوای اسناد با در نظر گرفتن معنای ساختاری درون اسناد به منظور کاوش سریع و خوشه‌بندی کارآمد روی مجموعه اسناد است. اساس خوشه‌بندی اسناد XML بر مبنای استفاده از معیاری است که تعیین کننده میزان شباهت بین اسناد می‌باشد. که این شباهت می‌تواند از جنبه ساختار، محتوا، مفهوم و یا ترکیبی از هر یک از آنها باشد. سپس با اعمال الگوریتم‌های خوشه‌بندی متداول و یا بهبود یافته آنها، گروه‌بندی اسناد صورت می‌گیرد. روش کلی کار در این مقاله ترکیب مدلهای Xclust و Xproj به منظور دقت بالا در خوشه‌بندی اسناد XML می‌باشد. نتایج بر روی دیتاست Sigmod نشان داده که با معیار Accuracy مدل ترکیبی به میزان ۹۰،۳۴ درصد در مقایسه با مدل‌های Xclust با ۸۰،۳۵ و Xproj با ۸۵،۶۹ ، میزان دقت بیشتری دارد.

واژگان کلیدی: خوشه‌بندی، اسناد متنیXML، مدل Xclust، مدل Xproj

مقدمه

امروزه اسناد متنی اساسی‌ترین و فرم اصلی اطلاعات در محیط اینترنت هستند. بنابراین نظارت، مدیریت اسناد متنی و استفاده از آن به‌عنوان منابع باارزش، به‌سرعت در حال افزایش یافتن است. تجزیه‌وتحلیل جریان متن دارای اهمیت فراوانی است و کاربردهای مختلف ازجمله استخراج گروه‌های خبری، تشخیص و ردیابی موضوع، سازمان‌دهی اسناد و شناسایی دارد. خوشه-بندی یکی از مهم‌ترین روش‌های تجزیه‌وتحلیل جریان متن است. امروزه حجم زیادی از دانش بشر در غالب متن‌های الکترونیکی وجود دارد. پس یافتن اطلاعات موردنظر از میان این کوه اطلاعات بسیار مشکل است. برای پاسخ‌گویی به این چالش روش‌های دسته‌بندی و خوشه‌بندی متون پیشنهاد شده‌اند. هدف آن‌ها سازمان‌بندی مجموعه بزرگ مستندات متنی مانند وب، درون تعداد دسته‌ها یا خوشه‌های معنی‌دار نسبتاً کمی می‌باشد. ازجمله کاربرد این روش‌ها در بازیابی اطلاعات از پایگاه‌های داده، پروسه راه‌حل‌های هوشمند تجارت و مدیریت سیستم‌های اطلاعاتی یا پورتال سازمانی می‌باشد. با رشد سریع علم و افزوده شدن صفحات وب و پیدایش ویکی‌ها و وبلاگ‌ها، استفاده از روش‌های خوشه‌بندی داده می‌تواند رویکرد مناسبی باشد که به‌صورت بدون نظارت و بدون نیاز به مجموعه آموزشی اولیه، خوشه‌بندی را انجام می‌دهند.

به‌واسطه قالب نیمه ساخت‌یافته و خاصیت خود توصیف بودن، XML به ابزاری مناسب جهت ذخیره و تبادل داده‌ها تبدیل گشته و امکان مدل کردن انواع مختلفی از داده‌ها را فراهم کرده است (J.H. Hwang and Keun Ho Ryu,2010). با توجه به گسترش روزافزون استفاده از اسناد XML اهمیت سازمان‌دهی این اسناد، مطالعه و بررسی این موضوع و ایجاد روش‌های بهبود خوشه‌بندی اسناد XML جهت استفاده مؤثرتر از آن‌ها ضروری می‌نماید. از خوشه‌بندی می‌توان به‌عنوان یکی از تکنیک‌های پردازش اسناد نام برد. در خوشه‌بندی اسناد XML تمرکز بر استخراج داده‌های مفید و کشف دانش درداده‌ها می-باشد (T. Bray and J.Paoli,2000). خوشه‌بندی اسناد، به‌عنوان یکی از روش‌های یادگیری ماشین بدون نظارت، در زمینه-های مختلف پردازش زبان‌های طبیعی از قبیل بازیابی اطلاعات و خلاصه‌سازی متون کاربرد گسترده‌ای دارد. به‌عنوان‌مثال در موتورهای جستجو، خوشه‌بندی اسنادی که از نتایج موتور جستجو به دست می‌آید تأثیر قابل‌ملاحظه‌ای در بهبود دقت بازیابی اطلاعات خواهد داشت. استخراج ویژگی‌های مهم، مدل کردن و ترکیب ساختار و محتوای اسناد با در نظر گرفتن معنای ساختاری درون اسناد به‌منظور کاوش سریع و خوشه‌بندی کارآمد روی مجموعه اسناد XML بسیار حیاتی است. اساس خوشه‌بندی اسناد XML بر مبنای استفاده از معیاری است که تعیین‌کننده میزان شباهت بین اسناد می‌باشد که این شباهت می‌تواند از جنبه ساختار، محتوا، مفهوم و یا ترکیبی از هر یک از آن‌ها باشد.

پردازش اسناد، یکی از شاخص‌های بسیار مهم در دنیای اطلاعات است. خوشه‌بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه‌شده است. خوشه‌بندی قابلیت ورود به فضای داده و تشخیص ساختارش را امکان‌پذیر می‌نماید. لذا به‌عنوان یکی از ایده آل‌ترین مکانیزم‌ها، برای کار با دنیای عظیم داده‌ها محسوب می‌شود. خوشه‌بندی، یافتن ساختاری در مجموعه‌ای از داده‌ها است که در خوشه‌های متفاوت قرار می‌گیرند(J. Kim and Hyoung Joo Kim,2004). به‌بیان‌دیگر می‌توان گفت که خوشه‌بندی قرار دادن داده‌ها در گروه‌هایی است که اعضای هر گروه از زاویه خاصی شبیه یکدیگرند. درنتیجه شباهت بین داده‌های درون هر خوشه حداکثر و شباهت بین داده‌های درون خوشه‌های متفاوت حداقل می‌باشد. معیار شباهت در اینجا، فاصله بوده یعنی نمونه‌هایی که به یکدیگر نزدیک‌ترند در یک خوشه قرار می‌گیرند (G. Costa et al,2013). به‌عنوان نمونه در خوشه‌بندی اسناد دوری و یا نزدیکی داده‌ها متناسب با تعداد کلمه‌های مشترکی که در دو سند وجود دارد و یا در خوشه‌بندی سبد خرید مشتریان، فاصله بر اساس شباهت خرید تعیین می‌شود (J.H. Hwang and Keun Ho Ryu,2010). لذا محاسبه فاصله بین دو داده در خوشه‌بندی بسیار مهم می‌باشد؛ زیرا کیفیت نتایج نهایی را دستخوش تغیر قرار خواهد داد.

هدف خوشه‌بندی، که گاهی بانام یادگیری بدون نظارت شناخته می‌شود، تقسیم مجموعه‌ای از داده‌های یک سند XML به خوشه‌های مختلف است. آیتم‌های قرارگرفته در هر خوشه، در مقایسه با آیتم‌های درون سایر خوشه‌ها، میزان «همانندی»

بیشتری را به اشتراک می‌گذارند. روش‌های خوشه‌بندی با در نظرگیری محتوا، معنا و ساختار اسناد و روابط سلسله مراتبی بین گره‌های این اسناد، معیاری به‌عنوان همانندی را به کار می‌گیرند تا میزان همانندی و یا عدم همانندی بین دو سند را تخمین بزنند. در این مقاله بر مبنای ترکیب روش‌های Xclust (M.J.Zaki,2002) و XPROJ (C.C. Aggarwal et al,2007) مدلی جدید برای خوشه‌بندی اسناد متنی XML بر روی مجموعه داده‌های Sigmod (www.cs.washington.edu,2016) ارائه می‌کنیم.

ساختار کلی مقاله را بشرح زیر سازماندهی کردیم: در بخش دوم تحقیقات قبلی در رابطه با خوشه‌بندی اسناد XML را توضیح می‌دهیم، در بخش سوم مدل پیشنهادی را توضیح می‌دهیم، در بخش چهارم به ارزیابی و نتایج خواهیم پرداخت و نهایتا در فصل پنجم، نتیجه‌گیری و کارهای آینده را توضیح می‌دهیم.

کارهای قبلی

تاکنون تحقیقات زیادی راجع به خوشه‌بندی اسناد XML انجام شده است و از الگوریتم‌های متفاوتی برای این کار استفاده شده است. در این بخش کارهای محققان که برای خوشه‌بندی اسناد XML ارائه شده است را توضیح خواهیم داد.

S.GULATI و همکارش الگوریتم‌های Xclust، S-Grace، Xproj و Xcleaner را برای خوشه‌بندی اسناد XML ارزیابی و مقایسه کرده‌اند (S. Gulati and G. Munjal,2015). در مدل Xclust خوشه‌بندی بر مبنای DTD انجام شده است. در ابتدا، گروه‌هایی باید شناسایی شوند که به DTD ها متصل هستند و احتمال همانندی آنها از نظر ساختاری و معنایی وجود دارد. بدین شکل برنامه طراحی شده می‌تواند عوامل مشابه را بکار گیرد تا در صورت بروز نامنظمی، توجه ویژه‌ای را به DTD ها معطوف کند و به مجموعه‌ای از DTD های هماهنگ دست پیدا کند. با توجه به این که سازگارسازی DTD های دارای احتمال همانندی در سطح ساختاری یا معنایی نیازمند ساختارسازی مجدد است، بنابراین عمل وفق دادن DTD های کاملا همانند، امری ساده‌تر خواهد بود. عمل خوشه‌بندی به صورت بازگشتی به DTD های یک خوشه انجام می‌شود. این فرآیند تا آن جا ادامه پیدا می‌کند که فرمی قابل مدیریت از تمامی DTD ها به دست آمده باشد. مدل S-Grace برای خوشه‌بندی اسناد XML بر مبنای گراف پیشنهاد شده است. الگوریتم Xproj مجموعه شاخصی از زیرساختارها را برای خوشه‌های مشابه مورد استفاده قرار می‌دهد. در واقع این الگوریتم خوشه‌بندی، بخشی از یک الگوریتم ساختار محور است که با ایجاد مجموعه‌ها، به بهترین شکل از همانندی‌های ساختاری اسناد مختلف بهره می‌برد. الگوریتم Xcleaner، این الگوریتم با استخراج الگوهای ویژگی شناختی مجموعه داده‌های ورودی و مقایسه ساده آنها، اسناد را خوشه‌بندی می‌کند. اندازه‌گیری تشابه اسناد XML بسیار حیاتی است.

مدلی برای خوشه‌بندی اسناد XML با استفاده از ماتریس همسایگی ارائه شده است (Xue-Liang Zhang et al,2010). در این ماتریس اسناد با هم مقایسه می‌شوند و دقت تشابه آنها در ماتریس نوشته می‌شود و بر مبنای آن اسنادی که با هم تشابه یا تا حدودی با هم یکسان هستند شناسایی و استخراج می‌شوند. ارزیابی و نتایج برروی دیتاست ACM Sigmod انجام شده است. نتایج نشان می‌دهد که دقت ماتریس همجواری در مقایسه با مدل‌های دیگر بیشتر است.

خوشه‌بندی اسناد متنی با استفاده از وزن‌دهی به اسناد ارائه شده است (N.K. Nagwani and A.Bhansali,2010). در این مدل ساختار و محتویات فایل XML مقایسه و ارزیابی شده است. برای وزن دهی به اسناد از معیار فاصله استفاده شده است. ارزیابی و نتایج بر روی دیتاست Wikipedia XML انجام شده است. وزندهی به ساختار، محتویات و عناصر داده شده است. و برای ارزیابی از معیارهای Precision، Recall و F-Measure استفاده شده است. دقت میعارها به ترتیب برابر ۰,۷۶، ۰,۹۱ و ۰,۸۳ است.

خوشه‌بندی اسناد متنی با استفاده از شبکه‌های خود سازمانده پیشنهاد شده است (M. Hagenbuchner et al,2006). هدف از استخراج اطلاعات از داده‌ها که روش‌های متعدد و متنوعی از یادگیری رابطه‌ای گرفته تا آمار و شبکه‌های عصبی مصنوعی را در برمی‌گرد، یافتن دانشی از داده‌های پیچیده با ابعاد بالا و حجم بالاست که بدون رایانه قابل تحلیل نیستند. تجزیه و تحلیل خوشه‌ای یکی از روش‌های استخراج اطلاعات از داده‌ها و همچنین از روش‌های یادگیری بدون نظارت است که برای خوشه-بندی داده‌ها با توجه به شباهت یا درج نزدیکی مورد استفاده قرار می‌گیرد و می‌تواند داده‌ها را به دسته‌های همگن و متمایز تقسیم کند. روش‌های متنوعی و الگوریتم‌های متفاوتی برای خوشه‌بندی وجود دارد که یکی از آنها نقشه‌های خودسازمانده است. شبکه‌های خود سازمانده از انواع شبکه‌های عصبی مصنوعی با یادگیری بدون ناظر می باشند که در تحلیل فضاهای پیچیده توانایی زیادی دارند. به طور کلی اگر داده‌های در فضای ورودی به یکدیگر نزدیک باشند، در نقشه حاصل از شبکه خودسازمانده نیز نزدیک به یکدیگر باقی می‌مانند. به همین دلیل نقشه‌های خودسازمانده می‌تواند فضای ورودی با ابعاد بزرگ را در یک نقشه دو بعدی، با حفظ ساختار توپولوژیکی داده های ورودی داشته باشند. ارزیابی و نتایج بر روی دیتاست IMDB انجام گرفته است. نتایج نشان می‌دهد که دقت مدل در خوشه‌بندی برابر ۹۸,۶۶ درصد است و در برخی موارد به ۱۰۰درصد دقت دست یافته است.

خوشه‌بندی اسناد XML با استفاده از تکنیک منطق فازی پیشنهاد شده است (J. Liu and X.X. Zhang,2014). منطق فازی یک روش تصادفی و یا احتمالی نمی‌باشد و در حقیقت این روش خود یک نظام خاص برای مواجه شدن با موقعیت‌های دارای ابهام و غیر قطعی معرفی می‌کند. ویژگی اساسی نظریه مجموعه فازی نمایش داده‌های غیر قطعی و همچنین عملیات و برنامه‌ریزی ریاضی می‌باشد. هر مجموعه فازی می‌تواند توسط یک تابع عضویت نشان داده شود که بیانگر کلاس عضویت عنصر x در مجموعه مرجع X به مجموعه فازی A است. اگر درجه عضویت یک عنصر از مجموعه برابر صفر باشد، آن عضو کاملاً از مجموعه خارج است و اگر درجه عضویت یک عضو برابر با 1 باشد، آن عضو کاملا در مجموعه قرار دارد. در مدل پیشنهادی به هر المان یک مقدار در بازه ۰ تا ۱ نسبت داده شده است. نتایج نشان داده که مدل منطق فازی، دقت بالایی در تشابه اسناد دارد.

مدل FXProj بر پایه فازی برای خوشه‌بندی اسناد XML ارائه شده است (T. Ji et al,2011). در این مدل از قوانین فازی برای وابستگی اسناد استفاده شده است و برای تعریف اسناد از ساختار درختی استفاده شده است. یکی از مسائل بسیار با اهمیت مطرح در خوشه بندی داده‌ها، محاسبه میزان فاصله میان اشیاء (عدم شباهت) است که می‌تواند دارای هزینه‌های پردازشی و ورودی/خروجی بسیار زیادی باشد. در این مقاله با انجام عملیات پیش پردازشی بر روی درخت، سرعت الگوریتم در حذف نویزها و عملیات خوشه بندی بهبود داده شده است. نتایج نشان داده است که دقت معیار F-Measure در مدل FXProj در مقایسه با مدل‌های دیگر بیشتر است.

خوشه‌بندی اسناد XML با استفاده از تکنیک خوشه‌بندی سلسله مراتبی و قوانین فازی پیشنهاد شده است (Chun-Ling et al,2010). در روش خوشه‌بندی سلسله مراتبی، به خوشه‌های نهایی بر اساس میزان عمومیت آنها ساختاری سلسله مراتبی، معمولا به صورت درختی نسبت داده می‌شود. از روش خوشه‌بندی سلسله مراتبی جهت تعیین گروه‌های همگن و از شاخص ضریب همبستگی جهت انتخاب مناسب‌ترین روش خوشه بندی استفاده شده است. ارزیابی و نتایج بر روی دیتاست‌های Reuters و Hitech انجام شده است. در مرحله اول عملیات مربوط به استخراج متون انجام می‌گیرد و در مرحله دو بر مبنای قوانین فازی و تابع عضویت فازی که در این مدل از تابع عضویت ذوزنقه‌ای استفاده شده است انجام گرفته است. و در مرحله سوم بر مبنای ماتریس همجواری و درخت تشابه اسناد، عملیات خوشه‌بندی اسناد انجام می‌گیرد.

خوشه‌بندی اسناد XML با استفاده از درخت MST و روابط معنایی انجام شده است (L. Song et al,2009). الگوریتم MST در دو فاز انجام می‌شود. در مرحله اول المان‌های سند شناسایی می‌شوند و در مرحله دوم هر المان‌ها به صورت مجزا

کوچکترین زیر درخت فراگیر را برای خود می‌سازد. با تشکیل زیردرخت‌ها بر مبنای معیار تشابه هر سند با اسناد دیگر مقایسه می‌شود. ارزیابی و نتایج بر روی دیتاست‌های Niagara و IBM XML انجام گرفته است. نتایج نشان داده است که مدل MST از دقت معیار بالایی برخوردار است.

مدل S-Grace (W. Lian et al,2004) برای خوشه‌بندی اسناد XML بر مبنای گراف پیشنهاد شده است. الگوریتم S-Grace یک مدل سلسله مراتبی برای خوشه‌بندی اسناد XML است که از الگوریتم خوشه‌بندی Rock[1] استفاده می‌کند. با توجه به صفات تعریف‌شده در الگوریتم S-Grace، احتمال می‌رود که فرمول خوشه‌بندی میزان فاصله‌ی حقیقی در الگوریتم S-Grace نتواند بخوبی بر برخی از فایل‌ها، عمل کند. الگوریتم Rock در چنین شرایطی می‌تواند مؤثر باشد. علی‌رغم این‌که در یک‌فاصله‌ی دور، بسیاری از صفات فایل به‌موقع بسته نمی‌شوند، بااین‌حال ارتباط با همتاهای تکرارشونده در سایر گروه‌ها حفظ می‌شود. بنابراین بهتر است که این صفات از خوشه‌های یکسان انتخاب شوند. عموماً گراف‌های ساختاری به‌دست‌آمده از اسناد در خوشه انتخابی SG، محاسبه و ذخیره می‌شوند. فرآیند پیش‌ـخوشه‌بندی با بکارگیری متد درهم‌سازی منجر به تولید SG می‌شود. یک جفت گراف ساختاری درون SG ازاین‌رو با یکدیگر همتا به شمار می‌روند که میزان فاصله‌ی آن دو از آستانه‌ی تحمل پیشنهادی کمتر است. با در نظر گرفتن این فاصله می‌توان میزان اختلاف بین دو گره تقریباً مشابه، شامل گراف‌های ساختاری و SG را محاسبه و الگوریتم‌های انتخاب بر اساس فاصله‌ی درونی را بررسی کرد. الگوریتم Rock با استفاده از خصیصه‌ی فراپیوند، دو گروه مناسب را برای ادغام درون سلسله‌مراتب الگوریتم انتخاب می‌کند. با در اختیار داشتن زمان گراف‌های ساختاری، متغیر b مربوط به SG نشان‌دهنده‌ی تعداد دفعات تکرار همتاها خواهد بود. همچنین متغیر b نشان‌دهنده‌ی احتمال همسایگی گراف‌های ساختاری در سلسله‌مراتب خواهد بود که می‌تواند منجر به طولانی شدن آستانه‌ی فاصله شود. در الگوریتم S-Grace، تعداد همتاهای یک گراف ساختاری به‌سادگی با تعداد اسناد مرتبط با هر آیتم محاسبه می‌شود. در فرمول‌بندی این الگوریتم، می‌توان یک سند XML دارای طراحی مناسب را به شکل یک گراف طراحی، و یا حتی گراف ساختاری تبدیل کرد؛ به‌علاوه، با طرح‌ریزی فاصله‌ی بین دو سند XML می‌توان تعداد رابطه‌ی المان/زیرالمان را کمی‌سازی کرد.

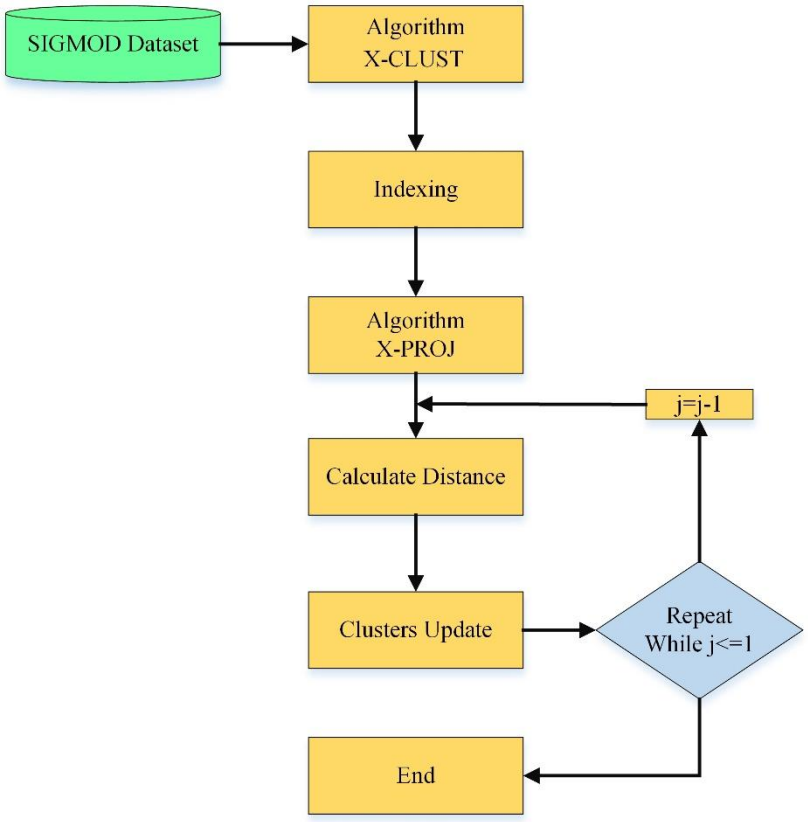
مدل پیشنهادی

استاندارد XML یک مجموعه ثابتی از عناصر نیست بلکه با استفاده از آن کاربران می‌توانند عناصر و یا برچسب‌های مورد نیاز خود را تعریف و در محیط‌های اطلاعاتی دیگر استفاده کنند. این استاندارد برای حل مشکلات مربوط به انتشار الکترونیکی حجم گسترده‌ای از داده‌ها طراحی شده است. برچسب گذاری اسناد متنی XML اولین و کلیدی‌ترین مرحله در پردازش پرس-وجوی XML محسوب می‌شود. الگوریتم‌های برچسب گذاری درختی از طریق تخصیص برچسبی منحصر به فرد به هر گره درخت که یک سند XML نشان دهنده آن است، روند پردازش پرس‌وجوی XML را تسهیل می‌نمایند. رابطه ساختاری بین دو گره درختی از طریق مقایسه دو برچسبی شناسایی می‌شود که همانند هم می‌باشند. بدون الگوریتم‌های برچسب گذاری درختی، پردازش پرس‌وجوی XML می‌تواند سخت باشد.

در مدل ترکیبی در مرحله اول، ابتدا داده‌ها از مجموعه داده Sigmod (www.cs.washington.edu,2016) خوانده و در مرحله دوم با استفاده از الگوریتم Xclust (M.J. Zaki,2002) ساختار درختی را تشکیل و عناصر فایل XML را به اعداد صحیح تغییر نام می‌دهیم و عملیات ایندکس‌گذاری بر روی عناصر انجام می‌گیرد. در مرحله سوم از روش مبتنی بر فراوانی کلمات استفاده می‌کنیم که یک روش بسیار کاربردی برای وزندهی به کلمات کلیدی است. و در مرحله چهارم با استفاده از الگوریتم Xproj (C.C. Aggarwal et al,2007) عملیات مربوط به خوشه‌بندی را انجام می‌دهیم. در الگوریتم Xproj به هر برچسب

[1] Robust Clustering Algorithm for Categorical Attributes

یک مقدار وزنی اختصاص داده می‌شود و عملیات خوشه‌بندی بر مبنای فاصله شباهت انجام می‌شود. وزن نودها در ساختار درختی الگوریتم Xproj محاسبه می‌شود و برمبنای نزدیکی خوشه‌بندی انجام می‌گیرد. در شکل(۱)، فلوچارت مدل ترکیبی نشان داده شده است.



شکل(۱): فلوچارت مدل ترکیبی برای خوشه‌بندی اسناد XML

در الگوریتم Xclust (M.J. Zaki,2002) از یک ساختار سطحی جهت نمایش داده ساختار XML استفاده شده است. در این ساختار تنها از عناصر و یا همان برچسب‌ها که در اسناد XML وجود دارند استفاده می‌شود و محتوای بین آنها و صفات آنها نادیده گرفته می‌شود. در روش Xclust درخت حاصل از سند XML، به صورت قالب ساختاری گره‌ها نمایش داده می‌شود. سپس با استفاده از این قالب‌ها میزان شباهت اسناد بدست می‌آید و با روش افزایشی، عمل خوشه‌بندی انجام می‌گیرد. ساختار سطحی که توسط نمایش درختی بدست می‌آید، نمایانگر گره‌های موجود در هر یک از سطوح درخت می‌باشد. درخت برچسب‌ها به طور حریصانه از بالا به پایین به روش زیر درخت ایجاد می‌شود.

الگوریتم Xproj (C.C. Aggarwal et al,2007) یک الگوریتم تفکیکی که بر مبنای ساختار درختی که برای خوشه‌بندی اسناد XML استفاده می‌شود. این الگوریتم برچسب‌های اسناد XML را به صورت ساختار درختی تشکیل می‌دهد و ارتباط آنها را بر مبنای فاصله تشابه پیدا می‌کند.

برای بدست آوردن TF باید کلمات کلیدی عناصر ارزیابی شوند که کدام عناصر در داخل سند حاوی آن کلمه‌ها هستند و هر کلمه چند بار تکرار شده است. روش TF طبق معادله(۱) تعریف می‌شود که (t_k, d_i) برابر تعداد تکرار هر کلمه t_k در سند d_i است (X. ZHANG et al,2011).

$$w_{ki} = tf(t_k, d_i) = \begin{cases} (t_k, d_i) & t_k \in \text{vector of } d_i \\ 0 & t_k \notin \text{vector of } d_i \end{cases} \quad (۱)$$

برای خوشه‌بندی عناصر باید معیاری برای تعیین فاصله بین عناصر مشخص شود. اگر یک بردار عناصر به‌صورت $< a_1(x), a_2(x), ..., a_n(x) >$ باشد باید از فاصله اقلیدسی طبق معادله (۲) برای به دست آوردن فاصله بین دو عنصر x_i و x_j استفاده می‌کنیم.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (۲)$$

نتایج مدل ترکیبی، می‌بایست در مرحله ارزیابی مورد تحلیل قرار گیرد تا بتوان دقت آن را تعیین نمود و در پی آن کارایی مدل ترکیبی را مشخص کرد. این معیارها را می‌توان هم برای مجموعه داده‌های آموزشی در مرحله یادگیری و هم برای مجموعه رکوردهای آزمایشی در مرحله ارزیابی محاسبه نمود. در این مقاله از معیارهای Precision، Recall، F-Measure و Accuracy برای ارزیابی استفاده کردیم (R.S. Michalski et al,1998).

$$Precision = \frac{TP}{TP + FP} \quad (۳)$$

$$Recall = \frac{TP}{TP + FN} \quad (۴)$$

$$F - Measure = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad (۵)$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (۶)$$

پارامتر TN بیانگر تعداد اسنادی است که دسته واقعی آن‌ها منفی بوده و الگوریتم خوشه‌بندی نیز دسته آن‌ها را به‌درستی منفی تشخیص داده است.TP بیانگر تعداد اسنادی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم خوشه‌بندی نیز دسته آن‌ها را به‌درستی مثبت تشخیص داده است.FP بیانگر تعداد اسنادی است که دسته واقعی آن‌ها منفی بوده و الگوریتم خوشه‌بندی دسته آن‌ها را به‌اشتباه مثبت تشخیص داده است.FN بیانگر تعداد اسنادی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم خوشه‌بندی دسته آن‌ها را به‌اشتباه منفی تشخیص داده است. معیار Recall نسبت تعداد نمونه‌هایی که به‌درستی خوشه آن‌ها تشخیص داده شده را به‌کل نمونه‌ها نشان می‌دهد. تفاوت معیار Precision با معیار Recall در مخرج کسر است. در مخرج کسر Recall تعداد نمونه‌های مثبت واقعی قرار می‌گیرد. اما در Precision تعداد پیشگویی‌های مثبت قرار می‌گیرد. بنابراین Recall نشان می‌دهد که چه نسبتی از مثبت‌های واقعی به‌درستی مثبت تشخیص داده‌شده‌اند. اما Precision نشان می‌دهد که چه نسبتی از مثبت‌ها واقعاً مثبت هستند. معیار F-Measure از ترکیب معیارهای Precision و Recall به دست

می‌آید و در مواردی استفاده می‌شود که نتوان اهمیت ویژه‌ای را برای هر یک از دو معیار Precision و Recall نسبت به یکدیگر قائل شد. مهم‌ترین معیار برای تعیین کارایی یک الگوریتم خوشه‌بندی معیار Accuracy می‌باشد. این معیار دقت کل یک خوشه‌بندی را محاسبه می‌کند. این معیار نشان‌دهنده این است که چند درصد از کل مجموعه داده‌ها به‌درستی خوشه‌بندی شده است.

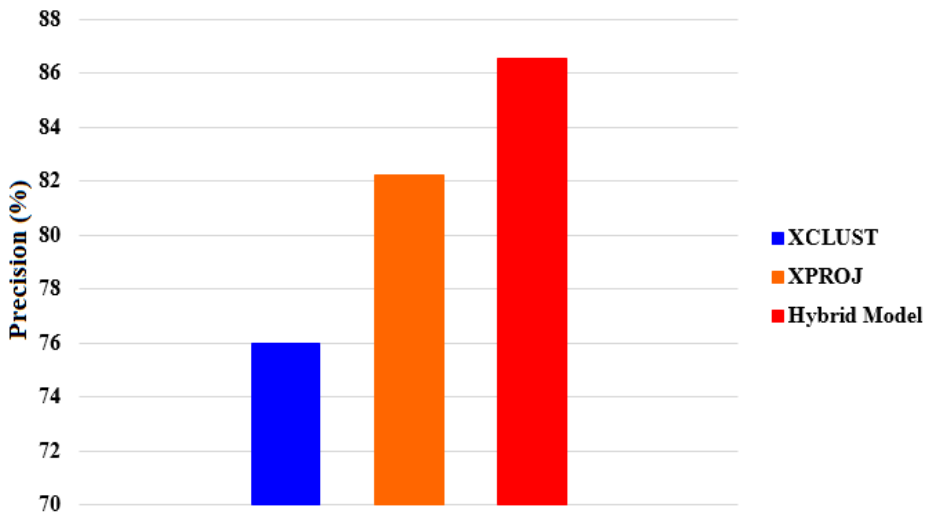
یافته ها

پیاده‌سازی مدل پیشنهادی در محیط برنامه‌نویسی C#.NET 2013 انجام شده است و نتایج حاصل از مدل پیشنهادی با مدل‌هایی که برای خوشه‌بندی اسناد متنی پیشنهاد شده‌اند مقایسه شده است. سی‌شارپ یک زبان شیءگرا و سطح بالا از خانواده زبان‌های چارچوب دات نت، شرکت مایکروسافت است. زبان سی‌شارپ، یک زبان برنامه نویسی چند الگویی است و منظم شده مدل‌های تابعی و شیء گرا می‌باشد. ارزیابی و نتایج بر روی دیتاست Sigmod (www.cs.washington.edu,2016) که حاوی ۱۱۵۲۶ المان و ۳۷۳۷ صفت است انجام شده است. در جدول(۱)، مقایسه دقت مدل ترکیبی و مدل‌های Xclust و Xproj بر روی مجموعه داده Sigmod نشان داده شده است.

جدول(۱): مقایسه مدل ترکیبی با مدل‌های X-Clust و X-Proj

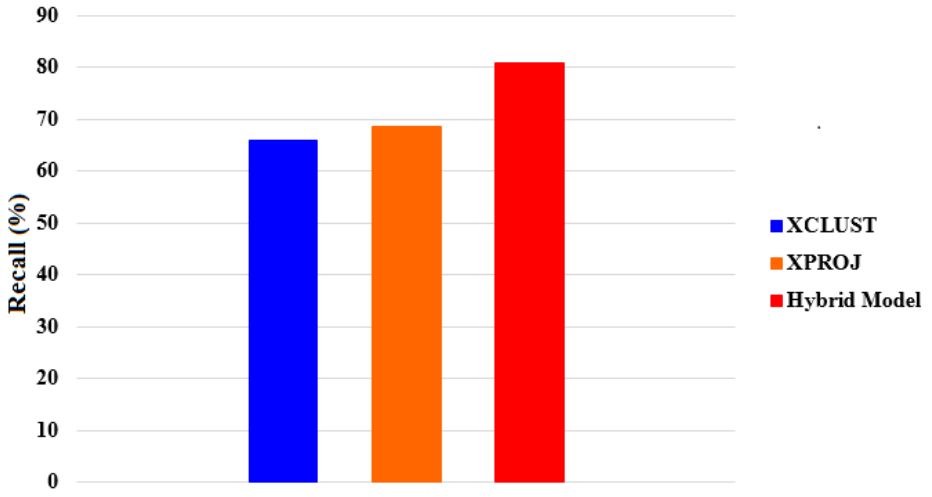
مدل‌ها	معیارها			
	Precision	Recall	F-Measure	Accuracy
Xclust	76.00	65.98	70.63	80.35
Xproj	82.24	68.68	74.85	85.69
Hybrid Model	86.56	80.91	79.18	90.34

در شکل(۲)، نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار Precision نشان داده شده است.



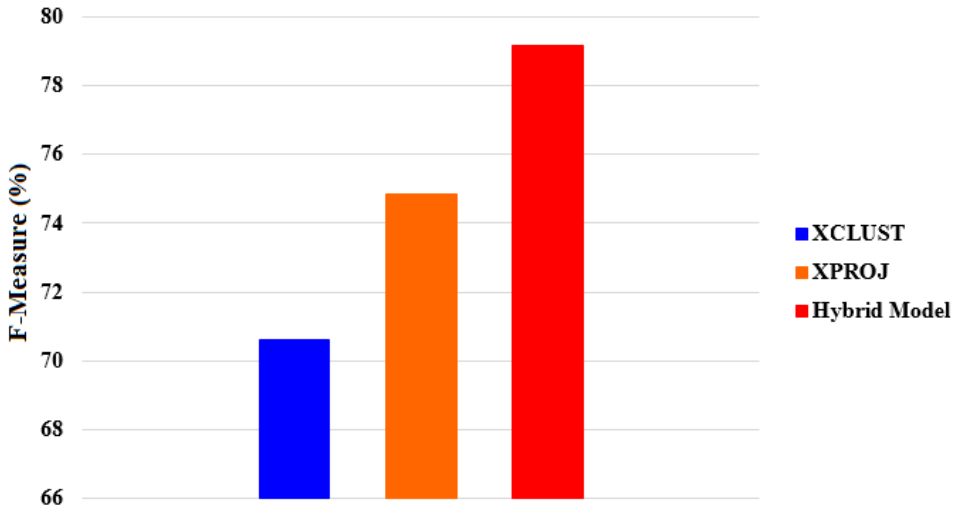
شکل(۲):نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار Precision

در شکل(۳)، نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار Recall نشان داده شده است.



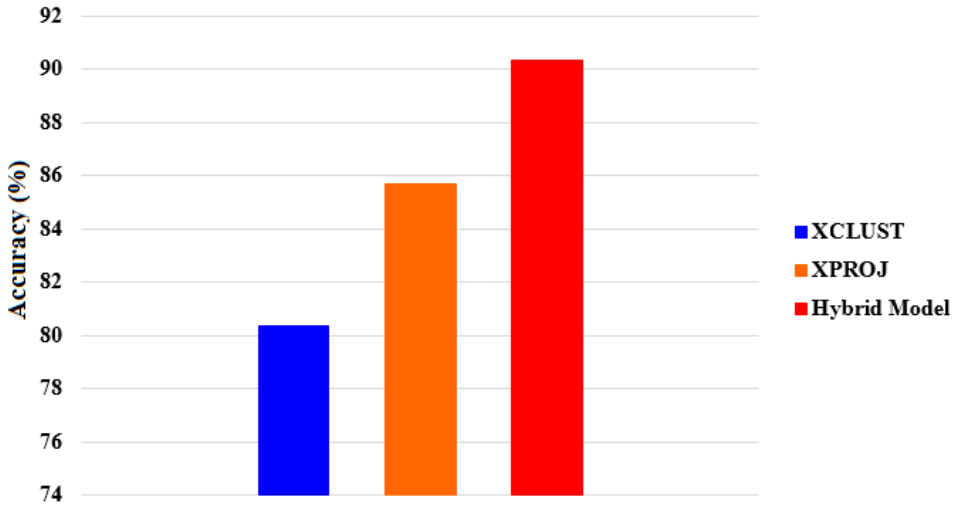
شکل(۳):نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار Recall

در شکل(۴)، نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار F-Measure نشان داده شده است.



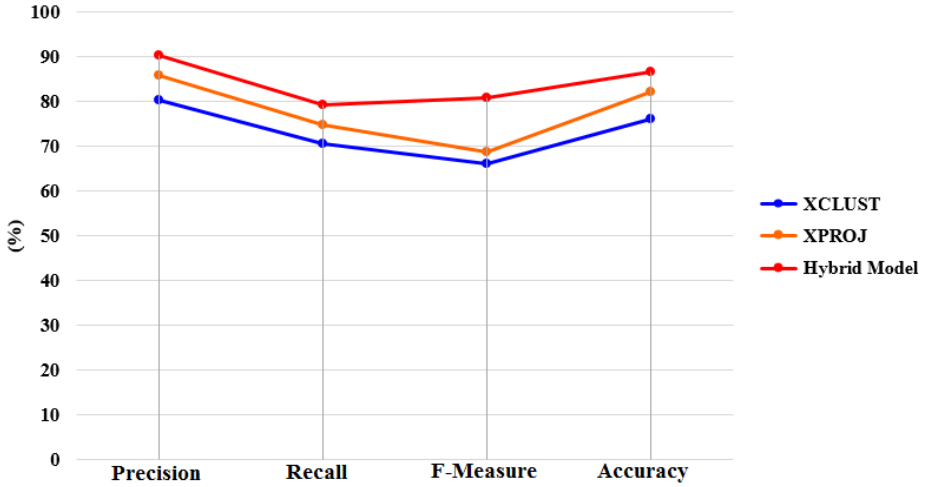
شکل(۴):نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار F-Measure

در شکل(۵)، نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار Accuracy نشان داده شده است.



شکل(۵):نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیار Accuracy

در شکل(۶)، نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای معیارهای Precision، Recall، F-Measure و Accuracy نشان داده شده است.



شکل(۶):نمودار مقایسه‌ی مدل ترکیبی با مدل‌های Xclust و Xproj بر مبنای تمامی معیارها

همانطور که در شکل(۶)، مشاهده می‌کنید دقت مدل ترکیبی در مقایسه با مدل‌های Xclust و Xproj بیشتر است. و دقت مدل Xproj در مقایسه با مدل Xclust بیشتر است.

بحث و نتیجه‌گیری

در این مقاله یک مدل ترکیبی بر مبنای مدل‌های Xclust و Xproj برای خوشه‌بندی اسناد XML پیشنهاد شد. با در نظر گرفتن روش فراوانی کلمات در بین اسناد XML، نتایج نشان داد که نرخ دقت خوشه‌بندی در مدل ترکیبی در مقایسه با مدل‌های Xclust و Xproj بیشتر است. مدل ترکیبی شامل یک قسمت فراوانی کلمات و تشابه اسناد بر مبنای فاصله است که کلمات کلیدی واقع در متن را استخراج کرده، سپس با استفاده از وزندهی و برمبنای مدل Xproj، میزان شباهت بین اسناد XML تخمین می‌گردد. برای بررسی میزان شباهت بین نودهای درخت Xproj از فاصله اقلیدسی استفاده شده است. اسناد XML به دلیل شامل بودن تعداد زیادی عناصر مشترک دارای هم‌پوشانی بسیاری هستند.

وجود کلمات مشترک بین اسناد متنی XML، ارزیابی دقیق میزان شباهت متن‌ها را مشکل می‌کند. در نتیجه در این مقاله با ارائه مدل ترکیبی Xclust و Xproj مدلی جدید برای خوشه‌بندی اسناد متنی XML و با لحاظ کردن تعداد فراوانی کلمات ارائه دادیم. نتایج بر روی دیتاست Sigmod نشان داد که دقت معیار Accuracy در مدل ترکیبی برابر ۹۰،۳۴ و در مدل‌های Xclust و Xproj بترتیب برابر ۸۰،۳۵ و ۸۵،۶۹ است. در آینده می‌توان با ترکیب الگوریتم یادگیری ماشین همانند K-Means با مدل‌های الگوریتمی همانند Xclust، دقت تشخیص در خوشه‌بندی اسناد متنی XML را با تشخیص فاصله شباهت بهبود داد.

منابع:

J.H. Hwang, Keun Ho Ryu, A weighted common structure based clustering technique for XML documents, Journal of Systems and Software, Vol. 83, Issue 7, pp. 1267-1274, 2010.

T. Bray, J. Paoli, Extensible markup language (XML) 1.0, 2nd edn. Sperberg-McQueen CM University of Illinois at Chicago and text encoding initiative. Sun Microsystems Inc, Eve Maler, 2000

J. Kim, Hyoung-Joo Kim, A partition index for XML and semi-structured data, Data & Knowledge Engineering, Vol. 51, Issue 3, pp. 349-368, 2004.

G. Costa, G. Manco, R. Ortale, E. Ritacco, Hierarchical clustering of XML documents focused on structural components, Data & Knowledge Engineering, Vol. 84, pp. 26-46, 2013.

M.J. Zaki, Efficiently mining frequent trees in a forest, in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, ACM: Edmonton, Alberta, Canada, p. 71-80, 2002.

C.C. Aggarwal, & et al., 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, ACM: San Jose, California, USA. p. 46-55, 2007

http://www.cs.washington.edu/research/xmldatasets/ [Last Access: Sep-11-2016]

S. Gulati, G. Munjal, Algorithms for Clustering XML Documents: A Review, 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, IEEE, pp. 654-658, 2015

Xue-Liang Zhang, T. Yang, Bao-Quan Fan, X. Wang, Jin-Mao Wei, A Novel Method for Measuring Structure and Semantic Similarity of XML Documents Based on Extended Adjacency Matrix, International Conference on Applied Physics and Industrial Engineering, Physics Procedia 24, pp.1452-1461, 2012.

N.K. Nagwani, A. Bhansali, Clustering Homogeneous XML Documents Using Weighted Similarities on XML Attributes, IEEE, pp. 369-372, 2010

M. Hagenbuchner, A. Sperduti, A.C. Tsoi, F. Trentini, F. Scarselli, and M. Gori, Clustering XML Documents Using Self-organizing Maps for Structures, INEX 2005, LNCS 3977, pp. 481-496, 2006.

J. Liu, X.X. Zhang, Data integration in fuzzy XML documents, Information Sciences, pp. 1-16, 2014

T. Ji, X. Bao, and D. Yang, FXProj-A Fuzzy XML Documents Projected Clustering Based on Structure and Content, ADMA 2011, Part I, LNAI 7120, pp. 406-419, 2011

Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang, Mining fuzzy frequent item sets for hierarchical document clustering, Information Processing and Management, Vol. 46, pp. 193-211, 2010

L. Song, J. Ma, J. Lei, D. Zhang, and Z. Wang, Semantic Structural Similarity Measure for Clustering XML Documents, WISM 2009, LNCS 5854, pp. 232-241, 2009

W. Lian, D. Wai-lok Cheung, N. Mamoulis, and Siu-Ming Yiu, An Efficient and Scalable Algorithm for Clustering XML Documents by Structure, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 1, pp. 82-96, 2004

X. ZHANG, T. WANG, X. LIANG, F. AO, Y. LI, A Class-based Feature Weighting Method for Text Classification, Journal of Computational Information Systems, 8(3): 965–972, 2011.

R.S. Michalski, I. Bratko, M. Kubat, Machine Learning and Data Mining: Methods and Applications, New York: Wiley, 1998

کارگاه‌های آموزشی مرکز اطلاعات علمی

آموزش مهارت های کاربردی در تدوین و چاپ مقاله

اصول تنظیم قراردادها

مقاله نویسی علوم انسانی