خوشه بندی XML معنایی

چکیده

افزایش دسترس پذیری منابع اطلاعاتی XML ناهمگون تعداد مسائل مرتبط با اینکه چگونه داده‌های نیمه ساختار یافته نشان داده شوند و مدیریت شوند را افزایش داد. اگر چه منابع XML می‌توانند محتوا و ساختار مناسبی را نشان دهند، اسناد XML متفاوت ممکن است در اصل معنا شناسی مربوطه را با تعریف ذهنی از تگ‌های نشانه گذاری رمزگذاری کند. کشف دانش برای استنتاج سازمان معنایی اسناد XML به چالش اصلی در مدیریت داده XML تبدیل شد. در این زمینه، مسئله خوشه بندی داده XML را بر طبق ساختار و به همین ترتیب ویژگی‌های محتوای غنی با دانش هستی شناسی واژگان بررسی می‌کنیم. فریم ورکی را برای خوشه بندی ساختارهای XML منسجم به صورت معنایی بر اساس مدل نمایش تراکنشی پیشنهاد می‌کنیم. آزمایش‌ها بر مجموعه داده واقعی بزرگ شواهدی را ارائه می‌دهند، مبنی بر اینکه رویکرد پیشنهادی در تشخیص گروه داده XML به شدت موثر است و ساختار و یا پیوندهای درونی محتوا را نشان می‌دهد.

1. مقدمه

XML به عنوان نیروی محرکی برای نمایش و تبادل داده در وب معرفی شد. براستی، سیمای خود توصیف و نیمه ساختاریافته XML مدل کردن طیف گسترده‌ای از داده‌ها را به عنوان اسناد XML، به منظور تحقق وعده‌های وب نسل بعد، امکان پذیر ساخت.

منابع داده XML ساختارها و محتوای متفاوتی را نشان می‌دهند. تگ نشانه گذاری، که نقش پایه را برای تحمیل ساختار به سند بازی می‌کنند، عوامل ذهنی را که نام نویسنده را در اطلاعات برنامه نویسی هک می‌کند را منعکس می‌کند. در نتیجه، داده XML به صورت متفاوت اعلام شده ممکن است " از لحاظ معنایی" به درجه خاصی مربوط باشد.

در همچین زمینه‌ای، چالش، استنتاج معنایی از اسناد XML برطبق اطلاعات معنایی در دسترس، یعنی ساختار و ویژگی‌های محتوا است. این مسئله چندین دامنه کاربردی جالب دارد، مانند ادغام منابع داده و پردازش پرس و جو، که می‌تواند به صورت یکپارچه در هر نوع داده نیمه ساختار یافته تعمیم داده شود. برای مثال، تشخیص قرابت‌هایی ساختاری و معنایی در میان داده XML می‌توانند به تکنیک‌هایی برای شاخص گذاری داده کمک کند، و بنابراین فضای پژوهشی را کمتر کند و طراحی طرح‌های پرس و جو را بهبود بخشد.

مانند یک وظیفه داده کاوی اکتشافی اساسی، خوشه بندی راه حل‌های طبیعی را برای کشف ویژگی‌های مشترک و جنبه‌های خاص نشان داده شده توسط اسناد XML نشان می‌دهد. به هر حال، پیچیدگی ذاتی داده نیمه ساختاریافته نیازمند تلاش ناچیزی برای تعیین فریم ورک خوشه بندی موثر است. استخراج ویزگی‌های مهم، مدل سازی محتوا و ساختار اسناد، تعریف مفاهیم مناسب همگن بین اسناد تنها برخی از مسائل بررسی شده هستند.

سهم. در این مقاله از طریق تحلیل عمیق محتوا و ویژگی‌های ساختاری در داده چگونگی ربط معنایی داده‌های XML بررسی می‌شود. آنچه اخیراً برای کاووش داده XML پیشنهاد شده است معرفی نمادی از تاپل‌های درختی در تعریف مدل نمایش XML است که نگاشت درخت اسناد XML را در داده تراکنشی اجازه می‌دهد. مفهوم تاپل درخت به خوبی از شناخت معنایی زیرساختارهای منسجم اسناد XML گرفته شده است، علاوه بر این، یک نمایش XML منطقی، مسطح، را ممکن می‌سازد که برای برآورده سازی نیازمندی‌هایی برای خوشه بندی داده XML برطبق اطلاعات محتوا و ساختار بسیار مناسب است. سهم ما می‌تواند به شکل زیر خلاصه شود:

1. با تدبیر ویژگی‌های مناسب برای داده XML، بر اطلاعات محتوای استخراج شده از عناصر متنی و اطلاعات ساختاری خاص از مسیرهای تگ تمرکز می‌کنیم. هر دو نوع اطلاعات نحوی با دانش ارائه شده توسط هستی شناسی

واژه بدست می‌آیند. به ویژه، برای مورد ساختاری، روش ابهام زدایی حس کلمه[1] جدیدی را برای انتخاب مناسب ترین حس برای هر نام تگ در زمینه مسیر درخت XML انتخاب می‌کند. ویژگی‌های XML به آیتم‌های تاپل درخت XML ابلاغ می‌شود.

2. یک مدل تراکنشی را برای نمایش تاپل‌های درختی XML استخراج شده از یک مجموعه اسناد XML تصور می‌کنیم. این مدل در مبنای یک فریم ورک خوشه بندی XML معنایی است. یک رویکرد تمرینی موثر را بر اساس الگوریتم طراحی شده برای دامنه تراکنش XML می پذیریم اگر چه فریم ورک پیشنهادی برای هر روش خوشه بندی پارامتریک تصور می‌شود.

3. چندین آزمایش را بر مجموعه داده‌های بزرگ انجام دادیم که ارزیابی توانایی فریم ورک را در اجرای خوشه بندی محتوا محور یا ساختار محور، و به همین ترتیب کشف خوشه‌هایی با انواع "ترکیبی" را هدف قرار می‌دهد. شواهد تجربی استدلال کردند که فریم ورک پیشنهادی به شدت موثر است و مقیاس پذیری خوبی را نشان می‌دهد.

کارهای مرتبط. اخیراً رویکردهای متعددی برای کاووش داده XML توصیه شده است، که اغلب با توجه به ساختار بر خوشه بندی اسناد XML تمرکز می‌کند. این رویکردها اساسا با چندین برنامه کاربردی در مدیریت داده ساختار یافته، به خصوص در محیط وب، که تقاضا برای راه حل‌های موثر و کارآمد شناسایی شباهت‌های ساختاری در میان داده نیمه ساخت یافته افزایش می‌یابد، انگیزش می‌یابند. در این زمینه، یک ویرایش از راه دور آگاه از XML برای سنجش شباهت‌های ساختاری در میان اسناد XML پیشنهاد شده است؛ و الگوریتم خوشه بندی سلسله مراتب استانداردی برای ارزیابی اینکه چقدر اسنادهای خوشه به DTD مربوطه خود نزدیک هستند، اعمال شد. به هر حال، به طور کل، محاسبه فاصله ویرایش درخت به نوبه خود غیر عملی است، که نیازمند تعداد درجه دومی از مقایسه‌های بین عناصر اسناد است.

[1] word sense disambiguation

یک بینش مهم در وظایف دسته بندی نظارت شده داده XML از نقطه نظر ساختاری در [19] ارائه شده است. نویسنده یک تکنیک یادگیری ماشین را پیشنهاد داده است که از استخراج ساختارهای فرعی که مکرر در اسناد XML به منظور تعریف نقش‌های دسته بندی رخ می‌دهد استفاده می‌کند.

اخیراً، اهمیت خلاصه سازی داده XML افزایش می‌یابد، اساسا با هدف تعریف ترکیب‌های برای خوشه‌ها با ساختار مشابه است. رویکردها از مدل‌های نمایش مبتنی بر گراف و تطبیق درخت پیشنهادی استفاده می‌کنند. XSketch یک مدل خلاصه اصلاح شده است که مسئله تخمین انتخاب مسیر موجود را در محیط کلی داده XML با گراف ساختاریافته با مقادیر عناصر بررسی می‌کند. ساخت یک خلاصه دقیق بر اساس فرضیات آماری است که فقدان مسیر دقیق و ارزش اطلاعات را در خلاصه جبران می‌کند. به هر حال، ایجاد یک XSketch یک مسئله NP سخت تشریح شده است، بنابراین یک استراتژی پالایش اکتشافی باید استفاده شود.

نیاز برای سازمان دهی داده XML بر طبق ویژگی‌های ساختاری و محتوایی؛ با توجه به افزایش ناهمگونی منابع XML چالش برانگیز شده است. به هر حال، کاووش داده XML از نقطه نظر ترکیب محتوا/ساختار هنوز در مرحله اولیه است، و هیچ رویکردی برای ارائه قابلیت‌های موثر برای خوشه بندی XML معنایی وجود ندارد. اولین تلاش در [6] داده شده است، که اعمال تکنیک خوشه بندی تفکیکی را به اسناد XML نشان داده شده در یک مدل فضای برداری با ویژگی‌های مبتنی بر تگ و متن اعمال می‌کند.

یک نمایش جایگزین، به نام BitCube، در [18] به عنوان شاخص بیت مپ 3 بعدی سه تایی < سند، مسیر عناصر-XML، کلمه> ارائه داده است. شاخص‌های BitCube می‌توانند برای تقسیم اسناد به خوشه ها، با استفاده از فاصله بیتی و سنجش شهرت دستکاری شوند. به منظور افزایش سرعت پاسخ به پرس و جو، عملیات برش/تاس/پیش بینی برای زیر بخش‌های نتیجه شده از فاز خوشه بندی اجرا می‌شود. به هر حال، هیچ تصمیم مهمی توسط مولفین در مورد بهبود ممکن خوشه بندی سند ارائه نشده است. به طور کل، رویکرد از معایب معمولی مدل‌های پیش نمایش بولین، مانند فقدان معیار تطبیق جزیی و سنجش طبیعی رتبه بندی اسناد رنج می‌برد.

تولید ویژگی‌ها برای داده XML در [16] عمیق تر بررسی شده است، جایی که تفسیرها، ساختارها و اطلاعات هستی شناسی با هم ترکیب می‌شوند. به هر حال، در اینجا تمرکز بر ایجاد ویژگی‌های مناسب در راستای اهداف دسته بندی نظارت شده داده XML است.

طرح مقاله. باقی مقاله به شرح زیر سازمان یافته است. بخش 2 اصطلاح شناسی و نمادهای سودمندی، و پیش زمینه ضروری بر مفاهیم تاپل درختی، آیتم، و تراکنش برای دامنه داده XML ارائه می‌دهد. بخش 3 تشریح می‌کند که چگونه ویژگی‌های XML به صورت معنایی به اطلاعات نحوی در آیتم‌های تاپل درختی XML با دانش هستی شناسی واژگان دست می‌یابد. تاکید ویژه این بخش بر روش‌های جدید ابهام زدایی از حس تگ است. بخش 4 فریم ورکی را برای خوشه بندی تراکنش‌های XML ارائه می‌دهد. بخش 5 ارزیابی آزمایش‌های را گزارش می‌دهد که اثربخشی فریم ورک خوشه بندی را گزارش می‌دهد. بخش 6 نتایج برجسته و نکات امیدوار کننده برای پژوهش‌های آینده بیان می‌کند.

2. پیش زمینه

2.1 مسیرها و درخت‌های XML

یک درخت T یک چندتایی $T = \langle r_T, N_T, E_T, \lambda_T \rangle$ است، $N_T \subseteq \mathbb{N}$ به مجموعه‌ای از گره‌ها اشاره دارد، $r_T \in N_T$ ریشه متمایز T است، $E_T \subseteq N_T \times N_T$ به مجموعه یال‌های (بدون دور) اشاره دارد، و $\lambda_T : N_T \mapsto \Sigma$ تابعی از ارتباط یک گره با یک تگ در الفبای Σ. است. فرض کنید Tag، Att، و Str الفبایی از نام تگ ها، صفات، و رشته‌ها هستند. در یک درخت XML، XT یک جفت $XT = \langle T, \delta \rangle$، به طوریکه : 1) T یک درخت تعریف شده در الفبای $\Sigma = Tag \cup Att \cup \{\mathtt{S}\}$ است، که نماد $\mathtt{S} \notin Tag \cup Att$ برای اشاره به مدل محتوای **#PCDATA** استفاده می‌شود؛ 2) با توجه به $n \in Leaves(T) \Leftrightarrow n \in N_T,\ \lambda_T(n) \in Att \cup \{\mathtt{S}\}$؛ 3) $\delta : Leaves(T) \mapsto Str$ تابعی از رشته‌های مرتبط با گره برگ T است.

یک مسیر p XML یک توالی $p = s_1.s_2 \dots .s_m$ از نمادها در $Tag \cup Att \cup \{S\}$ است. نماد s_1 با نام تگ عنصر ریشه سند متناظر است. یک مسیر XML می‌تواند دو نوع باشد: مسیر تگ ؛ اگر $s_m \in Tag,$ ، یا مسیر کامل باشد، اگر $s_m \in Att \cup \{S\}.$ برقرار باشد. به $\mathcal{P}_{XT}$ به عنوان مجموعه‌ای از مسیرهای کامل در XT اشاره می‌کنیم.

فرض کنید $XT = \langle T, \delta \rangle$ یک درخت XML است، و $p = s_1.s_2 \dots .s_m$ یک مسیر XML است. کاربرد p برای XT شناسایی مجموعه گره‌های $p(XT) = \{n_1, \dots, n_h\}$ است به طوری که، برای هر $i \in [1..h]$ ، دنباله‌ای از گره ها، یا مسیر گره، $np_i^p = [n_{i_1}, \dots, n_{i_m}]$ با ویژگی‌های زیر وجود دارد: 1) $n_{i_1} = r_T$ و $n_{i_m} = n_i$ ؛ 2) $n_{i_{j+1}}$ فرزند n_{i_j}، برای هر $j \in [1..m\text{-}1]$ ؛ 3) $\lambda(n_{i_j}) = s_j$ برای هر $j \in [1..m]$. است.

علاوه بر این، می‌گوییم که یک مسیر به یک درخت XML یک پاسخ حاصل می‌کند، که بسته به نوع مسیر تعریف می‌شود. در مورد مسیر تگ p، پاسخ p بر XT دقیقاً مجموعه‌ای از شناسه گره $p(XT)$ است، بنابراین داریم $\mathcal{A}_{XT}(p) \equiv p(XT)$. برای یک مسیر کامل p، پاسخ p بر XT به عنوان مجموعه‌ای از مقادیر رشته‌ای مرتبط با گره‌های برگ شناخته شده با p تعریف می‌شود، به همین ترتیب داریم $\mathcal{A}_{XT}(p) = \{\delta_T(n) \mid n \in p(XT)\}$. برای یک مسیر کامل p، پاسخ p بر XT به عنوان مجموعه‌ای از مقادیر رشته‌ای مرتبط به گره‌های برگ شناخته شده توسط pتعریف می‌شود، $\mathcal{A}_{XT}(p) = \{\delta_T(n) \mid n \in p(XT)\}$.

2.2 تاپل‌های درخت XML.

تاپل‌ها مفهوم تاپل را در یک پایگاه داده رابطه‌ای همانند سازی می‌کنند و بسطی از وابستگی‌های تابعی برای محیط XML پیشنهاد می‌کنند. در یک پایگاه داده رابطه‌ای، یک تاپل یک تابع از هر صفت با مقداری از دامنه متناظر است. برطبق [8]، تعریف‌های زیر را ارائه می‌دهیم:

تعریف 2.1 با توجه به XT از درخت XML، یک تاپل درخت τ یک زیر درخت ماکسیمال XT است، به طوری که بر هر (تگ یا کامل) مسیر p در XT، پاسخ $\mathcal{A}_\tau(p)$ حداقل یک عنصر است.

به $\mathcal{T}_{XT}$ به عنوان مجموعه‌ای از تاپل‌های درخت از XT اشاره می‌کنیم.

مستقیما، یک تاپل درخت یک نمایش درخت (یا زیر درختی) از مجموعه کاملی از مفاهیم متمایز است که بر طبق معنا شناسی ساختاری درخت اصلی با هم همبسته هستند. علاوه بر این، تاپل‌های درخت استخراج شده از همان درخت، ساختار یکسانی را حفظ می‌کنند، در حالی که روش‌های متفاوتی از محتوای مربوطه ساختاری را منعکس می‌کند که می‌تواند طبیعتآ از درخت اصلی استنباط شده باشد.

مثال 1. درخت XML نشان داده شده در شکل 1 را در نظر بگیرید، که دو مقاله علمی را از آرشیو DBLP نشان می‌دهد. هر گره داخلی دارای تگ منحصربفردی است که به نام تگ اشاره می‌کند،در حالی که هر گره برگ نیز با نام و مقدار یک صفت برچسب خورده است، یا نماد S و رشته مربوطه با مدل محتوای #PCDATA متناظر است. مسیر پاسخ‌ها می‌تواند به سادگی محاسبه شود: برای مثال، مسیر dblp.article.title مجموعه‌ای از شناسه‌های گره $\{n_8, n_{22}\}$ را حاصل می‌کند، در حال که مسیر dblp.article.author.S مجموعه رشته {'Hartmut Liefke', 'Dan Suciu'} را کسب می‌کند.

سه تاپل درخت می‌تواند از مثال درخت (شکل 2) استنتاج شود. یک تاپل درختی با شروع از زیررخت راست با ریشه در عنصر dblp استخراج شود. دو تاپل درخت در عوض با شروع از زیردرخت چپ با ریشه در dblp استخراج می‌شود، از آنجایی که در این زیر دخت دو مسیر dblp.article.author وجود دارد، هریک مسیر پاسخ متماییزی را متناظر با نویسنده مقاله کسب می‌کنند.

2.3 یک مدل تراکنشی برای تاپل‌های درخت XML.

در مقدار زیادی از داده‌های ساختاریافته در دسترس، بخش مربوطه با داده تراکنشی نشان داده می‌شود، برای مثال دنباله‌ای با طول متغیر از اشیا با صفات دسته ای. نکته جالب در تحلیل این دامنه خاص از بسیاری از مسائل کاربردی چالش برانگیز ظهور می‌کند (برای مثال تحلیل لاگ‌های دسترسی وب) که می‌تواند با استفاده از داده تراکنشی به

رسمیت شناخته شود. با توجه به مجموعه $\mathcal{I} = \{e_1, \ldots, e_m\}$ از مقادیر دسته‌ای متمایز، یا آیتم‌ها، یک پایگاه داده تراکنشی یک مجموعه چند گانه از تراکنش‌های $tr \subset \mathcal{I}$ است.

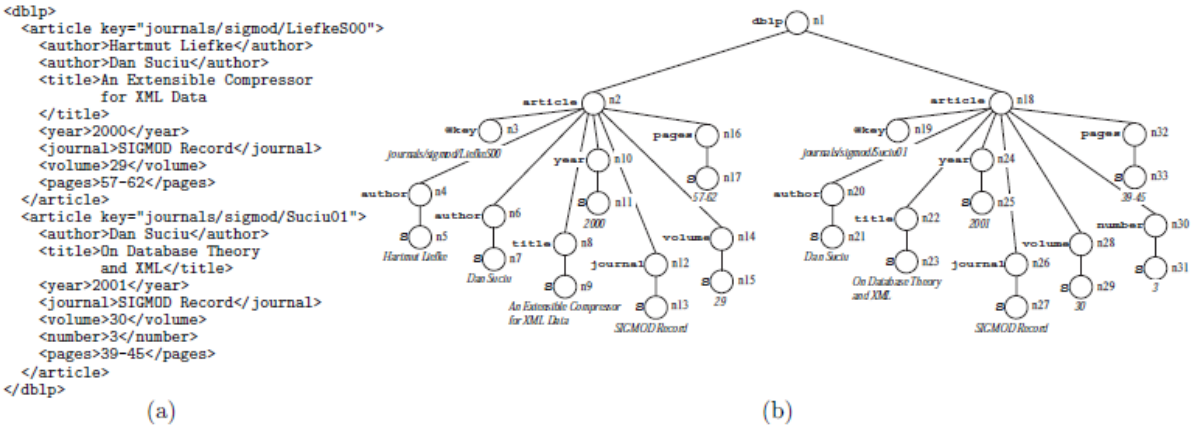
در تنظیمات ما، دامنه آیتم بر روی همه عناصر در مجموعه‌ای از تاپل‌های درخت XML ساخته می‌شود، که مجموعه‌ای از پاسخ‌های متمایز مسیرهای کامل اعمال شده به تاپل‌های درخت است. یک تراکنش با مجموعه‌ای از آیتم‌های مرتبط با عناصر برگ یک تاپل درخت خاص مدل می‌شود. بینش پشت همچین مدلی اساسا در تعریف خود تاپل نهفته است: هر مسیر اعمال شده به سه تاپل پاسخ منحصربفردی را دریافت می‌کند، بنابراین هر ایتم در یک تراکنش اطلاعاتی را بر مفهومی نشان می‌دهد که از آیتم‌های دیگر در همان تراکنش متمایز است.

تعریف 2.2. فرض کنید τ یک تاپل درخت باشد و $p \in \mathcal{P}_\tau$ یک مسیر کامل در τ. در دوتایی $e = (p, \mathcal{A}_\tau(p))$ یک آیتم تاپل درخت را در τ تعریف می‌کند.

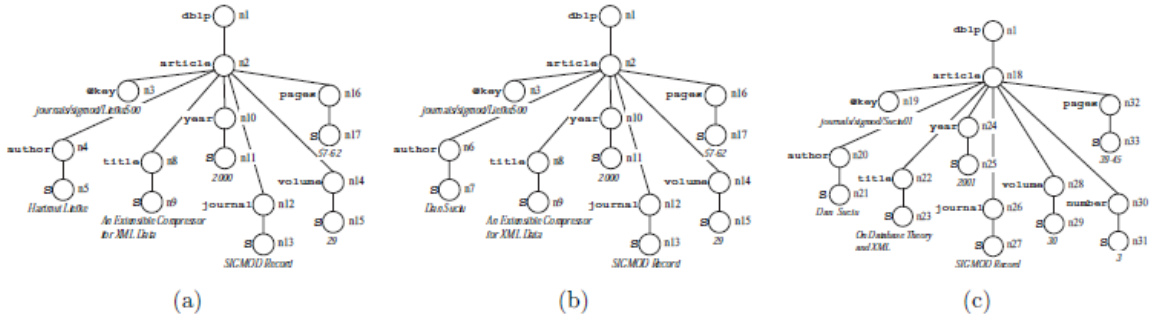
فرض کنید $\mathcal{I}_\tau = \{(p, \mathcal{A}_\tau(p)) \mid p \in \mathcal{P}_\tau\}$ به مجموعه‌ای از آیتم‌های یک تاپل درخت τ اشاره می‌کند. با توجه به XT درخت XML، مجموعه‌ای از آیتم‌های مرتبط به عنوان $\mathcal{I}_{XT} = \bigcup_\tau \mathcal{I}_\tau$، با $\tau \in \mathcal{T}_{XT}$ تعریف می‌شود. در همین قیاس، با توجه به مجموعه $\mathcal{XT} = \{XT_1, \ldots, XT_n\}$ از درخت‌های XML، مجموعه آیتم‌های مرتبط، یا دامنه آیتم، $\mathcal{I}_{\mathcal{XT}} = \bigcup_{XT \in \mathcal{XT}} \mathcal{I}_{XT}$ تعریف می‌شود.

مجموعه $\mathcal{I}_\tau$ از آیتم‌های تاپل درخت در τ به تراکنش XML مرتبط با τ، اشاره دارد. با توجه به مجموعه $\mathcal{XT} = \{XT_1, \ldots, XT_n\}$ از درخت‌های XML، مجموعه داده S تراکنش XML به عنوان $\mathcal{S} = \bigcup_{XT \in \mathcal{XT}} \mathcal{S}_{XT}$ اشاره می‌شود، به طوری که $\mathcal{S}_{XT} = \{\mathcal{I}_\tau \mid \tau \in \mathcal{T}_{XT}\}$ برقرار است.

برخی از نشانه‌ها می‌تواند با توجه به تاثیر تراکنش‌های XML بر جنبه‌های مهم عناصر از دست رفته و تکه تکه شده‌ای تاکید کند.



شکل 1 : مثالی از سند DBLP XML و درخت آن



شکل 2: تاپل‌های درخت استخراج شده از درخت XML شکل 1 (b) (page:صفحه؛ Journal: مجله؛ Volume: جلد؛ year:سال؛ author:نویسنده؛ title: عنوان)

ناهمگنی بالا در الگوهای رابطه‌ای، که می‌تواند از الگوهای XML تحت معیار نگاشت مشتق شود، در کل، مقادیر از دست رفته زیادی را با خود به همراه دارد؛ تراکنش‌ها بر این عیب سنتی غلبه می‌کنند، چرا که آیتم‌ها همیشه به مقادیر غیر خالی (non-null) اشاره دارند. تکه تکه شدن مربوطه، ممکن است نتیجه منفی تجزیه بیش از حد اسناد XML اصلی باشد، و باعث سربار جدی در ارزیابی پرس و جو شود؛ به هر حال، تاپل‌های درخت XML به صورت قابل ملاحظه‌ای به عنوان "ویژگی‌هایی" برای خوشه بندی داده نیمه ساختار یافته عمل می‌کند، در حالی که پردازش بیشتر پرس و جو می‌توانند بر اسناد XML اصلی انجام شود.

مثال 2. به منظور مدل کردن تاپل‌های درخت به عنوان یک تراکنش، هر تاپل می‌تواند به دو مسیر متمایز و پاسخ مربوطه تقسیم شود، که در شکل 3(a) نشان داده شده است. برای مثال، اپلیکیشن dblp.article.@key مقدار

صفات ’journals/sigmod/LiefkeS00‘ را متناظر با گره n_3 از τ_1 حاصل می‌کند، آیتم e_1 سپس با این مسیر-پاسخ همراه است. هم اکنون، پاسخ $dblp.article.journal.S$ رشته ’SIGMOD Record‘ متناظر با گره‌های $n_{27}\ of\ \tau_3$ و $n_{13}\ of\ \tau_1\ and\ \tau_2$ است.

زمانی که دامنه آیتم به صورت کامل تعریف می‌شود، یک تراکنش به هر تاپل درخت با نگاشت دوتایی مسیر-پاسخ آن در آیتم‌های متناظر نسبت داده می‌شود. یک نمایش تراکنشی از تاپل‌های درخت شکل 2 در شکل 3(c) نشان داده شده است.

3. استخراج ویژگی‌های XML

3.1 ویژگی‌های ساختار

مسیرهای تگ مبنای طبیعی را برای استخراج ویژگی‌های ساختاری از داده XML نشان می‌دهند. به هر حال، اگر چه اطلاعات مهمی می‌تواند از تگ‌های XML استنتاج شود، عوامل ذهنی سَبک نویسندگان را در کدینگ اطلاعات برای XML منعکس می‌کند، بنابراین سازگاری اطلاعات در میان داده XML معمولاً تضمین نمی‌شود. ایده کلیدی در فراتر رفتن از یک استفاده مستثنی از متن نهفته است، برای مثال نام تگ، با نگاشت آن به مفاهیم مرتبط معنایی و هر مفهوم به فضای هستی شناسی مربوط است و با یک معنی لغوی، یا حس، مربوط به نام تگ نشان داده می‌شود. دانش هستی شناسی واژگان می‌تواند از این رو برای غنی سازی معنایی ویژگی‌ها استخراج شده از نام‌های تگ XML مورد استفاده قرار گیرد.

path (p)	$\mathcal{A}_{T_1}(p)$	*node ID*
dblp.article.@key	'journals/sigmod/LiefkeS00'	n_3
dblp.article.author.S	'Hartmut Liefke'	n_5
dblp.article.title.S	'An Extensible Compressor ...'	n_9
dblp.article.year.S	'2000'	n_{11}
dblp.article.journal.S	'SIGMOD Record'	n_{13}
dblp.article.volume.S	'29'	n_{15}
dblp.article.pages.S	'57-62'	n_{17}

path (p)	$\mathcal{A}_{T_2}(p)$	*node ID*
dblp.article.@key	'journals/sigmod/LiefkeS00'	n_3
dblp.article.author.S	'Dan Suciu'	n_7
dblp.article.title.S	'An Extensible Compressor ...'	n_9
dblp.article.year.S	'2000'	n_{11}
dblp.article.journal.S	'SIGMOD Record'	n_{13}
dblp.article.volume.S	'29'	n_{15}
dblp.article.pages.S	'57-62'	n_{17}

path (p)	$\mathcal{A}_{T_3}(p)$	*node ID*
dblp.article.@key	'journals/sigmod/Suciu01'	n_{19}
dblp.article.author.S	'Dan Suciu'	n_{21}
dblp.article.title.S	'On Database Theory ...'	n_{23}
dblp.article.year.S	'2001'	n_{25}
dblp.article.journal.S	'SIGMOD Record'	n_{27}
dblp.article.volume.S	'30'	n_{29}
dblp.article.number.S	'3'	n_{31}
dblp.article.pages.S	'39-45'	n_{33}

(a)

item ID	*associated node IDs*
e_1	n_3
e_2	n_5
e_3	n_9
e_4	n_{11}
e_5	n_{13}, n_{27}
e_6	n_{15}
e_7	n_{17}
e_8	n_7, n_{21}
e_9	n_{19}
e_{10}	n_{23}
e_{11}	n_{25}
e_{12}	n_{29}
e_{13}	n_{31}
e_{14}	n_{33}

(b)

tr_1	e_1 e_2 e_3 e_4 e_5 e_6 e_7
tr_2	e_1 e_8 e_3 e_4 e_5 e_6 e_7
tr_3	e_9 e_8 e_{10} e_{11} e_5 e_{12} e_{13} e_{14}

(c)

شکل 3 : نمایش تراکنشی از تاپل‌های درخت شکل 2؛ (a) تاپل‌های درخت با مسیرها و پاسخ ها، (b) دامنه آیتم، و (c) مجموعه تراکنشی

با توجه به افزایش دامنه و دسترسی عمومی، پایگاه داده واژگان WordNet در این کار به عنوان پایگاه دانش هستی شناسی استفاده می‌شود. گروه کلمات WordNet با همان معنی در کلاس‌های هم ارزی، به نام synsets (مجموعه مترادف) استفاده می‌شود.

هر مجموعه مترادف مفهومی را نشان می‌دهد و با تشریح متنی مختصری به نام "تفسیر" (gloss)، توصیف می‌شود. به طور ویژه، synsets اسم از طریق is-a (فراشمول/معنا شمول) و رابطه part-of (جز واژه/کل واژه) به هم مرتبط هستند. ما انتظار داریم که اساسا از بخش اسم WordNet استفاده کنیم، چرا که اسم‌ها به شدت بیشتر برای یادداشت داده XML استفاده می‌شوند.

ابهام زدایی حس تگ. نگاشت نام تگ در یک فضای مفهوم هستی شناسی نیاز است که مسئله تصمیم گیری برای مناسب ترین موارد برای هر نام تگ را بررسی کند. این می‌تواند با اجرای ابهام زدایی حس کلمه (WSD) به انجام رسد، که کلمه‌ای را با یک سنجش مبتنی بر زمینه‌ای که کلمه در آن ظاهر می‌شود نسبت می‌دهد.

رویکرد پیشنهادی برای ابهام زدایی حس نام‌های تگ شامل انتخاب با توجه به مناسب‌ترین حس، یک مسیر- متن است. یک مسیر-متن با شبکه معنایی ایجاد شده بر همه موارد ممکن مرتبط با تگ‌های مسیر خاص ساخته شده است. برای هر مسیر تگ $p = t_1.t_2. \ldots .t_n$، یک گراف وزن دار مستقیم $SG(p)$، به نام گراف synset از p، برای ابهام زدایی از حس تگ‌هایی در p استفاده شده است. $SG(p)$ به شرح زیر تعریف می‌شود:

- گره‌ها دوتایی $\langle t_i, \sigma \rangle$، با $i \in [1..n]$ و $\sigma \in senses(t_i)$ هستند، که $senses(t_i)$ به مجموعه حس‌ها برای کلمه (تگ) t_i اشاره دارد؛ علاوه بر این؛ گره‌های اضافی مبدا (source) و حفره (sink) در راستای ملاقات معمولی در گراف، داده شده است.
- یال‌ها ارتباطی بین گره‌های پیوسته $(\langle t_i, \sigma \rangle, \langle t_{i+1}, \rho \rangle)$، با $i \in [1..n\text{-}1]$ است؛ علاوه بر این، یال‌های $(source, \langle t_1, \sigma \rangle)$ و $(\langle t_n, \rho \rangle, sink)$، برای هر $\sigma \in senses(t_1), \rho \in senses(t_n)$ برقرار هستند.
- وزن‌های یال برای انعکاس ارتباط معنایی بین مفاهیم مرتبط با هر دو گره $\langle t_i, \sigma \rangle, \langle t_{i+1}, \rho \rangle$ منعکس می‌شود؛ وزن‌ها بر یال‌ها شامل مبدا یا حفره‌ای است که به 0 تنظیم می‌شود.

زمانی SG(p) ایجاد شد، ابهام زدایی نام‌های تگ در p با یافتن مسیر حداکثر وزن در SG(p) انجام می‌شود، لذا مناسبترین حس، موارد متناظر با این مسیر گراف هستند. در مورد چندین مسیر گراف با حداکثر وزن، موردی ترجیح داده می‌شود که می‌تواند با استفاده از ترتیب خطی عرضه شده دیکشنری از synsets بر اساس شناسه لغوی برای حواس استفاده شود. مشاهده اینکه محاسبه مسیر با حداکثر وزن در تعدادی از یال‌ها، خطی است؛ با توجه به لایه‌هایی از گراف synset شکل می‌گیرد.

جنبه مهم ساخت گراف synset چگونگی محاسبه وزن‌های یال، و چگونگی محاسبه ارتباط معنایی بین حواس هستند. ما این جنبه را به صورت دقیق بررسی می‌کنیم.

در WSD مبتنی بر دیکشنری فرض این است که ممکن ترین مورد برای اختصاص کلمات هم رخداد متعدد حداکثرسازی ارتباط در میان حس‌های انتخاب شده است. از این دیدگاه، روش پیشگام Lesk [10] از کلمه هدف با انتخاب معنا ابهام زدایی می‌کند و تفسیر بیشترین تعداد از کلمات را با تفسیر مرتبط با کلمات مجاور به اشتراک می گذارد.

استفاده از هستی شناسی لغوی، مانند WordNet، اتخاذ روابط معنایی را با تکیه بر بهره وری سلسله مراتب مفاهیم در کنار تفسیری از دیکشنری اتخاذ می‌کند. الگوریتم پایه Lesk می‌تواند از این رو به منظور اتخاذ مزایای شبکه روابط ارائه شده در WordNet اتخاذ شود. این ایده در یک معیار ارتباط معنایی بین حس کلمات بر اساس مفاهیم همپوشان تفسیر شده به رسمیت شناخته می‌شود، که معیاری از شایستگی‌های تطبیق عبارت است و وزن آن‌ها بسیار سنگین تر از تطبیق کلمات تکی است. معیار همپوشانی تفسیر بسط یافته دو مفهوم را به عنوان ورودی می‌گیرد (دو WordNet synsets) و امتیاز همپوشانی تفسیر را محاسبه می‌کند، در اینجا به امتیاز (score)، به عنوان مجموعه‌ای از اندازه مربع همپوشانی متمایز بین تفسیرها اشاره می‌شود. یک همپوشانی زمانی که حداکثر دنباله از کلمات اشتراک گذاشته شده رخ دهد تشخیص داده می‌شود. در نهایت ارتباط معنایی بین دو synsets a و b سنجیده می‌شود، همپوشانی توابع امتیاز دهی تفسیر نه تنها برای مقایسه تفسیر a و تفسیر b بلکه برای جفت تفسیرها از synsets مقایسه می‌شود که a و b مستقیماً از طریق یک رابطWordNet اصلی به هم مرتبط هستند. به صورت دقیق تر، مقایسه شامل ترکیب فراشمول a با فراشمول b’s، معنا شمول a با معنا شمول b؛ فراشمول a با b، و a با فراشمولb’s است. جزییات بیشتر را می‌توانید در [4] ببنید.

در ادامه این راستا، مکانیسم بالا برای امتیازدهی همپوشانی تفسیرها را حفظ می‌کنیم، در عین حال به صورت متفاوتی از طبقه بندی مفهوم WordNet برای تعیین روابط بین synsets استفاده می‌شود. از یک طرف؛ ترکیبات متفاوت روابط بر synsets دیده می‌شود. براستی، نام‌های تگ‌ها در یک مسیر XML به ترتیبی که برای استنتاج سلسله مراتب تحلیل مجاز است رابطه دارد/ از طرفی دیگر، تنها جهت مستقیم را اتخاذ نمی کنیم بلکه به سمت synset هدف هم مسیرهایی را اتخاذ می کنیم.

فرض کنید WSR به مجموعه انتخابی از روابط WordNet synset اشاره دارد. فرض کنیم WSR شامل روابط زیر است : فراشمول، معنا شمول، جز واژه و کل واژه. با توجه به روابط $r \in \mathcal{WSR}$ و a synset، تابع w را طوری تعریف می کنیم که، به r و a اعمال می‌شود، و مجموعه w(r,a) از synsets مستقیماً از یک a از طریق روابط WordNet r حاصل می‌شود.

تابع w می‌تواند تا جایی بسط یابد که شامل synsets است، که غیر مستقیم به زیر مجموعه هدف در یک فاصله داده شده در طبقه بندی WordNet اشاره دارد. با توجه به دو a synsets و a' و یک رابطه WordNet ثابت، یک دسته بندی از a به a' منجر می‌شود. با توجه به یک رابطه $r \in \mathcal{WSR},$، یک a synsets. یک مقدار عدد صحیح $\overline{d},$ ؛ مجموعه‌ای از زوج‌های (a', d) است؛ به طوری که a' synsets مرتبط به a، از طریق r، در طول مسیر $d \leq \overline{d}$ به شرح زیر تعریف می‌شود:

$$\omega^*(r, a, \overline{d}) = \begin{cases} \bigcup_{a' \in \omega(r,a)} \omega^*(r, a', \overline{d} - 1) & \text{if } \overline{d} > 1 \\ \omega(r, a) & \text{if } \overline{d} = 1 \end{cases} .$$

فرض کنید $a = \langle t_i, \sigma \rangle$ و $b = \langle t_{i+1}, \rho \rangle$ دو synsets متناظر با گره‌هایی در یک گراف synsets است، و فرض می‌کند $\overline{d}$ یک مقدار عدد صحیح است. وزن یال بین a و b به شرح زیر محاسبه می‌شود:

$$\begin{aligned} & weight(a, b, \overline{d}) = score(a, b) \\ & + \sum_{(b', d) \in \omega^*(\text{hype}, b, \overline{d}) \ \vee \ (b', d) \in \omega^*(\text{holo}, b, \overline{d})} score(a, b') \times \text{f}(d) \\ & + \sum_{(a', d) \in \omega^*(\text{hypo}, a, \overline{d}) \ \vee \ (a', d) \in \omega^*(\text{mero}, a, \overline{d})} score(a', b) \times \text{f}(d), \end{aligned}$$

در این رابطه f(d) یک تابع است که با افزایش مقدار d به صورت یکنواخت کاهش می‌یابد. در محیط آزمایشی ما، f(d) را یک تابع نمایی معکوس ثابت می‌کنیم، در حالی که حداکثر مقدار فاصله $\overline{d}$ برابر با 3 است. توجه داشته باشید که weight(a, b) مانند weight(b, a) است. این عدم تقارن در زمینه مسیر XML - در تناظر با نام تگ که معمولا یک فراشمول یا معنا شمول از نام تگ مربوط به b را نشان می‌دهد- توجیه شده است.

3.2 ویژگی‌های محتوا

ویژگی‌های محتوا با کشف الگوهایی از عناصر متنی تولید می‌شوند. ما به واحد محتوای متنی (اختصار TCU) به عنوان متن استخراج شده از هر گره برگ یک درخت XML اشاره می‌کنیم (برای مثال، یک محتوای عنصری #PCDATA یا یک مقدار صفت). TCUها می‌توانند با پذیرش یک مدل کیسه‌ای از کلمات نشان داده شوند، و در معرض تحلیل لغوی و معنایی قرار گیرد. مورد اول انتخاب ویژگی‌های معنایی قابل توجه (عبارت شاخص)، را با ابزارهای عملیات پیش پردازش متن با زبان خاص هدف قرار می‌دهد، مانند حذف stopwords و ریشه یابی کلمات. علاوه بر این، تحلیل معنایی می‌تواند برای نفوذ بر مسئله ابهام زدایی لغوی با بررسی درجه‌ای از چند معنایی عبارات اعمال شود.

دو وظیفه تحلیل متن وزن ارتباطی برای عبارات شاخص را تعیین می‌کند. براستی، ما می‌توانیم ارتباط نحوی و ارتباط معنایی را در نظر بگیریم. ارتباط نحوی معمولاً برطبق ارتباطی بر فرکانس وقوع عبارت، با توجه به زمینه محتوا وزن می‌شود. از طرفی دیگر، اطلاعات نحوی می‌توانند "به صورت غنی از نظر معنایی" به مفاهیم عباراتی که از نظر معنایی نادر هستند توسل جویند. این نکته را در بخش بعدی بررسی می کنیم.

4. خوشه بندی تراکنشی XML

4.1 تشابه آیتم در تاپل درختی XML

در تنظیم ما ویژگی‌های XML در آیتم‌های تاپل درختی تعبیه شده‌اند. مفهوم شباهت بین آیتم‌های تاپل درختی تابعی از تشابه بین ساختار چشم انداز آن‌ها و ویژگی‌های محتوایی هستند.

تعریف 4.1. فرض کنید e_i و e_j دو آیتم تاپل درخت باشند. تابع تشابه آیتم تاپل درخت به شرح زیر تعریف می‌شود:

$$sim(e_i, e_j) = f \times sim_S(e_i, e_j) + (1 - f) \times sim_C(e_i, e_j),$$

در این رابطه sim_S $(resp.$ $sim_C)$ به تشابه $(resp. content)$ ساختاری بین آیتم‌ها اشاره دارد، و $f \in [0..1]$ فاکتوری است که نفوذ بخش ساختاری را با تشابه کلی وفق می‌دهد.

علاوه بر این، فرض کنید $\gamma \in [0..1]$ آستانه تشابه است. می‌گوییم که دو آیتم تاپل درختی e_i و e_j γ-matched هستند اگر و تنها اگر دارای یک "تطبیق" در درجه‌ای بزرگتر یا برابر γ باشد، به همین ترتیب داریم $sim(e_i, e_j) \geq \gamma$ توجه داشته باشید که تشابه آیتم در تاپل درخت به عنوان تابع خطی تعریف می‌شود چرا که مستقیماً ما را قادر می‌سازد که سهم‌های متفاوت آمده از ویژگی‌های محتوایی و ساختاری را کنترل و درک کنیم. در ادامه بینشی را در تشابه ساختاری و محتویی بین آیتم‌های تاپل درختی کسب می‌کنیم.

تشابه در ساختار: بین دو آیتم تاپل درخت e_i و e_j ارزیابی شده در مقایسه با مسیرهای تگ چشم انداز آن‌ها و محاسبه تشابه میانگین بین حس بهترین تگ‌های منطبق ارزیابی می‌شود. با توجه به دو مسیر p_i, p_j و یک تگ $t \in p_i$، به $bm(p_j, t) = \{t' \in p_j \mid \nexists t'' \in p_j,\ t'' \neq t',\ sim(t.\sigma, t''.\sigma) > sim(t.\sigma, t'.\sigma)\}$ به عنوان مجموعه‌ای از تگ‌های p_j اشاره شد که t دارای بهترین تطبیق است. فرض می کنیم که هر هدف t با مناسب ترین مورد $t.\sigma$ مرتبط است که با اجرای یک فرآیند از ابهام زدایی حس تگ انتخاب شده است، که در بخش 3.1 تشریح شد.

تعریف 4.2: e_i و e_j دو آیتم در تاپل درخت هستند، و $p_i = t_{i_1}.t_{i_2}\ldots\ .t_{i_n},\ p_j = t_{j_1}.t_{j_2}\ldots\ .t_{j_m}$ مسیر تگ‌های مربوطه است. تشابه ساختاری بین e_i و e_j به شکل زیر تعریف شده است

$$sim_S(e_i, e_j) = \frac{1}{n+m}\Bigg(\sum_{t \in p_i} \sum_{t' \in bm(p_j,t)} \frac{sim(t.\sigma, t'.\sigma)}{|bm(p_j, t)|} + \sum_{t \in p_j} \sum_{t' \in bm(p_i,t)} \frac{sim(t.\sigma, t'.\sigma)}{|bm(p_i, t)|} \Bigg),$$

در این رابطه $sim(t.\sigma, t'.\sigma)$ تشابه بین حس‌های مربوطه تخصیص داده شده به تگ t و t’ را محاسبه می‌کند.

به منظور تعیین معیار مناسب تشابه بین حس نام تگ، از طول مسیرها در طبقه بندی مفهوم و هم رخدادی حس استفاده می‌کنند. این رویکرد در دو مشاهده اصلی نهفته است.

در یک سلسله مراتب از مفهوم ها، معیار مبتنی بر مسیر برای تعیین درجه‌ای از دو مفهوم مرتبط مجاز است. به هر حال، طول مسیر باید به صورت متفاوتی بسته به محل مفهوم در سلسله مراتب تفسیر شود، چرا که یک مفهوم بالاتر

در سلسله مراتب عمومی تر است. یک معیار مبتنی بر مسیر مناسب در [14] تعریف شده است، که بر مفهوم کمترین رده بندی مشترک (lcs) برای محاسبه مناسب ترین مورد به اشتراک گذاشته شده توسط دو حس تمرکز می‌کند.

علاوه بر روابط معنایی بین حس کلمه، به تخمین ویژگی حس علاقه داریم. به صورت دقیق تری، ویژگی حس σ را در سراسر وقوع هر نام تگ مرتبط با σ در مجموعه داده شده تخمین می‌زنیم.

در مواجهت با مشاهدات بالا، تشابه بین حس σ_1 and σ_2 با ترکیب معیار هم رخدادی و مبتنی بر مسیر محاسبه می‌شود.

تعریف 4.3. فرض کنید S مجموعه‌ای از تراکنش‌های XML باشد؛ $\sigma_1,\ \sigma_2$ دو حس تگ باشد. تشابه حس بین σ_1 و σ_2 با توجه به S به شرح زیر تعریف شده است:

$$sim(\sigma_1,\sigma_2) = \frac{2 \times depth(lcs(\sigma_1,\sigma_2))}{depth(\sigma_1)+depth(\sigma_2)} \times \frac{freq(\sigma_1,\sigma_2,\mathcal{S})}{freq(\sigma_1,\mathcal{S})+freq(\sigma_2,\mathcal{S})-freq(\sigma_1,\sigma_2,\mathcal{S})},$$

در این معادله $depth(\sigma)$. فاصله از گره مفهوم برای σ تا ریشه سلسله مراتب است، $freq(\sigma_i,\mathcal{S})$ تعداد تراکنش XML در S است که شامل تگ t_i است به طوری که $t_i.\sigma = \sigma_i$، و $freq(\sigma_1,\sigma_2,\mathcal{S})$ به تعداد تراکنش‌های XML در S اشاره می‌کنند که شامل تگ t_i و t_j است، به طوری که $t_i.\sigma = \sigma_1$ و $t_j.\sigma = \sigma_2$ برقرار است.

تشابه با محتوا. عبارات شاخص از TCU استخراج می‌شوند که می‌تواند بر طبق روابط معنایی و نحوی وزن شود. از نقطه نظر نحوی، دو معیار آماری را در نظر میگیریم: تراکم عبارت در یک متن و نادر بودن عبارت در مجموعه متون. تابع وزن tf.idf متداول تنها برای در نظر گرفتن هردو معیار تعریف شده است. در تنظیمات ما، با توجه به هر عبارت w_j و هر TCU u_i، وزن tf.idf به شکل زیر محاسبه می‌شود:

$$tf.idf(w_j,u_i) = freq(w_j,u_i) \times \log\left(\frac{N}{n_j}\right),$$

در این معادله $freq(w_j, u_i)$ به تعداد وقوع w_j در u_i اشاره دارد، N کل تعداد TCUها است که در مجموعه تاپل‌های درخت شمرده شده است، و n_j تعداد TCUهای شامل w_j است.

تابع وزن عبارت معمولی تنها کلمات را برای ارزیای ارتباط بین کلمات در نظر می‌گیرد. ایده ما غنی سازی تابع tf.idfبا بهره برداری از نادر بودن عبارت از نظر معنایی است. این مفهوم را با توسل به درجه‌ای از چند معنایی بودن یک عبارت تعریف می‌کنیم، به صورتی که رابطه عبارت از تعداد معانی که یک عبارت دارد بکاهد. به صورت رسمی، نادر بودن معنایی عبارت w به شکل زیر ارزیابی می‌شود

$$s\text{-}rarity(w) = \log\left(\frac{\text{MAX\_POLYSEMY}}{|senses(w)|}\right),$$

در این معاله $senses(w)$ به مجموعه معانی w اشاره دارد و MAX_POLYSEMY یک اشاره ثابت به تعدادی از معنی دارترین معانی کلمه در WordNet اشاره دارد. تابع لگاریتمی شامل اثرات طرفداری عبارت با معنای کمتر معمولی است. علاوه بر این، فرض می‌کنیم که هر عبارت حداقل یک معنی دارد، حتی اگر آن عبارت در دیکشنری مرجع ارائه نشده باشد.

تابع نادر بودن معنایی با توجه به موقعیت عبارت در مجموعه واحدها با محتوای متنی ثابت است. بنابراین، برای هر عبارت w_j، $s\text{-}rarity(w_j)$ همان مقدار می‌تواند با مقدار وابسته tf.idf برای بسیاری از دوتایی‌های (w_j, u_i) ترکیب شده باشد.

تعریف 4.4. وزن مرتبط عبارت با توجه به w_j ، یک واحد محتوای متنی u_i به شکل زیر محاسبه شده است:

$$relevance(w_j, u_i) = tf.idf(w_j, u_i) \times s\text{-}rarity(w_j).$$

یک معیار به طور ویژه مناسب برای ارزیابی تشابه بین اسناد؛ تشابه کسینوسی است: این تقریبی را از نظر کسینوس زاویه‌ای از دو شکل سند عمومی در یک فضای چند بعدی استفاده می‌کند. هر TCU u_i می‌تواند با بردار عبارت $\vec{u}_i$ که jامین مولفه شامل مقدار $relevance(w_j, u_i)$. است رابطه داشته باشد. با توجه به دو عنصر XML

e_i and e_h ، تشابه محتوا بین e_i and e_h با سنجش تشابه کسینوسی بین بردار عبارات مرتبط با TCUهای مربوطه محاسبه می شود: $sim_C(e_i, e_h) = \frac{\vec{u}_i \cdot \vec{u}_h}{\|\vec{u}_i\| \times \|\vec{u}_h\|}$.

4.2 الگوریتم میانگین – XTrK

تاپل‌های درخت XML مدل شده به عنوان تراکنش می‌تواند به صورت موثری با اعمال یک الگوریتم تفکیکی توصیه شده برای دامنه تراکنش XML خوشه بندی شود.

یک مسئله خوشه بندی تفکیکی شامل بخش‌هایی از یک مجموعه $\{x_1, \ldots, x_n\}$ از اشیای k گروه غیر تهی است که هر یک شامل زیرمجموعه همگن از اشیا است. یک کلاس مهم از رویکردهای تفکیکی بر اساس مفهوم خوشه، یا نمایش است: هر شی x_i نسبت داده شده به خوشه C_j برطبق فاصله آن از مقدار C_j است، که آن را مرکز C_j گوییم.

یک الگوریتم تفکیکی ساده اما موثر برای داده تراکنشی جنریک خوشه‌ای میانگین-TrK است. این الگوریتم شامل دو فاز اصلی است. در اولین فاز، به عنوان روش مبتنی بر مرکز سنتی برای محاسبه $k + 1$ خوشه کار می‌کند: با انتخاب k شی به عنوان مرکز خوشه اول آغاز می‌شود؛ سپس انتساب مجدد تکراری هر شی باقی مانده به نزدیکترین خوشه تا زمانی که مراکز خوشه تغییر نکند ادامه می‌یابد. $(k+1)$امین خوشه را، خوشه سطل زباله[۲] گویند، که برای اشیای درک شده، ایجاد شده است، برای مثال، اشیایی که با هیچ خوشه‌ای اشتراک ندارند و برای همین به هیچ یک از k خوشه اول انتساب نمی یابند. فاز دوم مسئول تقسیم بازگشتی خوشه سطل زباله به تعداد خوشه‌های کوچک است.

الگوریتم میانگین- TrK یک نماد مناسب از مرکز خوشه را برای استفاده از یک مدل نمایش فشرده برای اساس تقویت اشتراک تراکنش‌ها در یک خوشه ارائه داده است. براستی، زمانی که دو تراکنش را مقایسه می کنیم به دنبال آیتم‌های "مشترک" هستیم. ابتدا، می توانیم تشابه بین دو تراکنش را مستقیما با تناسب تعداد آیتم‌های مشترک و تناسب عکس با آیتم‌های متفاوت ارزیابی کنیم. میانگین-TrK از ضریب جاکارد برای محاسبه تشابه بین دو مجموعه به عنوان نسبت کاردینالیتی اشتراک آن‌ها با کاردینالیتی واحد آن‌ها استفاده می‌شود.

---

[2] trash cluster

به هر حال، محاسبه اشتراک دقیق بین تراکنش‌های آیتم‌های تاپل درخت ممکن است موثر نباشند. دو آیتم حتی اگر یکسان نباشند؛ ممکن است ساختارهای مربوطه و اطلاعات محتوا را به اشتراک گذارند. در راستای این هدف، مفهوم اشتراک استاندارد بین مجموعه آیتم‌ها را با توانایی اتخاذ حتی حداقل تشابهات از ویژگی‌های محتوایی و ساختاری عناصر XML ارتقا دادیم.

تعریف 4.5. فرض کنید tr_1 و tr_2 دو تراکنش XML هستند، و $\gamma \in [0..1]$ یک آستانه تشابه است. مجموعه آیتم‌های اشتراک- γ بین tr_1 و tr_2 به شرح زیر تعریف شده است:

$$match^{\gamma}(tr_1, tr_2) = match^{\gamma}(tr_1 \rightarrow tr_2) \cup match^{\gamma}(tr_2 \rightarrow tr_1),$$

$$where$$

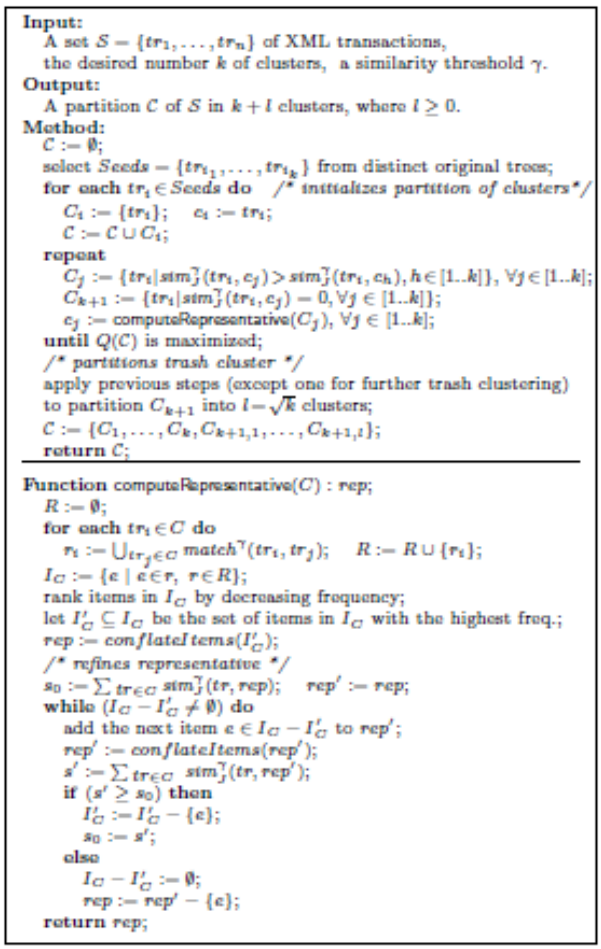
$$match^{\gamma}(tr_i \rightarrow tr_j) = \{e \in tr_i \mid \exists e_h \in tr_j,\ sim(e, e_h) \geq \gamma, \nexists e' \in tr_i, sim(e', e_h) > sim(e, e_h)\}.$$

مجموعه آیتم‌های اشتراک- γ اشتراک بین تراکنش را به درجه بالاتر یا برابر با آستانه تشابه γ نگاشت می‌کنند.

تعریف 4.6. فرض کنید tr_1 و tr_2، دو تراکنش XML باشند، و $\gamma \in [0..1]$ آستانه تشابه بین tr_1 و tr_2 باشد که به شرح زیر تعریف می‌شود:

ما پذیرفتیم که الگوریم میانگین-TrK بر تلاش اصلی درک مفاهیم مناسب نزدیک در میان تراکنش‌های XML و نمایش خوشه تمرکز می‌کند. شکل 4 فازهای اصلی الگوریتم نتیجه، به نام میانگین-XTrK است، که ویژگی برجسته آن در زیر بحث شده است.

در درجه اول، تابع تشابه XML ضریب جاکارد سنتی را به عنوان معیار تقریبی بین تراکنش‌ها به اشتراک می‌گذارد.



Input:
A set $S = \{tr_1, \ldots, tr_n\}$ of XML transactions,
the desired number k of clusters, a similarity threshold γ.
Output:
A partition $\mathcal{C}$ of S in $k+l$ clusters, where $l \geq 0$.
Method:
$\mathcal{C} := \emptyset$;
select $Seeds = \{tr_{i_1}, \ldots, tr_{i_k}\}$ from distinct original trees;
for each $tr_i \in Seeds$ **do** /* *initializes partition of clusters* */
$C_i := \{tr_i\}$; $c_i := tr_i$;
$\mathcal{C} := \mathcal{C} \cup C_i$;
repeat
$C_j := \{tr_i | sim^{\gamma}_J(tr_i, c_j) > sim^{\gamma}_J(tr_i, c_h), h \in [1..k]\}, \forall j \in [1..k]$;
$C_{k+1} := \{tr_i | sim^{\gamma}_J(tr_i, c_j) = 0, \forall j \in [1..k]\}$;
$c_j :=$ computeRepresentative(C_j), $\forall j \in [1..k]$;
until $Q(\mathcal{C})$ is maximized;
/* *partitions trash cluster* */
apply previous steps (except one for further trash clustering)
to partition C_{k+1} into $l = \sqrt{k}$ clusters;
$\mathcal{C} := \{C_1, \ldots, C_k, C_{k+1,1}, \ldots, C_{k+1,l}\}$;
return $\mathcal{C}$;

Function computeRepresentative(C) : rep;
$R := \emptyset$;
for each $tr_i \in C$ **do**
$r_i := \bigcup_{tr_j \in C} match^{\gamma}(tr_i, tr_j)$; $R := R \cup \{r_i\}$;
$I_C := \{e \mid e \in r,\ r \in R\}$;
rank items in I_C by decreasing frequency;
let $I'_C \subseteq I_C$ be the set of items in I_C with the highest freq.;
$rep := conflateItems(I'_C)$;
/* *refines representative* */
$s_0 := \sum_{tr \in C} sim^{\gamma}_J(tr, rep)$; $rep' := rep$;
while $(I_C - I'_C \neq \emptyset)$ **do**
add the next item $e \in I_C - I'_C$ to rep';
$rep' := conflateItems(rep')$;
$s' := \sum_{tr \in C} sim^{\gamma}_J(tr, rep')$;
if $(s' \geq s_0)$ **then**
$I'_C := I'_C - \{e\}$;
$s_0 := s'$;
else
$I_C - I'_C := \emptyset$;
$rep := rep' - \{e\}$;
return rep;

شکل 4: الگوریتم میانگین- XTrK

در درجه اول، تابع تشابه تراکنش XML ضریب جاکارد سنتی به عنوان یک ابزار تخمین در میان تراکنش‌ها جایگزین می‌شود.

دوم؛ نمایش هر خوشه C با شروع از مجموعه آیتم‌های مشترک $^{\gamma}$- در میان همه تراکنش‌ها در C محاسبه می‌شود. به صورت دقیق تر، برای هر تراکنش در C، واحد مجموعه آیتم مشترک $^{\gamma}$- با توجه به همه دیگر تراکنش‌ها در C بدست آمده است، این هیچ وابستگی را در بررسی تراکنش‌ها منعکس نمی‌کند. سپس، یک نماینده اولیه با انتخاب مجموعه اجتماع آیتم‌ها با بالاترین فرکانس محاسبه می‌شود. نماینده اولیه؛ ممکن است فرمی از تاپل‌های درخت را داشته باشد، چرا که برخی از آیتم‌ها ممکن است به یک مسیر اشاره کنند اما پاسخ‌های متفاوتی داشته باشمد. تابع conflateItems به باقی آیتم ها، برای هر زیر مجموعه از I آیتم اشتراک گذاری یک مسیر p اعمال شده است، که

یک آیتم را به دست می‌آورد که دارای p به عنوان مسیر و الحاق محتویات به آیتم‌ها در I به عنوان محتوای آن است. سرانجام، یک اکتشاف حریصانه نمایش فعلی را با اضافه کردن تکراری آیتم‌هایی با بیشترین تکرار تا زمانی که مجموعه تشابه دو به دو بین تراکنش و نمایش بیشتر از حداکثر نشود، تصحیح می‌کند. دوباره، هر تصحیحی باید تضمین کند که نمایش نتیجه شده یک تاپل درخت است.

علاوه بر این، تاپل‌های درخت میانگین- XTrK انتخاب شده به عنوان مرکز خوشه اولیه به موارد آمده از اسناد XML متفاوت محدود می‌شود. این استراتژی منفعت ساختار خوشه‌ها با تشابه داخلی پایین را هدف قرار می‌دهد.

سرانجام، معیار موجود در میانگین- XTrK نیازمند کیفیت بخش‌های خوشه نتیجه شده است که حداکثر شده‌اند. به طور معمول در یک خوشه بندی، این به معنی این است که انسجام خوشه حداکثر شده است در حالی که تشابه بین خوشه‌ها حداقل شده است. با اکتشاف تشابهات جفت جفت بین تراکنش ها، که از نمایش خوشه، تشابه بین خوشه‌ای و درون خوشه‌ای (برای بخش C از خوشه K) استفاده شده است به شکل زیر تعریف شده است:

$$IntraSim(\mathcal{C}) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|C_i|}\sum_{tr\in C_i} sim_J^{\gamma}(tr, c_i)$$

$$InterSim(\mathcal{C}) = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} sim_J^{\gamma}(c_i, c_j)}{\frac{1}{2}k(k-1)}.$$

کیفیت خوشه بندی C به عنوان $Q(\mathcal{C}) = IntraSim(\mathcal{C}) - InterSim(\mathcal{C})$ تعریف شود. اگر فرض کنیم که هردو تشابه داخل خوشه و بین خوشه‌ای دارای مقادیری در $[0..1]$ هستند، و $IntraSim(\mathcal{C}) \geq InterSim(\mathcal{C})$ خوشه بندی خوب است، سپس $Q(\mathcal{C})$ از $[0..1]$ است.

واضح است که یک جایگزین، یک معیار خروج ارزان تر، بررسی می‌کند که آیا با توجه به تکرار قبلی مرکز خوشه در تکرار فعلی تغییر می‌کند یا نه.

به هر حال، استفاده از یک معیار اعتبار به ما اجازه می‌دهد که وظایف خوشه را ارزیابی کنیم.

5. ارزیابی آزمایشی

توصیف داده. برای ارزیابی فریم ورک خوشه بندی پیشنهادی، پایگاه داده XML واقعی را با داشتن ویژگی‌های متفاوت بر طبق سه جنبه اصلی در نظر می‌گیریم: ترکیب مجموعه داده (یک سند واحد یا مجموعه اسناد)، پیچیدگی ساختاری از نظر درجه عناصر تو در تو، با توجه به اثرات اندازه و تعداد عناصر متنی. جنبه دوم یک دلیل مهم است که باید مجموعه داده‌های مصنوعی را در نظر نگیریم: براستی، آن‌ها عناصری شامل متن‌های زبان طبیعی منسجم ارائه نمی‌دهند، اما حداقل از یک ژنراتوراتوماتیک استفاده می‌کنند که کلماتی که بیشتر از همه در یک متن نثر ثابت رخ داده اند را به عنوان اشتراکی برای تقلید از توزیع متن آماری واقعی در نظر می‌گیرند.

سه منبع داده XML برای انجام آزمایش‌ها استفاده شدند: آرشیو DBLP، کانال خبری Reuters RSS، و PubMed. DBLP (http://dblp.unitrier.de/xml/) یک کتابشناسی دیجیتال است که اساسا به مقالات روزنامه‌ای، مقالات کنفرانسی، کتاب ها، فصل‌های کتاب، و علوم کامپیوتر مربوط است. نسخه XML این آرشیو با یک سند بزرگ واحد نشان داده می‌شود، که می‌تواند (برطبق DTD) به هزاران سند XML تجزیه شود. کانال خبری Reuters RSS (http://www.microsite.reuters.com/rss/) به کسب و کار، سرگرمی، تکنولوژی، علم مربوط است. هر خبر Reuters RSS از لیستی از آیتم‌ها تشکیل شده است، که هریک یک هدر، یک توصیف، و یک لینک دسترسی به مقاله کامل را دارند. PubMed (http://www.ncbi.nlm.nih.gov/entrez/) سرویسی از کتابخانه ملی پزشکی است، و شامل بیش از 15 میلیون استناد برای مقالات پزشکی است. این معیار از ژورنال‌های اضافی و مدلاین در زمینه بیولوژی و پزشکی آمده است. ما مجموعه‌ای را با جمع آوری اسناد XML به صورت پویا که در نتیجه پرس و جو برای "پروتئین" در موتور جستجوی سایت و دسترسی به همه فیلدهای موجود در .PubMed بدست آمدهاست را جمع آوری می‌کنیم.

سه مجموعه انواع متفاوتی از ساختار و محتوا را ارائه می‌دهند، چرا که آن‌ها خدمات آموزنده مختلفی را با منبع اصلی آن منعکس می‌کنند. Reuters RSS مثال خوبی از خدمات RSS (تشکیل ارتباط‌های بسیار ساده) است، که برای اشتراک گذاری محتوای وب مانند هد لاین خبری به شهرت رسیده است. اسناد در مجموعه Reuters RSS ساختار

بسیار منظمی دارند، یا میانگین عمقی نزدیک به حداکثر عمق دارد. DBLP و PubMed به کتابشناسی علمی اشاره دارد. از چشم اندازی ساختاری DBLP تغییرات بیشتری را نشان می‌دهند، اگر چه با عمق میانگین کوچک مشخص می‌شود و تشریح متن کاملا کوتاهی محدود به عنوان مقاله، موضوع رخداد (نام کامل کنفرانس)، و نام نویسنده را ارائه می‌دهد. PubMed سخت ترین مجموعه از سه مجموعه است، که عناصر عمیقا تو در تو را شامل عناصر حاوی چکیده گزارش با متن‌های بلند را نشان می‌دهد. توجه داشته باشید که موضوعات زیست شناسی مولکولی و پزشکی ممکن است بی نهایت پیچیده باشند: آن‌ها معمولآ شامل عباراتی هستند که در گونه‌های مرتبط به اشتراک گذاشته شده‌اند، و به همین ترتیب عبارات متعددی به صورت ویژه برای گونه‌های زیستی خاص استفاده می‌شود، و ارتباطات جدید ممکن است در میان موضوعات مرتبط قبلی کشف شود.

نتایج و تنظیمات آزمایشی. آزمایش‌ها به منظور تست توانایی فریم ورک پیشنهادی در کسب اهداف زیر اجرا شدند:

1. خوشه بندی ساختار محور: تمایز بین کلاس‌هایی با ساختار همگن از تاپل‌های درختی XML.

2. خوشه بندی محتوا محور: شناسایی کلاس‌هایی از تاپل‌های درختی XML که یک محتوا را به اشتراک می گذارند.

3. خوشه بندی محتوا محور/ ساختار محور: این عمومی ترین هدف است و از سناریوهای متفاوت تشکیل شده است، که از تشخیص ساختار مشترک در سراسر موضوعات متفاوت، یا برعکس، برای شناسایی کلاس‌هایی از سه تاپل که هردو موضوعات مشترک را پوشش می دهند و به یک گروه ساختاری تعلق دارند تشکیل شده است – برای مثال، تاپل‌های درخت DBLP که بر "منطق محاسباتی" کار می‌کنند باید با خوشه‌های مشخصی بسته به اینکه آیا با مقالات کنفرانسی، مقالات ژورنالی، یا کتاب‌ها متناظر هستند یا نه؛ گروه بندی شوند.

از هر سه مجموعه قبلی تشریح شده، یک مجموعه تست را انتخاب کردیم. علاوه بر این، مجموعه داده ترکیبی را با اتخاذ یک سند از مجموعه داده‌های DBLP، Reuters RSS و PubMed همانندسازی کردیم. جدول 1 آمارهایی را در مورد هر مجموعه تست نشان می‌دهد، شامل اطلاعات دردسترس از فاز استخراج تاپل درخت؛ تعداد تراکنش‌ها (تاپل‌های درختی) که از اسناد اصلی در مجموعه، اندازه دامنه آیتم، تعداد عبارات شاخص بدست آمده است. عبارات

شاخص ویژگی‌های محتوا برای TCU در تاپل‌های درختی استخراج شده هستند؛ که در معرض فاز پیش پردازش شامل تحلیل لغوی، حذف stopwords، و ریشه یابی کلمات هستند.

جدول 1:آمارهای مجموعه تست (داده، اندازه، تعداد سند، تعداد ترنس، تعداد آیتم، تعداد عبارت)

data	*size*	*#docs*	*#trans.*	*#items*	*#terms*
DBLP	1444KB	3000	5140	6789	7329
Reuters RSS	3428KB	572	5653	7725	9251
PubMed	4247KB	1000	5331	6409	14380
Hybrid-Data	3229KB	1170	2893	6042	12935

جدول 2. خلاصه نتایج کلیدی خوشه بندی (نوع: content: محتوایی؛ hybrid: ترکیبی؛ structure: ساختاری)

data	*type*	f	γ	*#clust.*	*quality*
DBLP	content	0.1-0.2	0.6	65	0.97
DBLP	hybrid	0.5	0.6	83	0.955
DBLP	structure	0.7-0.8	0.7	5	1.0
Reuters RSS	content	0.1-0.2	0.55-0.65	144	0.968
Reuters RSS	hybrid	0.4-0.5	0.65-0.7	177	0.97
PubMed	content	0.1-0.2	0.65	11	0.91
PubMed	hybrid	0.5	0.7	10	0.895
Hybrid-Data	content	0.1-0.2	0.55	94	0.955
Hybrid-Data	hybrid	0.4	0.65	58	0.953
Hybrid-Data	structure	0.8-0.9	0.7	7	0.99

در ادامه بر نتایج آزمایشی اصلی بحث می‌کنیم، که در جدول 2 خلاصه شده است. خوشه بندی ساختار محور در زمانی که فاکتور f برابر یا بالاتر از 0.7 است حاصل می‌شود، در حالی که خوشه بندی محتوا محور نیازمند f برابر با 0.3 یا بالای 0.3 است. مقدار میانه f به رفتار خوشه بندی محتوا محور/ساختار محور فریم ورک منجر می‌شود. برای اختصار،جدول 2 بیشترین نتایج را برای هر تست نشان می‌دهد. برای نمونه، اولین ردیف جدول بهترین نتایج ارائه شده توسط میانگین XTrK را در زمانی که خوشه بندی محتوا محور بر DBLP اجرا می‌شود ارائه می‌دهد: این نتایج به تنظیمات خاص $(f \in [0.1..0.2],\ \gamma = 0.6)$ اشاره دارد؛ در واقع؛ نتایج کیفی بسیار مشابه و تعداد قابل مقایسه‌ای از خوشه‌های محاسبه شده بدست می‌اید، برای مثال مقدار γ نزدیک به 0.6 است.

در مجموعه تست DBLP؛ فریم ورک به صورت مناسبی پنج کلاس ساختاری را با پوشش اسنادی در مجموعه‌هایی شامل مقالات ژورنالی، مجموعه مقالات کنفرانس، مقاله‌ای در مجموعه مقالات، کتاب ها، و فصل‌های کتاب را سازمان داده است. تست نوع محتوا یک سناریوی متنوع تر را ارائه می‌دهد، که شصت و پنج موضوع برای اتخاذ خوشه‌هایی با اندازه متفاوت در نظر گرفته شده‌اند. با نگاه به نمایشی از خوشه‌های محاسبه شده، دریافتیم که عباراتی که در محتوای عناصر ظاهر می‌شوند، مانند عنوان، عنوان کتاب، مجله، سری، و نویسنده در ارتقای موضوعات مرتبط به صورت ویژه‌ای باعث تبعیض می‌شوند. تگ‌های خوشه به موضوعات کلی مرتبط هستند، مانند "برنامه نویسی منطقی"، "پایگاه داده وب"، " مدل‌های تراکنش پایگاه داده"، یا " استدلال غیریکنواخت" ؛ اما به موضوعات خاصتری مانند " همزمانی در سیستم‌های نرم افزاری"، " زبان و ماشین درختی"، "الگوریتم‌های تکاملی"، " مدل سازی شی گرا پایگاه داده چند رسانه ای"، بیشتر مرتبط هستند. حتی نتایج جالب تری با تست‌های خوشه بندی "ترکیبی" ارائه می‌شود. در برخی از موارد، تراکنش‌هایی که ویژگی‌های محتوایی را به اشتراک می گذارند در خوشه‌های مجزایی گروه بندی می‌شوند که گروه‌های ساختاری متفاوت را منعکس می‌کند. برای مثال، مقالات کنفرانسی و فصل‌های کتاب بر " سیستم‌های پایگاه داده چند رسانه ای" در دو خوشه جدا گروه بندی می‌شود. در موارد دیگر، اثرات بخش بندی با توجه به ویژگی محتوا صورت می‌گیرد: مقالات کنفرانسی بر " واسط‌ها و سیستم‌های تطبیقی" بسته به ویژگی‌های موضوعات بر زیر موضوعات "واسط‌ها و سیستم های تطبیقی" در خوشه‌های متفاوتی گروه بندی می‌شوند.

تست‌ها بر Reuters RSS نتایج مرتبطی را اساسا از نقطه نظر محتوا ارائه دادند. برای این مجموعه داده، خوشه بندی ساختاری خالص خیلی حس نمی‌شود، چرا که ساختار اسناد یک الگوی از پیش تعیین شده را برای یک مقاله خبری عمومی طراحی می‌کند. برخلاف DBLP، هیچ تگی وجود ندارد که گروه‌های ساختاری را نشانه گذاری کند، اگر چه این موارد به طور ضمنی به عنوان عناوینی از کانال‌های RSS مشخص می‌شود. اسناد در مجموعه تست انتخاب شده با یکی از ده کانال RSS مرتبط هستند (برای مثال "جهان"، " کسب و کار"، " علوم"، "ورزش"، "سیاست"، "تکنولوژی"). این کانال‌ها می‌توانند به عنوان یک زمینه گسترده در نظر گرفته شوند؛ با این حال، برخی از کلاس‌های مرتبط و متناظر ممکن است همپوشانی داشته باشند: برای نمونه، "خبرهای داخلی" و "خبرهای جهانی" ممکن است

در کانال "خبرهای برتر" گنجانیده شوند، و به همین ترتیب کانال‌های " خبرهای علمی" و "خبرهای سلامتی" ممکن است مقالات جدیدی را به اشتراک گذارند. خوشه‌های محاسبه شده سرانجام بر موضوعاتی از محدوده موضوعات مورد نظر از "کمک‌های سونامی" تا " ناسا و شاتل فضایی" اخیر، از "خدمات جدید برای تلفن‌های همراه" تا "رخداد Live8"، از "مسابقات تنیس 2005 ویمبلدون" تا " جاه طلبی‌های هسته‌ای کره شمالی" هستند.

PubMed، ما بر خوشه بندی محتوا محور تمرکز می‌کنیم. تگ‌های خوشه برطبق محتوای AbstractText، ArticleTitle و دیگر عناصر مرتبط با اطلاعات مجلات پزشکی تولید شدند. خوشه‌های محاسبه شده موضوعات مرتبط را پوشش می‌دهند، برای مثال، "بررسی پروتئوم" در " تومور نوروبلاستوما" یا در "انتخاب کروماتوگرافی"، "سرطان مولکولی" و چندین رشته فرعی "فیزیولوژی" (برای مثال غدد ترشح داخلی، "متابولیسم"، "قلب و فیزیولوژی گردش خون"، "فیزیولوژی سلولی و ملکولی ریه"). با توجه به پیچیدگی موضوعات PubMed، فریم ورک خوشه بندی هنوز اثربخشی خوبی را نشان می‌دهد، که در جدول 2 گزارش شده است.

عدم تجانس مجموعه Hybrid-Data به ما اجرای تست‌های ساختار محور را اعلام می‌کند. به خصوص هفت کلاس را انتظار داریم، که پنج تای آن‌ها DBLP، یک کلاس PubMed، و یک کلاس .Reuters RSS است. اغلب نتایج کیفیتی خوب توسط میانگین- XTrK ارائه می‌شود. تست‌های ترکیبی و محتوا محور دقت بالایی را نشان می‌دهند، که از گرایش عمده نشان داده شده در تست‌هایی بر مجموعه تست‌های متمایز پیروی می‌کنند.

علاوه بر f؛ نوع خوشه بندی در آستانه γ کاربردی است. براستی، مقادیر بالا γ (مثال 0.7) برای اجرای موثر خوشه بندی ساختاری کافی است، به خصوص زمانی که تمایز بین گروه‌های ساختاری هموار باشد، در مقابل، انعطاف پذیری بیشتر گروه (برای مثال مقدار کمتر γ) برای اتخاذ چندین جنبه معنایی از یک مجموعه داده مناسب است.

یک مشاهده نهایی ارزشمند در مورد سر و کار داشتن با دیگر سنجش‌های میانگین-XTrK است. خوشه‌های سطل زباله در هر تست تولید شدند، و افزایش در اندازه (تعداد خوشه ها) بخش‌های نتیجه شده از 10٪ (در PubMed و Reuters RSS) تا 25٪ (در DBLP و Hybrid-Data) را یافتیم. خوشه بندی سطل زباله همیشه کیفیت

بخش‌های کلی را بهبود می‌بخشد، اگر چه مزایای بدست آمده ممکن است کوچک باشند (2٪ در DBLP، 1٪ ددر Hybrid-Data) یا کاملاً مرتبط باشند (زیر 1٪ در PubMed و Reuters RSS).

6. نتیجه گیری و کارهای آینده

فریم ورک خوشه بندی جدیدی را برای سازمان معنایی داده XML ارائه دادیم. ویژگی‌هایی را برای نمایش مناسب اطلاعات ساختاری و محتوایی از اسناد XML بررسی کردیم. ویژگی‌های پشتیبانی از پایگاه دانش لغوی را غنی ساختیم، که نقش اصلی را در استنتاج معانی XML بازی می‌کند. از مفاهیم تاپل درخت برای استخراج ساختار معنایی منسجم از اسناد XML استفاده می‌کنیم، نشان دادیم که تاپل‌های درخت XML به آسانی به عنوان تراکنش مدل می‌شوند. یک رویکرد خوشه بندی تفکیکی توسعه یافته است و به دامنه تراکنشی XML اعمال شده است. ارزیابی خوشه بندی اثربخشی بسیار بالایی را برای مجموعه داده واقعی بزرگ نشان می‌دهد، که استدلال می‌کند که آیتم‌های تاپل درختی XML ویژگی‌های قدرتمندی برای خوشه بندی XML معنایی موثر به همراه دارد.

جهات آشکاری برای تحقیقات آینده وجود دارد. برخی از این موارد در راستای تحکیم جنبه‌های خاصی از چارچوب است، مانند توسعه الگوریتم خوشه بندی جدید که قادر است بهترین تناسب را با مدل تراکنش XML داشته باشد و در عین حال حداقل نیازمندی‌های مقیاس پذیری، کشف خوشه در زیر فضا، و تگ گذاری آگاهانه خوشه را برآورده سازد. علاوه بر این، نقش دانش هستی شناسی در پشتیبانی از تشخیص روابط در میان داده XML نیازمند بررسی بهتر است. بنابراین، باید ترکیب هستی شناسی کاربردی در فریم ورک خوشه بندی برای بهره بردن از مدل‌های مفهومی بسط یافته که نه تنها اشیای XML، روابط و محدودیت‌های آن را نشان می‌دهد؛ بلکه "نقش ها"یی را با توجه به چگونگی ظاهر شدن اشیا در یک منبع XML نشان می‌دهد.

References

[1] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: from relations to semistructured data and XML*. Morgan Kaufmann Publishers, 1999.

[2] M. Arenas and L. Libkin. A Normal Form for XML Documents. *ACM Trans. Database Systems*, 29(1):195–232, 2004.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press Books. Addison-Wesley, 1999.

[4] S. Banerjee and T. Pedersen. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proc. IJCAI*, pages 805–810, 2003.

[5] G. Costa, G. Manco, R. Ortale, and A. Tagarelli. A Tree-based Approach to Clustering XML Documents by Structure. In *Proc. PKDD*, pages 137–148, 2004.

[6] A. Doucet and H. A. Myka. Naive Clustering of a Large XML Document Collection. In *Proc. INEX Annual ERCIM Workshop*, pages 81–88, 2002.

[7] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[8] S. Flesca, F. Furfaro, S. Greco, and E. Zumpano. Repairs and Consistent Answers for XML Data with Functional Dependencies. In *Proc. Int. XML Database Symposium (XSym)*, pages 238–253, 2003.

[9] F. Giannotti, C. Gozzi, and G. Manco. Clustering Transactional Data. In *Proc. PKDD*, pages 175–187, 2002.

[10] M. Lesk. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from a ice cream cone. In *Proc. ACM SIGDOC Int. Conf. on Systems Documentation*, pages 24–26, 1986.

[11] W. Lian, D. W. Cheung, N. Mamoulis, and S.-M. Yiu. An Efficient and Scalable Algorithm for Clustering XML Documents by Structure. *IEEE Trans. Knowledge Data Engineering*, 16(1):82–96, 2004.

[12] A. Nierman and H. V. Jagadish. Evaluating Structural Similarity in XML Documents. In *Proc. ACM SIGMOD WebDB Workshop*, pages 61–66, 2002.

[13] N. Polyzotis and M. Garofalakis. Structure and Value Synopses for XML Data Graphs. In *Proc. VLDB*, pages 466–477, 2002.

[14] P. Resnik. Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[15] A. R. Schmidt, F. Waas, M. L. Kersten, D. Florescu, I. Manolescu, M. J. Carey, and R. Busse. The XML Benchmark Project. Technical report, INS-R0103, CWI, Amsterdam, The Netherlands, 2001.

[16] M. Theobald, R. Schenkel, and G. Weikum. Exploiting Structure, Annotation, and Ontological Knowledge for Automatic Classification of XML Data. In *Proc. ACM SIGMOD WebDB Workshop*, pages 1–6, 2003.

[17] J. Widom. Data Management for XML: Research Directions. *IEEE Data Engineering Bulletin*, 22(3):44–52, 1999.

[18] J. P. Yoon, V. Raghavan, V. Chakilam, and L. Kerschberg. BitCube: A Three-Dimensional Bitmap Indexing for XML Documents. *Journal of Intelligent Information Systems*, 17(1):241–252, 2001.

[19] M. J. Zaki and C. C. Aggarwal. XRules: An Effective Structural Classifier for XML Data. In *Proc. ACM SIGKDD*, pages 316–325, 2003.