دانشکده مهندسی برق و کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد در رشته‌ی مهندسی کامپیوتر- نرم‌افزار

طراحی و پیاده‌سازی یک الگوریتم کارآمد برای کاوش اقلام تکراری در جریان داده‌ها

به وسیله‌ی

فاطمه نوری

اساتید راهنما

دکتر محمد هادی صدرالدینی

دکتر کورش زیارتی

شهریور ماه ۱۳۹۰


بسم الله الرحمن الرحيم

بـــه نـــام خـــدا

اظـهارنـامـه

اینجانب فاطمه نوری (۸۷۰۹۳۱) دانشجوی رشته‌ی مهندسی کامپیوتر گرایش نرم-افزار دانشکده‌ی مهندسی 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 آن‌را 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: فاطمه نوری

تاریخ و امضا: ۳۰ / ۰۶ /۱۳۹۰

به نام خدا

طراحی و پیاده‌سازی یک الگوریتم کارآمد برای کاوش اقلام تکراری در جریان داده‌ها

به کوشش

فاطمه نوری

پایان‌نامه

ارائه شده به تحصیلات تکمیلی دانشگاه به عنوان بخشی از فعالیت‌های تحصیلی لازم برای اخذ درجه کارشناسی ارشد

در رشته‌ی

مهندسی کامپیوتر (نرم‌افزار)

از دانشگاه شیراز

شیراز

جمهوری اسلامی ایران

ارزیابی شده توسط کمیته پایان نامه با درجه: عالی

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

دکتر کورش زیارتی، دانشیار بخش مهندسی کامپیوتر(رئیس کمیته).....

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

دکتر ستار هاشمی، استادیار بخش مهندسی کامپیوتر .....

شهریور ماه ۹۰

تقدیم به

دو گوهر فروزان زندگیم پدر و مادر مهربانم
و همسر عزیزم

سپاسگزاری

پس از سپاس به درگاه خداوند سبحان که توفیق آموختن علم را به این بنده عنایت فرمود تشکر از اساتید فرهیخته به عهده اینجانب فرض است. نخست اساتید دانشمند و گرانقدر جناب آقای دکتر محمدهادی صدرالدینی و دکتر کورش زیارتی که زحمات فراوان و بی‌دریغ به پای من متحمل شدند و مرا با صبر و حوصله فراوانشان راهنمایی فرمودند و در طول دوران دانشجویی از فضل و دانش‌ایشان بهره‌های فراوان بردم. لذا ضمن آرزوی سلامتی و طول عمر برایشان، تشکر و امتنان قلبی خود را نسبت به این دو استاد گرانقدر ابراز می‌دارم و امیدوارم من نیز توانسته باشم انتظاراتشان را برآورده کنم. سپس به اساتید ارجمند و فرزانه، جناب آقای دکتر غلامحسین دستغیب و جناب آقای دکتر ستار هاشمی به خاطر اینکه تقبل نمودند که به عنوان اساتید مشاور در طول تدوین پایان نامه زحمات مرا پذیرا باشند و زحمت مطالعه پایان نامه مرا به عهده گرفتند، نهایت تشکر و قدردانی خود را اعلام می‌دارم و برایشان آرزوی سلامتی و طول عمر می‌کنم.

و همچنین از شرکت مخابرات استان شیراز به خاطر حمایت مالی که از پایان‌نامه اینجانب به عمل آورده، نهایت تشکر و قدردانی خود را اعلام می‌دارم و سرانجام، نهایت والاترین سپاس خود را از صمیم قلب به محضر خانواده عزیز و گرانقدر خود که همواره مشوق و یاری دهنده‌ام بودند ابراز می‌دارم اگر چه می‌دانم هیچ نوع تشکر و قدردانی، نمی‌تواند زحماتی را که آنان در راه به ثمر رساندن اینجانب متحمل شده‌اند را جبران نماید، و« اینهمه را از نظر لطف خدا می بینم».

چکیده

به کوشش

فاطمه نوری

داده کاوی به معنای استخراج داده و اطلاعات غیر صریح و احتمالاً سودمندی از حجم زیادی از داده‌ها می‌باشد، که در گذشته ناشناخته و پنهان بوده‌اند. با انجام عملیات داده کاوی دانش جالب و گاه غیر منتظره، نظم‌ها و الگوهای پنهان، یا اطلاعات سطح بالا می‌توانند از مجموعه‌ای از داده‌های مرتبط موجود در پایگاه داده استخراج شوند. یافتن اقلام تکراری یک عمل بسیار مهم در داده کاوی محسوب می شود. با استفاده از اقلام تکراری قوانین همبستگی استخراج و بیان می شوند.

یکی از بهترین الگوریتم‌های ارائه شده در این زمینه الگوریتم تک گذره به نام NewMoment می-باشد که مجموعه‌ایی از اقلام تکراری بسته را با روش پنجره‌ی کشویی حساس به تراکنش بدست می‌آورد. الگوریتم پیشنهادی ما به نام TMoment، نیز مجموعه‌ایی از اقلام تکراری بسته را با روش پنجره‌ی کشویی حساس به تراکنش بدست می‌آورد، با این تفاوت که الگوریتم قبل از روش بیتی استفاده می‌کند که در پایگاه داده‌های خلوت فضای زیادی را از دست می‌دهد، اما در روش پیشنهادی، خود تراکنش‌ها نگه داشته می‌شوند، که باعث کاهش فضای مصرفی و همچنین زمان اجرا می‌شود.

در این پایانامه کارایی الگوریتم جدید به همراه چند الگوریتم دیگر از لحاظ حافظه مصرفی و زمان اجرا با انجام تعدادی آزمایش مورد بررسی قرار گرفته است. ارزیابی‌های صورت گرفته نشان‌دهنده برتری الگوریتم جدید از لحاظ زمان اجرا و حافظه مصرفی است.

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فصل اول

مقدمه

۱-۱ مقدمه

امروزه حجم زیادی از داده‌ها در پایگاه داده‌های مربوط به شرکت‌ها، مراکز تجاری و دولتی ذخیره می‌شود. استفاده معمول از این داده‌ها انجام عملیات گزارش‌گیری برای کاربران و مدیران است. استفاده دیگری که امروزه از حجم انبوه داده‌های ذخیره شده در پایگاه‌های داده و انباره‌های داده می‌شود انجام عملیات داده کاوی است. در عملیات داده کاوی به دنبال الگوهای پنهان و احتمالا سودمند هستیم. برخی از این الگوها در انجام تصمیم‌گیری‌های به مدیران کمک می‌کنند و برخی دیگری برای کاربران و مشتریان مفید واقع می‌شوند. الگوهایی که در عملیات مختلف داده کاوی پیدا می‌شوند انواع گوناگونی دارند. یک نوع معروف و پرکاربرد از این الگوها قواعد وابستگی یا قوانین وابستگی نام دارد.

۱-۲ بیان مسئله

معروف‌ترین کاربرد قوانین وابستگی در تحلیل سبد خرید برای فروشگاه ها و مراکز تجاری است. به عنوان مثال پس از کاوش پایگاه داده مربوط به یک فروشگاه زنجیره‌ای ممکن است مشخص شود که مشتریانی که از این فروشگاه مربا می‌خرند به احتمال ۶۰٪ کره نیز خواهند خرید. یافتن چنین قواعدی می‌تواند در چیدن قفسه‌ها، راهنمای مشتریان و مسائل مدیریتی سودمند باشد. عملیات یافتن قواعد وابستگی را کشف یا کاوش قواعد وابستگی گویند.

داده کاوی[1] به معنای استخراج داده و اطلاعات غیر صریح و احتمالاً سودمندی از حجم زیادی از داده‌ها می‌باشد، که در گذشته ناشناخته و پنهان بوده‌اند. با انجام عملیات داده کاوی دانش جالب و گاه غیر منتظره، نظم‌ها و الگوهای پنهان، یا اطلاعات سطح بالا می‌توانند از مجموعه‌ای از داده‌های مرتبط موجود در پایگاه داده استخراج شوند و از زوایای مختلف مورد بررسی قرار گیرند. بنابراین پایگاه‌های داده حجیم را می‌توان به عنوان منابعی غنی و قابل اطمینان برای تولید و وارسی برخی اطلاعات در نظر گرفت.

[1] Data mining

یافتن اقلام تکراری[1] یک عمل بسیار مهم در داده کاوی محسوب می شود. با استفاده از این اقلام تکراری قوانین همبستگی استخراج و بیان می شوند. در یک دسته از تراکنش‌ها[2]، پشتیبان[3] اقلام، درصد تکرار اقلام در کل تراکنش‌هاست . یک عنصر تکراری است اگر پشتیبان آن بزرگتر یا مساوی یک مینیمم پشتیبانی شود، که کاربر در نظر گرفته است.

۱-۲ هدف و دامنه تحقیق

در سالهای اخیر استخراج جریان داده‌ها[4] بسیار مورد توجه محققان قرار گرفته است. با به وجود آمدن محیط‌های کاربردی جدید مانند آنالیز ترافیک شبکه، کاوش جریان‌های کلیک وب[5] و کشف نفوذ به شبکه جریان داده‌ها با سرعت بالا، پیوسته[6] و نامحدود وارد سیستم می-شوند و دیگر پردازش داده‌ها نمی‌تواند به صورت ایستا[7] انجام شود.

انواع مختلف داده‌های مربوط به سری‌های زمانی، داده‌هایی که در محیط‌های پویا تولید می-شوند مانند میزان مصرف انرژی، ترافیک شبکه، مبادلات سهام، ارتباطات از راه دور، جریان‌های کلیک وب، مراقبت‌های بصری و امور نظارتی در مسایل محیطی و آب و هوا، مثال‌های از جریان‌های داده‌ایی محسوب می‌شوند.

در جریان داده‌ها امکان استفاده از روش‌های مشابه که از پایگاه داده‌های سنتی استفاده می-کنند، وجود ندارد. به دلیل حجم زیاد داده‌ها در سیستم‌ها، داده‌ها را نمی‌توان به صورت ماندگار ذخیره کرد و به صورت جریان‌های موقت می‌باشند. در جریان داده‌ها هر اقلامی[8] را فقط یکبار می‌توان آزمایش کرد، همچنین به خاطر نامحدود بودن داده‌ها سیستم با محدودیت حافظه روبرو است و در ضمن نتایج کاوش باید خیلی سریع تولید شود.

---

[1] Frequent itemset

[2] transaction

[3] support

[4] Data stream

[5] Web click stream mining

[6] continuous

[7] static

[8] itemsets

در جریان داده‌ها یکی از موضوع‌های مهم کاوش اقلام تکراری[1] است که در سالهای اخیر بسیار مورد توجه محققان قرار گرفته است.

با توجه به اهمیت جریان داده‌ها که گفته شد در این تحقیق ما سعی براین داریم که الگوریتم جدیدی را برای کاوش اقلام تکراری در جریان داده‌ها مطرح کنیم.

۱-۳ ساختار پایان‌نامه

در این پایان‌نامه ابتدا به معرفی داده کاوی و جریان‌های داده‌ای می‌پردازیم. سپس از میان عملیات داده کاوی مسئله کشف قواعد داده‌ای را به طور دقیق بررسی خواهیم کرد. روش‌ها و الگوریتم‌های ارائه شده برای یافتن اقلام تکرار شونده که مهمترین گام برای کشف قواعد وابستگی است را بررسی خواهیم کرد. این الگوریتم‌ها از جنبه‌های مختلف مقایسه و مزایا و معایب هر کدام بررسی می‌شوند. در ادامه یک الگوریتم جدید برای کشف اقلام تکرار شونده در جریان داده‌ای ارائه می‌کنیم و در پایان ارزیابی و نتایج آزمایشات و نتیجه‌گیری نهایی را خواهیم داشت.

[1] Frequent itemset mining

فصل دوم

مبانی نظری

۲-۱ داده کاوی

اخیرا سرعت تولید و جمع‌آوری داده‌ها در پایگاه‌های داده به صورت روزافزونی زیاد شده است. استفاده گسترده از بارکد در فروش تولیدات، کامپیوتری شدن تعداد زیادی از کارهای تجاری، اداری و دولتی و پیشرفت در زمینه ابزار جمع‌آوری داده‌ها ما را با انبوهی از داده‌ها مواجه کرده است. امروزه پایگاه داده در زمینه‌های تجاری، اداری، علمی، مهندسی و زمینه‌های دیگر استفاده می‌شوند. تعداد چنین پایگاه داده‌ای به دلیل نیاز مبرم به جمع‌آوری و گزارش‌گیری از داده‌ها و همچنین وجود سیستم‌های قدرتمند پایگاه داده در حال افزایش است. این چنین رشد فزاینده‌ای در داده‌ها نیاز ضروری برای ابزار جدیدی که بتواند به طور بتواند به طور هوشمند و خودکار این داده‌ها پردازش کرده و به اطلاعات و دانش‌های سودمند تبدیل کند احساس می-شود. در نتیجه داده کاوی به یک زمینه تحقیقاتی با اهمیت فراوان تبدیل شده است.

از داده کاوی همچنین به عنوان کشف دانش در بانک‌های داده‌ای یاد می‌شود، که به معنی فرایند استخراج اطلاعات غیرصریح و احتمالا سودمندی از پایگاه داده است که در گذشته ناشناخته و پنهان بوده‌اند. با انجام عملیات داده کاوی دانش‌های جالب و گاه غیرمنتظره، نظم‌ها و الگوهای پنهان یا اطلاعات سطح بالا می‌توانند از مجموعه‌ای از داده‌های مرتبط در پایگاه داده استخراج شوند و از زوایای مختلف مورد بررسی قرار گیرند. بنابراین پایگاه‌های داده حجیم را می‌توان به عنوان منبعی غنی و قابل اطمینان برای تولید و وارسی برخی از دانش‌ها و اطلاعات در نظر گرفت.

کاوش اطلاعات و دانش از پایگاه داده حجیم به عنوان یک موضوع کلیدی برای محققینی که در زمینه پایگاه‌های داده و یادگیری ماشین کار می‌کنند و به فرصتی برای کسب در آمدهای بیشتر توسط شرکت‌های صنعتی و تجاری تبدیل شده است. دانش‌های کشف شده توسط داده کاوی می‌توانند در مدیریت اطلاعات، پردازش گزارش‌ها، انجام تصیم‌گیری‌ها و بسیاری زمینه‌های دیگر استفاده شوند. به علت وجود گسترده داده‌ها در حجم زیاد و نیاز مبرم به تبدیل این داده‌ها به اطلاعات و دانش مفید برای کاربردهای مختلف، داده کاوی در سال‌های اخیر توجه زیادی را به خود جلب کرده است[۱،۲].

داده کاوی موضوعی وابسته به کاربرد است و کاربردهای مختلف نیازمند روش‌ها و تکنیک-های داده کاوی مختلفی هستند. کاوش قواعد وابستگی، دسته‌بندی، خوشه‌بندی، پیش‌بینی و

تحلیل سری‌های زمانی از جمله مهمترین روش‌ها و تکنیک‌های داده کاوی به شمار می‌آیند. در ادامه هر کدام از روش‌ها به صورت خلاصه توضیح داده می‌شود.

کشف قواعد وابستگی در جریان‌های داده‌ای از جذابیت زیادی در بین متخصصان داده کاوی برخوردار است. در این تکنیک وابستگی‌ها و ارتباطات میان داده‌ها بدست می‌آید. نتیجه این عملیات داده کاوی دسته‌ای از قواعد است که به آنها قواعد وابستگی گفته می‌شود.

یکی دیگر از روش‌های مهم داده کاوی توانای انجام عملیات دسته‌بندی در حجم زیاد داده-هاست. این عملیات کاوش قوانین دسته‌بندی نیز نامیده می‌شود. در این روش اشیا موجود در یک پایگاه داده براساس مقادیر چند خصوصیت از آنها به دسته‌های مجزا تقسیم می‌شوند. در دسته‌بندی داده‌ها آزمایشی تحلیل می‌شوند. در دسته‌بندی داده‌ها مجموعه داده‌های آزمایشی برچسب کلاسها مشخص است. برای هر کلاس داده‌های آزمایشی مدلی براساس خصوصیات داده‌ها ساخته می‌شود. حاصل عملیات دسته‌بندی می‌تواند یک درخت تصمیم یا مجموعه‌ای از قوانین دسته‌بندی باشد که برای فهم بهتر داده‌های موجود در پایگاه داده و همچنین دسته-بندی داده‌هایی که در آینده به پایگاه داده اضافه می‌شوند به کار می‌رود. دسته‌بندی داده‌ها ارتباط تنگاتنگی با کاوش قواعد وابستگی دارد، به طوری که گاهی دسته‌بندی را به کمک قواعد وابستگی انجام می‌دهند.

به عنوان مثال برای فروشنده‌ی ماشین دسته‌بندی مشتریانش براساس تمایل و علاقه هر کدام به انواع مختلف ماشین مطلوب است به طوری که بتواند به مشتریانش خدمات بهتری ارائه دهد و کاتالوگ‌های محصولات جدید مورد نظر آنها را برایش بفرستد و درآمد حاصل از فروش خود را افزایش دهد. خوشه‌بندی داده‌ها از مهمترین روش‌های داده کاوی به شمار می-آید. در خوشه‌بندی مجموعه‌ای از داده‌ها گروه‌بندی می‌شوند. فرق خوشه‌بندی با دسته‌بندی داده‌ها در این است که در خوشه‌بندی بر خلاف دسته‌بندی تعداد کلاس‌ها در ابتدا مشخص نیستند. خوشه‌بندی داده‌ها براساس اصل مفهومی زیر صورت می‌گیرد:

حداکثر کردن شباهت‌های اعضای هر کلاس و حداقل کردن شباهت‌ها بین اعضای مربوط به کلاس‌های مختلف.

به عنوان مثالی از خوشه‌بندی مجموعه‌ای از کالاها را می‌توان در ابتدا به صورت مجموعه‌ای از کلاسهای مختلف خوشه‌بندی کرده و سپس مجموعه‌ای از قوانین را براساس این چنین دسته‌بندی نتیجه‌گیری کرد.

پیش‌بینی یکی دیگر از تکنیک‌های داده کاوی است که در آن مقادیر ممکن برای متغییرهای نامعلوم پیش‌بینی می‌شوند. در پیش‌بینی ابتدا داه‌هایی که به متغییر نامعلوم مربوط هستند به وسیله برخی تحلیل‌های آماری پیدا می‌شوند، سپس از برخی روش‌های هوشمند مانند شبکه-های عصبی و الگوریتم‌های ژنتیک برای انجام پیش‌بینی استفاده می‌شود. برای مثال مقدار حقوقی که یک کارمند می‌گیرد را می‌توان با استفاده از چگونگی توزیع حقوق کارمندان مشابه