دانشکده مهندسی برق و کامپیوتر
بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات

رساله دکتری
در مهندسی کامپیوتر
(سیستمهای نرم‌افزاری)

کاوش الگوهای تکراری در جریان های داده بر اساس مدل پنجره لغزنده

به کوشش:
محمود دی پیر

استاد راهنما:
دکتر محمد هادی صدرالدینی

آذر ۱۳۹۰

بسم الله الرحمن الرحيم

به نام خدا

اظهارنامه

اینجانب محمود دی پیر (۸۵۳۹۳۴) دانشجوی رشته‌ی مهندسی کامپیوتر گرایش سیستمهای نرم‌افزاری دانشکده‌ی مهندسی برق و کامپیوتر اظهار می‌کنم که این پایان نامه حاصل پژوهش خودم بوده و در جاهایی که از منابع دیگران استفاده کرده‌ام، نشانی دقیق و مشخصات کامل آن را نوشته‌ام. همچنین اظهار می‌کنم که تحقیق و موضوع پایان نامه‌ام تکراری نیست و تعهد می‌نمایم که بدون مجوز دانشگاه دستاوردهای آن را منتشر ننموده و یا در اختیار غیر قرار ندهم. کلیه حقوق این اثر مطابق با آیین‌نامه مالکیت فکری و معنوی متعلق به دانشگاه شیراز است.

نام و نام خانوادگی: محمود دی پیر

تاریخ و امضا:

۹۰/۱۱/۳۰

به نام خدا

کاوش الگوهای تکراری در جریان های داده بر اساس مدل پنجره لغزنده

به کوشش:

محمود دی پیر

پایان نامه
ارائه شده به تحصیلات تکمیلی دانشگاه شیراز به عنوان بخشی
از فعالیتهای تحصیلی لازم برای اخذ درجه دکتری

در رشته‌ی:

مهندسی کامپیوتر(سیستمهای نرم‌افزاری)

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی و تصویب شده توسط کمیته پایان نامه با درجه: عالی

دکتر محمدهادی صدرالدینی، دانشیار بخش مهندسی و علوم کامپیوتر و فنآوری اطلاعات (رئیس کمیته) .....

دکتر غلامحسین دستغیبی فرد، استادیار بخش مهندسی و علوم کامپیوتر و فنآوری اطلاعات.....

دکتر ستار هاشمی، استادیار بخش مهندسی و علوم کامپیوتر و فنآوری اطلاعات.....

دکتر طاهر نیکنام، دانشیار بخش مهندسی برق دانشگاه صنعتی شیراز.....

دکتر شاپور گلبهار حقیقی، استادیار بخش مهندسی برق دانشگاه شیراز.....

آذر ۱۳۹۰

سپاسگزاری:

اکنون که این پایان‌نامه به پایان رسیده است از زحمات بی دریغ استاد راهنما جناب آقای دکتر صدرالدینی تشکر و قدردانی می‌نمایم. انجام این پایان نامه بدون راهنمایی ایشان امکانپذیر نبود. همچنین از راهنمایی های اساتید مشاور جناب آقای دکتر دستغیبی فرد و جناب آقای دکتر هاشمی سپاسگزارم. از همسر و فرزندم که در این مدت مرا همراهی نموده و کوتاهی های مرا تحمل کردند، نیز تشکر و قدردانی می‌نمایم.

تقدیم به

همسر مهربانم

چکیده

کاوش الگوهای تکراری در جریان های داده بر اساس مدل پنجره لغزنده

به کوشش

محمود دی پیر

در مسئله کاوش الگوهای تکراری به دنبال مجموعه هایی هستیم که در تعداد قابل توجهی از تراکنش ها (رکوردها) دیده می شوند. کاوش الگوهای تکراری در جریان های داده کاربرد زیادی در صنعت، تجارت و علوم مختلف دارد. به دلیل سرعت زیاد، بی پایان بودن، حجم بالا و ماهیت تغییر کننده محتوای جریان های داده، یافتن الگوهای تکراری در این نوع داده ها مسئله ای چالش برانگیز است. مدل پنجره لغزنده یکی از مدل های محبوب و پرکاربرد برای حل این مسئله است، که در آن میزان ثابتی از تراکنش های جدید برای کاوش در نظر گرفته می شوند. کاهش میزان حافظه مصرفی، افزایش سرعت کاوش و تعیین اندازه پنجره مهمترین چالش های این مدل اند. ما در این رساله برای غلبه بر این چالش ها الگوریتم ها و تکنیک هایی ارائه داده ایم. نخست دو الگوریتم پیشنهاد داده ایم که در آنها محتویات پنجره به صورت پویایی نگهداری می شود و در صورت درخواست کاربر، مجموعه اقلام تکراری کاوش می شود. سپس الگوریتمی تقریبی ارائه داده ایم که همواره مجموعه اقلام تکراری را نگهداری و به روز رسانی می کند. علاوه بر این الگوریتم جدیدی معرفی کرده ایم که قادر است با استفاده از پنجره های زمانی، سرعت متغییر جریان داده ورودی را در نظر بگیرد. آزمایش های ما نشان می دهند که همگی این الگوریتم ها نسبت به نمونه های مشابه خود در بیشتر موارد، از نظر حافظه و سرعت بهتر عمل می کنند. ما همچنین الگوریتمی ارائه داده ایم که در آن اندازه پنجره لغزنده بر اساس میزان تغییر مفهوم در جریان داده ورودی تنظیم می شود. در نهایت معیار جدیدی را برای تشخیص بهتر میزان تغییر مفهوم ارائه داده ایم.

فهرست

فهرست شکل ها

فهرست جداول

فصل اول

مقدمه

۱-۱ مقدمه

مسئله کاوش الگوهای تکراری[۱] در پایگاه های داده در سال ۱۹۹۳ توسط آگروال و دیگران ارائه شد[۱]. این مسئله به علت کاربرد وسیع در تجارت، صنعت و علوم مختلف، موضوعی مهم و اساسی در زمینه کشف دانش و داده کاوی است. برای حل این مسئله الگوریتم پایه ای و شناخته شده ای به نام *ای.پریوری* (Apriori) وجود دارد[۲]. این الگوریتم، معروف ترین الگوریتم در این زمینه است که در نرم افزارهای داده کاوی تجاری نیز استفاده شده است. در سال های اخیر تحقیقات بسیار زیادی در مورد این مسئله انجام گرفته و پیشرفت های چشمگیری صورت گرفته است که حاصل آنها ارائه الگوریتم های کارا و مقیاس پذیر از یک طرف و مطرح شدن آن در کاربرد های مختلف بوده است. تحقیقات انجام شده عموماً در مورد کاوش الگوهای تکراری در پایگاه داده های ایستا بوده است به طوری که الگوریتم های مختلفی در این مورد ارائه شده اند [۲،۳،۴،۵،۶،۷]. اگرچه تعداد الگوریتم های ارائه شده بسیار زیاد است اما اکثر آنها غیر از ای. پریوری [۲]، اف. پی. گروت [۳] واکلات [۷] چندان معروف نبوده و کاربرد آنها فراگیر نشده است. در پایگاه داده های ایستا فرض بر این است که داده ها در طول عملیات کاوش تغییر نمی کنند و الگوریتم داده کاوی می تواند چندین بار داده های ورودی مشخص شده ای را بخواند. به عبارت دیگر کل پایگاه داده در زمان کاوش در اختیار است. یک الگو در یک پایگاه داده تکراری است اگر تعداد رخداد آن از یک حد آستانه بیشتر باشد. این حد آستانه توسط کاربر عملیات داده کاوی تعیین می شود. تحلیل سبد خرید معروف ترین و شناخته شده ترین کاربرد الگوهای تکراری است. در این مسئله کالاهایی که توسط مشتریان مختلف یک فروشگاه به صورت تکراری با هم خریداری می شوند، در پایگاه داده آن فروشگاه

[۱] Frequent pattern mining

کاوش می شوند. نتایج این نوع کاوش بسیار جالب است، به نحوی که گاهی باعث تعجب کاربران شده است.

کاوش الگوهای تکراری علاوه بر داده های ایستا، اخیراً در پایگاه داده های پویا و جریان های داده[۲] نیز مطرح شده است[۸]. در داده های جریانی فرض بر این است که داده ها به صورت پشت سرهم و بی وقفه به سیستم می رسند. در یک جریان داده، داده ها با سرعت و به شکلی بی پایان دریافت می شوند. برنامه های کاربردی و سیستم های زیادی وجود دارند که جریان داده تولید می کنند. از جمله آنها می توان برنامه های پایش[۳] کارایی شبکه، رکورد های جزئیات تماس در مخابرات، کالاهای فروخته شده در فروشگاه های زنجیره ای، سیستم اطلاعات هوا شناسی، پیام های پست الکترونیک، رکورد های ثبت در وب سرورها، سیستم دنبال کردن جریان کلیک در وب سایت ها و غیره را نام برد. انجام محاسبات مختلف آماری و داده کاوی، از جمله مسائل مطرح در مدل جریانی می باشند. در جریان های داده، در هر لحظه، حجم داده هایی که تاکنون رسیده و در آینده نیز خواهد رسید، بسیار زیاد است. بنابراین ذخیره آن امکانپذیر نیست. تنها یک خلاصه کوچک از آن را می توان محاسبه و نگهداری کرد و از مابقی داده ها باید صرف نظر کرد، و یا اینکه فقط داده های اخیراً رسیده و نتایج به روز شده را می توان در حافظه نگهداشت. حتی با فرض اینکه بتوان همه جریان داده را ذخیره کرد، بررسی و خواندن مجدد آن برای پردازش بیشتر امکانپذیر نیست. سرعت رسیدن داده ها به شکلی است که هر عنصر داده ای باید در همان زمان به صورت بلادرنگ پردازش و سپس از آن صرف نظر شود. بنابراین زمان و حافظه دو فاکتور مهم در کاوش جریان های داده هستند. فاکتور مهم دیگر چگونگی برخورد با مسئله تغییر مفهوم است. خصوصیات آماری مفهومی که می خواهیم پیش بینی و یا توصیف کنیم با گذشت زمان در یک جریان داده تغییر می کند به طوری که مدلی که با استفاده از داده های قبلی ساخته شده است، برای داده های جدید معتبر نبوده و یا با خطا همراه خواهد بود. در کاوش الگوهای تکراری مدل مورد نظر مجموعه اقلام تکراری هستند. تغییر مفهوم به صورت تبدیل الگوهای تکراری به غیر تکراری و برعکس با گذشت زمان نمود پیدا می کند. این تغییرات در طول زمان ممکن است چندین بار

---

[۲] Data streams
[۳] Monitor

اتفاق بیفتد که خود مسئله جریان کاوی الگوهای تکراری را مشکل تر می کند. به عبارت دیگر مجموعه ی کل الگوهای تکراری و مقدار تکرار آنها با گذشت زمان تغییر می کند.

تا کنون سه مدل برای کاوش الگوهای تکراری در جریان های داده ارائه شده اند. این سه مدل، عبارت از مدل نشانه[۴]، مدل پنجره لغزنده[۵] و مدل زوال[۶] (یا مدل محو شدن) هستند. در مدل نشانه، داده های رسیده از یک نقطه عطف زمانی تا کنون مورد نظر هستند. این نقطه عطف زمانی می تواند زمان شروع به کار سیستم باشد. در مدل پنجره لغزنده، مقدار مشخص شده ای از داده های اخیراً دریافت شده به منظور کاوش مدنظر قرار می گیرند. این مقدار مشخص شده می تواند عددی و یا زمانی باشد. مورد اول را پنجره تراکنشی و مورد دوم را پنجره زمانی گویند. به عبارت دیگر پنجره W روی یک جریان داده به معنی این است که |W| تراکنش اخیر و یا همه تراکنش های مربوط به |W| واحد زمانی اخیر مورد توجه هستند، که در اینجا |W| اندازه پنجره است. واضح است که تعداد تراکنش های پنجره در حالت اول ثابت و در حالت دوم مطابق سرعت دریافت داده ها می تواند متغییر باشد. در هر دو حالت، داده های جدید از یک طرف وارد پنجره شده و از طرف دیگر داده های قدیمی خارج می شوند، که به این عملیات لغزش پنجره به جلو گفته می شود. در مدل سوم یعنی مدل زوال از یک تابع زوال به منظور وزن دادن به عناصر داده ای بر اساس ترتیب رسیدن آنها از جریان داده، استفاده می شود به نحوی که داده های اخیر وزن بیشتری نسبت به داده های قدیمی تر می گیرند. در این مدل از داده هایی که وزن شان صفر می شود صرف نظر می شود. ما در این پایان نامه مسئله جریان کاوی الگوهای تکراری را در مدل پنجره لغزنده مورد توجه قرار داده ایم چون در این مدل به داده های اخیر اهمیت بیشتری داده می شود و علاوه براین نیاز به حافظه محدودی دارد. با مطالعه و مقایسه الگوریتم ها و همچنین تحقیقات قبلی انجام شده، سعی در ارائه روش های جدید و بهینه تر در این زمینه داریم.

۱-۲ تعریف مسئله

---

[۴] Landmark
[۵] Sliding Window
[۶] Decay

فرض کنید $I=\{i_1,i_2,\ldots,i_m\}$ مجموعه ای از اقلام باشد. همچنین فرض کنید DS جریانی از تراکنش ها به صورت پشت سرهم باشد، به طوریکه هر تراکنش زیر مجموعه ای از I است. برای هر مجموعه قلم (الگو) X که زیر مجموعه ای از I است، گفته می شود که تراکنش T در DS شامل الگوی X است اگر $X \subseteq T$ باشد. کسری از تراکنش ها در DS که شامل X باشند، به عنوان پشتیبانی X شناخته می شود. پشتیبانی مطلق یا تعداد تکرار X، تعداد تراکنش هایی در DS است که شامل این مجموعه قلم باشند. با داشتن یک حداقل حد آستانه پشتیبانی S، گوئیم مجموعه قلم X، تکراری است اگر حداقل S% از تراکنش های DS شامل X باشند. پنجره لغزنده روی جریان داده DS شامل |W| تراکنش اخیر در جریان داده DS است، که |W| در اینجا طول پنجره است. با گذشت زمان، پنجره با اضافه شدن تراکنش جدید و حذف تراکنش قدیمی، به طرف جلو حرکت می کند. جهت بدست آوردن کارائی بهتر می توان تراکنش ها را به صورت دسته ای اضافه و حذف نمود. یک پنجره روی جریان داده DS به صورت:

$$W_{n-|w|+1} = \{T_{n-|w|+1}, T_{n-|w|+2}, \ldots, T_n\} \qquad (۱-۱)$$

تعریف می شود که در آن $n-|w|+1$ و T_i به ترتیب شماره پنجره و i امین تراکنش در جریان داده DS هستند. در حقیقت پنجره شامل تراکنش های اخیراً رسیده جریان داده DS می باشد. یک مجموعه قلم به نام X در این پنجره تکراری است اگر رابطه $Sup(X) \geq |W|.S$ برقرار باشد که در این رابطه Sup(X) به معنی تعداد تکرار X در این پنجره است. بنابراین با داشتن پنجره W و حد آستانه پشتیبانی S تعیین شده بوسیله ی کاربر، مسئله یافتن همه ی الگوهای تکراری در این پنجره با استفاده از حداقل زمان و حافظه است. به عنوان مثال شکل ۱-۱، یک جریان داده از تراکنش ها را نشان می دهد. ستون سمت چپ شماره تراکنش ها و ستون سمت راست اقلام موجود در هر تراکنش است. دو پنجره W_1 و W_2 بر روی تراکنش هایی که تا کنون رسیده اند، در نظر گرفته شده اند.