دانشکده مهندسی

پایان نامه‌ی کارشناسی ارشد در رشته‌ی مهندسی کامپیوتر(نرم‌افزار)

کاوش الگوهای تکرار شونده در جریانهای داده

به کوشش

مینا معمار

استاد راهنما

دکتر محمدهادی صدرالدینی

شهریور ۱۳۹۰


بسم الله الرحمن الرحيم

به نام خدا

کاوش دسته‌داده‌های تکرارشونده در جریانهای داده

به کوشش:

مینا معمار

پایان نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی

از فعالیت‌های تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته:

مهندسی کامپیوتر- نرم افزار

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان نامه با درجه: عالی

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

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

دکتر اقبال منصوری، استادیار بخش مهندسی و علوم کامپیوتر.....

شهریورماه ۱۳۹۰

به نام خدا

اظهارنامه

اینجانب مینا معمار (۸۷۰۹۳۰) دانشجوی رشته‌ی مهندسی کامپیوتر گرایش نرم‌افزار دانشکده‌ی مهندسی اظهار می‌کنم که این پایان‌نامه 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 پایان‌نامه‌ام تکراری نیست و تعهد می نمایم 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: مینا معمار

تاریخ و امضاء: ۱۳۹۰/۰۶/۳۰

تقدیم به آیه‌های عشق و ایثار

پدر و مادر مهربانم

سپاسگزاری

پیش از هر چیز از خداوند متعال به خاطر تمامی نعمت‌هایی که هدیه کرده سپاسگزارم.
از استاد ارجمند، جناب آقای دکتر محمد هادی صدرالدینی که در طول انجام این پروژه همواره مشوق و راهنمای من بوده‌اند، صمیمانه تشکر و قدردانی می‌نمایم. همچنین از اعضای خانواده‌ام که در تمام دوران تحصیل با روی گشاده مرا یاری کرده‌اند نیز سپاسگزارم.
این پایان‌نامه تحت قرارداد همکاری پژوهشی به شماره ۵۰۰/۴۷۰۹/ت مورخ ۱۳۸۹/۰۴/۰۷ از حمایت‌های مالی و معنوی مرکز تحقیقات مخابرات ایران بهره‌مند شده است.در اینجا لازم می‌دانم از حمایت این مرکز تشکر و قدردانی نمایم.

چکیده

روشی سریع برای یافتن دسته‌داده‌های تکراری در جریان‌های داده مبتنی بر مدل پنجره لغزان بیتی

به کوشش

مینا معمار

یافتن دسته‌داده‌های تکراری در جریان پیوسته تراکنشها یکی از مسائل حیاتی در کاربردهایی نظیر تحلیل سبد خرید، مانیتورینگ شبکه و پیش‌بینی فروش انبار می‌باشد. پنجره لغزان یکی از مدلهایی است که به دلیل مدیریت تغییر محتوا، مصرف حافظه محدود و سرعت پردازش کم به شکل گسترده در یافتن دسته‌داده‌های تکراری در جریانهای داده استفاده شده است. یک الگوریتم مبتنی بر مدل پنجره‌ای لغزان به یک ساختمان داده کارا احتیاج دارد که به سریعترین شکل ممکن با حذف و درج تراکنشها خود را به روز کند. به علاوه یک روش محاسباتی ابداعی نیز لازم است تا با تاخیر اندکی بعد از درخواست کاربر جهت مشاهده نتایج کاوش در یک پنجره، مجموعه دسته‌داده‌های تکراری را تولید کند. در این پایان‌نامه یک ساختمان داده کارا با نام دنباله بیتی بلوک‌بندی شده برای ذخیره و نگهداری محتویات پنجره معرفی شده است. به علاوه با یک تکنیک جدید این پنجره مورد پویش قرار گرفته تا مجموعه دسته‌داده‌های تکراری در پنجره جاری به شکلی کارا استخراج شوند. نتایج آزمایشات روی جریان داده‌های واقعی و مصنوعی نشان می‌دهد که این الگوریتم سریعتر از الگوریتمهایی است که اخیرا در زمینه کاوش جریانهای داده ارائه شده‌اند.

فهرست مطالب

فهرست جدول‌ها

فهرست تصویرها

فهرست نمودارها

فصل اول

مقدمه

۱-۱- داده‌کاوی

در سال ۱۹۸۲ John Naisbitt در کتاب خود Megatrends نوشت : « ما در اطلاعات غرق شده‌ایم ولی در فقر دانش به سر می‌بریم.» در دهه‌های اخیر، حجم داده‌های موجود هر ساله دو برابر می‌شود ولی دانشی که ما از این داده‌ها به دست می‌آوریم به همان سرعت نیست. زمینه داده‌کاوی تا حدی برای رفع این مشکل پدیدار شده است و تکنیکهای داده‌کاوی نشان داده است که در بسیاری از دامنه‌های کاربر در دنیای واقعی می‌توانند به شدت مفید باشند.

داده کاوی عبارت است از استخراج الگوهای جذاب و مخفی از حجم بزرگی از داده‌ها که در دیتابیسها، انباره‌های داده، و یا سایر محلهای ذخیره‌سازی اطلاعات نگهداری می‌شوند. به عبارت دیگر داده‌کاوی استخراج دانش از پایگاه داده‌ها می‌باشد. داده‌کاوی تجمیعی از تکنیکهای مختلف نظیر تکنولوژیهای بانک اطلاعاتی، آماری، یادگیری ماشین، شبکه‌های عصبی ، استخراج اطلاعات و سایر علوم مشابه می‌باشد.

طبق آنچه در [۳۷] آمده است : داده‌کاوی فرآیند کشف الگوها وارتباطات معناداری است که در دیتابیسهای بسیار بزرگ پنهان شده‌اند. همچنین در [۳۸] تعریفی که برای داده‌کاوی ارائه شده است عبارت است از: آنالیز شهودی بانکهای اطلاعاتی برای کشف ارتباطات غیرقابل انتظار و خلاصه سازی داده‌ها با یک شیوه جدید که برای صاحبان آن داده قابل درک و مفید باشد.

داده‌کاوی بخشی از فرآیندی با نام کشف دانش در پایگاه داده می‌باشد. این فرآیند شامل چندین مرحله می‌باشد که همگی قبل از اجرای مرحله داده‌کاوی اجرا می‌شوند، از قبیل انتخاب داده‌ها، پاکسازی داده‌ها، پیش‌پردازش و تبدیل داده‌ها[۳۹].

داده‌کاوی فرآیند جمع‌آوری داده، آنالیز و پیش‌بینی است. ابزارهای داده‌کاوی از متدهای تحلیلی سطح بالایی برای کشف ارتباطات و الگوهای ناشناخته در دیتاستهای بزرگ استفاده می‌کنند و رفتار و سمت و سوی آنها در آینده را پیش‌بینی می‌کنند. نتایج داه‌کاوی به کاربران امکان می‌دهد تا تصمیمات مبتنی بر دانشی اتخاذ کنند. فرآیند داده‌کاوی شامل سه مرحله زیر می‌باشد:

- **مرحله ۱** : اکتشاف. این مرحله معمولا با آماده سازی داده‌ها شروع می‌شود که شامل انتخاب داده‌ها، پاکسازی داده‌ها، تبدیل داده‌ها و انجام یک سری آنالیز مقدماتی روی داده‌ها می‌باشد. آنالیز مقدماتی روی داده‌ها با هدف شناسایی متغیرهایی که بیشترین ارتباط با یکدیگر را دارند انجام می‌شود تا بر اساس آن بتوان نوع و پیچیدگی تکنیک داده‌کاوی که در مرحله بعد انجام می‌شود را تعیین کرد.

- **مرحله ۲:** ساخت مدل. این مرحله شامل اعمال تکنیکهای مختلف روی داده‌های تست و انتخاب کاراترین تکنیک برای رسیدن به نتیجه مورد نظر می‌باشد.

- **مرحله ۳:** توسعه. در این مرحله تکنیک انتخاب شده روی داده‌های جدید اعمال می‌شود تا نتایج مورد نظر استخراج شوند.

بسته به دامنه کاربرد و نیازهای کاربر، داده‌کاوی می‌تواند کارکردها و وظایف مختلفی را ارائه دهد. شاخه‌های داده‌کاوی کاملا مختلف و متمایز از یکدیگر هستند چون دیتاستهای مختلف شامل الگوهای مختلفی هستند. در میان شاخه‌های داده‌کاوی، شاخه‌های زیر به شکل گسترده‌تری در کاربردهای واقعی به کار گرفته شده‌اند:

- **خوشه‌بندی**[1]: این تکنیک به دنبال مجموعه‌ای از گروه‌بندی‌ها برای توصیف داده‌ها می‌گردد. خوشه‌ها می‌توانند کاملا از هم مجزا باشند و یا شامل یک ساختار پیشرفته‌تر نظیر ساختار سلسله‌مراتبی جهت نمایش داده‌ها باشند.

[1] Clustering

- **طبقه‌بندی**[1]: هدف این دسته پیدا کردن تابعی است که هر یک از داده‌های موجود در دیتاست را به یکی از چند کلاس از پیش تعریف شده نگاشت کند.

- **شمارش تکرار**[2] **و کاوش قوانین انجمنی**[3]: این دسته به دنبال ارتباطات میان متغیرها یا شناسایی پراهمیت ترین مقادیر در دیتاست می‌گردند.

- **خلاصه سازی و رگرسیون:** این دسته شامل متدهایی برای یافتن یک توصیف مختصر برای کل داده‌های دیتاست یا زیرمجموعه‌ای از آن می‌باشد. مثال ساده برای این دسته محاسبه میانه و انحراف معیار برای فیلدها می‌باشد. مثال دیگر نیز یافتن تابعی است که داده‌ها را با کمترین خطا مدل کند.

داده‌کاوی یکی از زمینه‌های تحقیقاتی است که با توجه به توسعه سخت‌افزار و نرم‌افزار از اهمیت بالایی برخوردار شده است. داده، در واقع دارایی یک سازمان به شمار می‌آید، و مبرهن است که استفاده از این داده برای پیش‌بینی آینده می‌تواند بسیار جذاب باشد. داده‌کاوی روشی برای کمک به سازمانهاست که بتوانند از داده‌های خود نهایت استفاده را ببرند و در جاهایی که نیاز به تصمیم‌گیری دارند بتوانند از نتایج حاصل از داده‌کاوی بهره بگیرند. از آنجایی که سازمانها به صورت مداوم در حال رشد هستند، دیتابیسهای آنها نیز به همان نسبت در حال رشد هستند. در نتیجه تکنیکهای داده‌کاوی آنها به مرور زمان قادر به پوشش دیتابیسهای در حال رشد آنها نیستند و با شکست روبرو می‌شوند. بزرگ و پویا بودن طبیعت دیتابیسها است و تغییر در آنها به طور معمول اتفاق می‌افتد، داده‌های جدید به دیتابیس اضافه می‌شود و داده‌های قبلی حذف یا اصلاح می‌شوند. با توجه به این مسئله در داده‌کاوی توجه به این نکته که نتایج باید به روز باشند و با حداکثر داده‌های جاری سازگاری داشته باشند، ضرورت دارد. به عبارت دیگر سیستمهای داده‌کاوی بایستی با توجه به اینکه داده‌ها در طول زمان تغییر می‌کنند، نسبت به زمان حساس باشند و تغییرات و اصلاحاتی که در داده‌ها رخ می‌دهد روی نتایج آنها تاثیر بگذارد.

---

[1] Classification

[2] Frequency counting

[3] Association rule mining

۱-۲- جریان داده‌ها

سیستمهای مدیریت بانک اطلاعاتی(DBMS[1]) متعارف در بسیاری از کاربردهای واقعی که در آنها داده‌ها به صورت ارتباطات پایدار مدل شده‌اند، موفق بوده‌اند. همانطور که گفته شد در دهه‌های اخیر، دسته‌ای از کاربردها پدیدار شده‌اند که شامل پردازش حجم بزرگی از داده‌ها می‌باشند. داده‌هایی که در این کاربردها ظاهر می‌شوند به فرم جریان داده[2] هستند. این داده‌ها به صورت پیوسته و با ترتیب ثابتی تولید می‌شوند. حجم زیاد (معمولا نامحدود) داده‌ها که در جریان وارد می‌شود امکان ذخیره کردن کل جریان را روی رسانه‌های ذخیره‌سازی غیرممکن می‌کند، علاوه بر این در بسیاری از کاربردها سرعت ورود داده‌ها بسیار بالا است. مثالهایی که در ادامه می‌آید نمونه از این کاربردها است:

- شبکه‌های حسگر به شکل افزاینده‌ای در نظارتهای محیطی و ژئوفیزیکی، کنترل ترافیک، مسیریابی، دیدبانی در حال رایج شدن می‌باشند. اندازه‌گیریهایی که توسط حسگرها تولید و ذخیره می‌شود را می‌توان به شکل جریان داده پیوسته و نامحدود مدل کرد.

- فعالیتهای مالی و بازاری به صورت مداوم داده‌هایی مثل معاملات خرید و فروش، معاملات بازار بورس، قیمت گذاریهای لحظه‌ای و نرخ ارز خارجی را تولید می‌کنند. آنالیز آنلاین این داده‌ها می‌تواند فعالیتهای بازاری پراهمیت و الگوهای اقتصادی را شناسایی کند.

- در مخابرات در هر دقیقه حجم قابل توجهی از تماسهای تلفنی در حال تولید هستند. تحلیل بلادرنگ چنین لیست تماسهایی ممکن است الگوی مصرف مشتریان را آشکار کند و در نتیجه به بهبود کیفیت سرویسها کمک کند.

- در زمینه شبکه‌های کامپیوتری، از فعالیتهای جذابی که اخیرا مورد توجه قرار گرفته است، نظارت آنلاین و تحلیل ترافیک شبکه می‌باشد. کارهای قابل انجام در این زمینه عبارت است از ردیابی مصرف پهنای باند، آنالیز سیستمهای مسیریابی و کشف نفوذ به سرور.

---

[1] Database management system

[2] Data stream

سرعنوانهای بسته‌های IP که از وب سایتها جمع‌آوری می‌شود، می‌تواند به شکل جریان داده مدل شود.

DBMSهای قدیمی برای کاربردهای جریان داده‌ها مناسب نیستند. به عنوان مثال بسیاری از تکنیکهایی که در این DBMSهای قدیمی استفاده می‌شوند به چند بار پیمایش کل دیتاست احتیاج دارند و این در حالی است که با توجه به حجم نامحدود جریانهای داده ما فقط می‌توانیم یک بار این داده‌ها را بخوانیم، یک بار که یک داده پردازش و دور ریخته می‌شود، دیگر نمی‌توان این داده را بازگرداند. انجمن تحقیقات پایگاه داده برای تحقق امکان مدیریت و پردازش جریان داده‌ها، در حال مطالعه برای طراحی یک دسته جدید از سیستمها با عنوان سیستم مدیریت جریان داده (DSMS[1]) می‌باشد.

۱-۳- تاریخچه مختصر از جریان داده‌ها

اگرچه انجمن تحقیقات دیتابیس اخیرا به جریان داده‌ها علاقمند شده است ولی ایده جریان داده‌ها به حدود نیم قرن پیش برمی‌گردد. در سال ۱۹۶۰، Landin که در حال طراحی زبانهای محاسباتی پیاده‌سازی نشده بود واژه «جریان» را برای مدل کردن تاریخچه متغیرهای حلقه به کار برد [۲۹]. به هر حال در چند دهه بعد، جریان داده‌ها و تکنیکهای پردازش آنها بیشتر شناخته شدند و در گروه بندی «گردش داده[2]» قرار گرفتند. «گردش داده» عمدتا در زمینه توسعه تکنیکهای پردازش موازی و ارزیابی پتانسیلهای همزمانی در محاسبات استفاده می‌شد. گردش داده را می‌توان به عنوان یک مثال معیار برای جریان داده در نظر گرفت.

با ظهور کاربردهایی نظیر وب گسترده جهانی[3]، ارتباطات بی‌سیم، شبکه‌های حسگر و بسیاری کاربردهای مشابه دیگر، مدیریت و پردازش جریان داده یکی از داغ‌ترین موضوعات در دهه‌های گذشته شده است. سیستمهای مدیریت جریان داده (DSMS) زیادی توسعه داده

[1] Data stream management system

[2] Data flow

[3] World wide web

شده‌اند. سیستمهای دانشگاهی نظیر: Aurora [۳۰]، Atlas [۳۱] ، MAIDS [۳۲] و NiagaraCQ [۳۳]. علاوه بر این سیستمهای مدیریت جریان داده تجاری نیز برای پوشش دادن خصوصیات جدید جریانهای داده در حال شکل‌گیری هستند.

۱-۴- توزیع متغیر در جریانهای داده

در DBMSهای متعارف منطقی است که فرض کنیم داده‌ها ایستا هستند به این دلیل که در واقع داده‌ها نمونه‌هایی از یک توزیع ایستا می‌باشند، در حالی که این قانون برای جریان داده‌ها برقرار نیست. به عنوان نمونه، جریانهای داده سریع توسط فعالیتهای پیوسته و در طول دوره‌های زمانی طولانی تولید می‌شوند، بنابراین توزیع مقادیر داده‌های موجود در جریان می‌تواند به میزان قابل توجهی در طول زمان تغییر کند. به این مسئله با عنوانهای «سیر تکاملی داده[۱]»، «جریان پویا[۲]»، «داده تغییرکننده در زمان[۳]»، «داده تغییر محتوا دهنده[۴]» اشاره شده است[۳۴،۳۵،۳۶].

در طراحی هر مدل پردازشی جریان داده، بایستی برای حفظ کارایی و صحت تکنیک، به مسئله توزیع متغیر داده‌ها توجه شود. با تغییر در توزیع جریان داده این تغییر باید در نتایج پردازش جریان داده منعکس شود تا کاربر از این تغییر باخبر شود. به همین دلیل تمام تکنیکهای کاوش جریان داده بایستی از روشهای افزایشی جهت به روز نگه داشتن جوابهای خروجی استفاده کنند.

---

[1] Data evolution

[2] Dynamic stream

[3] Time-changing data

[4] Concept-drifting data