ISSN: 2745-3673
http://www.Rcsj.ir

مجلـه‌ی علمی پژوهش در علوم رایانه (سال چهارم)
شماره ۱۳ / بهار ۱۳۹۸ / ص ۷۰-۶۲

بررسی تکنیک‌های حفظ حریم‌خصوصی در داده‌کاوی

توحید ساکت
گروه کامپیوتر، واحد میانه، دانشگاه آزاد اسلامی، میانه، ایران.

نام نویسنده مسئول:
توحید ساکت

چکیده
داده‌کاوی بیشتر برای تأمین امنیت و برای شناسایی فعالیت‌های افراد خرابکار شامل جابه‌جایی پول و ارتباطات بین آنها و همچنین شناسایی و ردگیری خود آنها با بررسی سوابق مربوط به مهاجرت و مسافرت‌ها است. صاحبان داده‌ها به علت ترس از افشای اطلاعات شخصی و محرمانه خود توسط دیگران، چندان تمایلی جهت انجام داده‌کاوی روی داده‌های خود نداشته ولی این مطلب را نیز می‌دانند که بدون انجام داده‌کاوی به نتایج و دانش مفید از داده‌های یکدیگر دسترسی پیدا نمی‌کنند. مالکان داده باید اطلاعات خود را برای مطالعه و تحقیق در اختیار پژوهشگران قرار دهند و اطلاعات داده شده نباید باعث نقض حریم خصوصی افراد شود به همین دلیل باید به دنبال تکنیک‌هایی باشیم تا به مالکان داده‌ها این اطمینان را بدهیم که امکان تبادل و انتشار داده‌ها وجود دارد و می‌توان با حفظ حریم خصوصی، داده‌ها را در اختیار پژوهشگران قرار داد. استفاده از اینترنت و فناوری اطلاعات در کنار مزایایی که دارد، خطر برملا شدن اسرار خصوصی را به همراه دارد. از تکنیک‌های حفظ حریم‌خصوصی در داده‌کاوی می‌توان به تکنیک‌های تفاضل، قوانین انجمنی، انجمنی مخفی، توزیع، گمنامی، طبقه‌بندی، خوشه‌بندی و برون‌سپاری اشاره کرد. در این مقاله مفاهیم امنیت، روش‌های حفظ حریم‌خصوصی، داده‌کاوی و موضوعات مرتبط با این مفاهیم را توضیح داده و به بررسی تکنیک‌های حفظ حریم‌خصوصی در داده‌کاوری خواهیم پرداخت.

واژگان کلیدی: حریم‌خصوصی، امنیت، داده‌کاوی، قوانی انجمنی.

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۳ ، بهار ۱۳۹۸، ص ۷۰- ۶۲

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

با توجه به ناشناس بودن کاربران و سهولت استفاده از اینترنت تجاوز به حریم خصوصی افراد به سرعت افزایش یافته و صاحب نظران را در جهت حمایت از حریم خصوصی افراد سوق داده است. این تحقیق ابتدا به مفاهیم حریم خصوصی، اهمیت حریم خصوصی،امنیت و سابقه حریم خصوصی پرداخته سپس حریم‌خصوصی در ایران، اسناد بین المللی و دیگر کشورها را مورد بررسی قرار می‌دهد و در نهایت با ارائه مباحثی پیرامون اینترنت و حریم خصوصی و نیز چگونگی سیاست‌گذاری و حمایت از این حق به نتیجه‌گیری در حوزه مذکور، می‌انجامد. حریم خصوصی از جمله حقوقی است که انسان‌ها به دلیل نیازهای شخصی از یک طرف به آن وابسته‌اند و از طرف دیگر به دلیل ضرورت زندگی جمعی مکلف‌اند این حق را نسبت به دیگران به رسمیت بشناسند. اما امروزه با گسترش ابزارهای اطلاع رسانی و استفاده گسترده از اینترنت، این حق به یکی از چالش انگیزترین مسائل حقوق بشر تبدیل شده است [۱].

حفظ حریم خصوصی افراد از موضوعات بسیار با اهمیت برای افراد و سازمان‌ها است به خصوص زمانی که سازمان‌ها با امر داده‌کاوی مواجه هستند و باید اطلاعات کاربران خود را برای مراکز داده‌کاوی ارسال کنند تا دانش مورد نیاز آنها برای کاربردهای آینده از این داده‌ها استخراج شود. حال در مواجه با داده‌های با حجم بسیار زیاد، بکارگیری محیط‌های توزیع شده، مدیریت بهتری را به ارمغان خواهد آورد اما چالش جدید حفظ حریم خصوصی افراد در این محیط‌ها است، یعنی بکارگیری روشی برای اینکه اطلاعات حساس کاربران در حین ارسال اطلاعات آن ها به مراکز داده‌کاوی از دسترسی‌های غیرمجاز در امان بماند. یکی از تکنیک‌هایی که برای حفظ حریم خصوصی افراد در محیط‌های متمرکز پیشنهاد شده است. نقض حریم خصوصی می‌تواند زمینه ساز ایجاد مشکلاتی برا ی افراد گردد که این خود به صورت معضلی در بحث تهیه داده‌های آماری برای تحلیل گران تبدیل شده است. هر مدلی که در مورد حفظ حریم خصوصی داده‌های منتشر شده، دو هدف اساسی را در نظر می‌گیرد. این دو هدف حفظ حریم خصوصی و میزان سودمندی داده‌های منتشر شده است. مسئله ای که در اینجا باعث دشوار شدن کار می‌شود، متضاد بودن این دو هدف است، به عبارت دیگر در صورتی که مدلی ارائه شود که حریم خصوصی را به طور صد درصد تضمین نماید، میزان سودمندی داده‌ها مقدار صفر درصد خواهد بود، چرا که تنها در صورتی حریم خصوصی به طور کامل و بی نقص حفظ خواهد شد که هیچ داده‌ای منتشر نشود. حفظ حریم خصوصی مطلق نیز امکان پذیر نیست. دلیل عمده آن وجود حملات با دانش قبلی دشمن است. حریم خصوصی در شبکه‌های اجتماعی به وسیله گمنامی به دست می‌آید. در سال‌های اخیر تعاریف گوناگونی از گمنامی گراف ارائه و مورد مطالعه قرار گرفته است [۲].

در ترکیه و طبق قوانین این کشور هر کس حق دارد که خواهان احترام به زندگی خصوصی و خانوادگی‌اش باشد و حریم افراد و زندگی خانوادگی نباید مورد بی‌حرمتی قرار بگیرد مگر در مواردی که از طرف یک نهاد دولتی دستوری در این زمینه صادر باشد. در جمهوری اسلامی ایران و همچنین مسلمانان مؤظف شده‌اند که نسبت به افراد غیرمسلمان با اخلاق حسنه و قسط و عدل اسلامی عمل کنند و حقوق انسانی آنها را رعایت کنند و ازآنجا که حق حریم خصوصی از بدیهی‌ترین حقوق انسانی به شمار می‌رود، بنابراین، حق حفظ حریم‌خصوصی غیرمسلمانان در قانون اساسی جمهوری اسلامی ایران، جایگاه خاصی دارد.

قوانین کشور هند، به صراحت حق حفظ حریم خصوصی را به رسمیت نمی‌شناسد. با این حال، دیوان عالی کشور اعلام کرد که این حق به صورت تلویحی در قانون اساسی مورد اشاره قرار گرفته است و هیچ شخصی را نمی‌توان از زندگی و آزادی فردی محروم کرد، مگر بر ساس رویه‌های تعیین شده توسط قانون. در هند، هیچ قانون جامعی به منظور حفاظت اطلاعات وجود ندارد. قوانین کشور اردن، به صراحت حق حریم خصوصی افراد را عنوان و مطرح می‌کند اما این حق از قرن‌ها قبل در باورها و سنت اسلام وجود دارد که بر طبق آن، اعضای جامعه و زندگی آنها در مقابل انواع مزاحمت‌ها، مورد حمایت قرار می‌گیرد.

مالکان داده باید اطلاعات خود را برای مطالعه و تحقیق در اختیار پژوهشگران قرار دهند و اطلاعات داده شده نباید باعث نقض حریم خصوصی افراد شود به همین دلیل باید به دنبال روش‌هایی باشیم تا به مالکان داده‌ها این اطمینان را بدهیم که امکان تبادل و انتشار داده‌ها وجود دارد و می‌توان با حفظ حریم خصوصی، داده‌ها را در اختیار پژوهشگران قرار داد. استفاده از اینترنت و فناوری اطلاعات در کنار مزایایی که دارد، خطر برملا شدن اسرار خصوصی را به همراه دارد لذا اهمیت انجام این تحقیق احساس می‌شود.

پیشینه تحقیق

در سال ۲۰۱۳، Belwal و همکارانش، پشتیبانی و اعتماد قواعد حساس را بدون تغییر مستقیم پایگاه داده‌ی معین، کاهش دادند. با این حال، تغییر به طور غیر مستقیم می‌تواند از طریق پارامترهای در حال ترکیب مرتبط با معاملات پایگاه داده و قوانین انجمنی انجام شود [۶]. در سال ۲۰۱۱، مشکلات پیش رو در داده کاوی به طور گسترده‌ای در بسیاری از جوامع مانند پایگاه داده، کنترل افشای آماری و جامعه رمزنگاری توسط Nayak و همکاران سنجیده شدند [۷]. در سال ۲۰۱۱، Islam و همکاران یک معماری جدید شامل روش‌های

ISSN: 2745-3673
http://www.Rcsj.ir

مختلف ارائه کردند که تمام ویژگی‌ها در پایگاه داده را تحت تاثیر قرار داد. یافته‌های تجربی نشان داد که معماری ارائه شده در حفظ الگوهای اصلی در یک مجموعه داده مختل و آشفته، بسیار مؤثر است [۸]. در سال ۲۰۱۱، Mukkamala و همکاران مجموعه‌ای از روش-های نگاشت مبتنی بر فازی را در زمینه ویژگی‌های حفظ حریم خصوصی و توانایی حفظ همان ارتباط با سایر زمینه‌ها مقایسه کردند [۹].

در سال ۲۰۱۳، ارتباط روش حفظ حریم خصوصی در داده کاوی به طور کامل توسط Matwin مورد آنالیز و بحث قرار گرفته است. استفاده از روش‌های خاص، توانایی آنها را برای جلوگیری از استفاده تبعیض آمیز از داده‌کاوی نشان داد و پیشنهاد کردند که هر گروه نباید در تعمیم داده‌ها نسبت به جمعیت عمومی، هدف قرار داده شود [۱۰]. در سال ۲۰۱۳، Sachan و همکارانش راه حل‌های فعلی حفظ حریم خصوصی را برای خدمات ابری مورد آنالیز قرار دادند، که در آنها راه حل بر اساس اجزای رمزنگاری پیشرفته تعیین می‌شود. این راه حل، دسترسی ناشناس، توانایی جدا کردن و حفظ محرمانه بودن داده‌های منتقل شده را ارائه داد. در نهایت، این راه حل پیاده‌سازی می‌شود، نتایج تجربی به دست می‌آیند و عملکرد مقایسه می‌گردد [۱۱]. در سال ۲۰۱۳، Vatsalan و همکارانش روشی به نام «ارتباط ضبط حریم خصوصی» را بررسی کردند، که با حفاظت از حریم خصوصی، ارتباط پایگاه داده‌ها به سازمان‌ها را مجاز کرد [۱۲]. در سال ۲۰۱۲، Qi و ZONG چند روش موجود از داده‌کاوی را برای حفاظت از حریم خصوصی بسته به توزیع داده‌ها، الگوریتم‌های استخراج، و پنهان کردن داده‌ها و قوانین مرور کردند. با توجه به توزیع داده‌ها، در حال حاضر تنها چند الگوریتم برای حفاظت از حریم خصوصی در داده کاوی بر اساس داده‌های متمرکز و پراکنده استفاده می‌شود [۱۳]. در سال ۲۰۱۰، Vijayarani و همکاران در مورد روش‌های جامعه آماری، پایگاه داده و جامعه رمزنگاری، تحقیقی انجام دادند که نیاز به هزینه بالا دارد [۱۴].

روش

در این مطالعه، تحقیقات و مقالات خارجی چاپ شده تا سال ۲۰۱۶ به زبان انگلیسی که در زمینه تکنیک‌های حفظ حریم خصوصی در داده‌کاوی انجام شده بودند، مورد بررسی قرار گرفتند. این مطالعات از طریق بانک‌های اطلاعاتی Google Scholar، Ovid، Science Direct، Civilica و با استفاده از کلیدواژه‌های حریم‌خصوصی، داده‌کاوی، قوانین انجمنی و امنیت به دست آمد. نتیجه‌ی این جستجو دستیابی به تعدادی مقاله‌ی اصلی و مرتبط با موضوع بود که از این میان تعداد ۱۸ مطالعه وارد پژوهش شدند. شرایط ورود مقالات به مطالعه شامل تکنیک‌های حفظ حریم خصوصی در داده‌کاوی بود. هدف این مقاله بررسی تکنیک‌های حفظ حریم خصوصی در داده‌کاوی است و از چند قسمت تشکیل شده است که قسمت اول شامل مقدمه‌ای بر حفظ حریم خصوصی در داده‌کاوی است. قسمت دوم به پیشینه مطالعات انجام شده در این خصوص می‌پردازد. قسمت سوم روش را بیان می‌کند. قسمت چهارم شامل یافته‌ها است که از مطالعات منتخب فیش‌برداری شد و مطالب جمع‌آوری شده در حیطه "تکنیک‌های حفظ حریم خصوصی در داده‌کاوی" تقسیم‌بندی و خلاصه‌سازی شد و در قسمت چهارم به بحث و نتیجه‌گیری پرداخته می‌شود.

یافته‌ها

تکنیک‌های حفظ حریم خصوصی در داده‌کاوی

تکنیک‌های حفظ حریم خصوصی با تغییر داده‌ها برای پوشاندن و یا پاک کردن داده‌های حساس اصلی برای پنهان کردن آنها، از داده‌ها محافظت می‌کنند. به طور معمول، آنها بر اساس مفاهیم شکست حریم خصوصی، ظرفیت برای تعیین اطلاعات اصلی کاربر از اطلاعات اصلاح شده، از دست دادن اطلاعات و برآورد خسارت دقت داده‌ها می‌باشند. هدف اصلی این روش‌ها ارائه دقت و حفظ حریم خصوصی است. روش‌های دیگر که از تکنیک‌های رمزنگاری برای جلوگیری از نشت اطلاعات استفاده می‌کنند، از نظر محاسباتی بسیار گران روش‌های توزیع داده و پارتیشن بندی پراکنده‌ی افقی یا عمودی از طریق نهادهای متعدد استفاده می‌کنند. گاهی اوقات افراد تمایلی برای به اشتراک گذاشتن کل مجموعه داده‌ها ندارند و ممکن است با استفاده از انواع پروتکل‌ها تمایل به مسدود کردن این اطلاعات داشته باشند. منطق اصلی برای پیاده‌سازی چنین تکنیک‌هایی، حفظ حریم خصوصی افراد حین استخراج نتایج از کل داده‌ها است.

تکنیک قوانین انجمنی

مسائل مربوط به انتقال داده‌ها برای حفظ حریم خصوصی در روش داده کاوی و طبقه بندی انجمنی در یک سناریوی داده-افزایشی را در سال ۲۰۱۱ مشخص شد. به منظور حفظ استاندارد حریم خصوصی به نام گمنامی، یک الگوریتم زمان چند جمله‌ای افزایشی برای تبدیل داده‌ها ارائه شده است. هنگام ساخت یک مدل طبقه بندی انجمنی، کیفیت باز هم می‌تواند حتی تحت تبدیل حفظ شود. آزمایشات

ISSN: 2745-3673
http://www.Rcsj.ir

مختلفی برای ارزیابی عملکرد الگوریتم توسعه یافته انجام می‌شود و با الگوریتم غیر افزایشی مقایسه می‌گردد. این الگوریتم به گونه‌ای ایجاد می‌شود که در هر مسئله، کارآمدتر باشد.

کاوش قوانین انجمنی، که در سال ۱۹۹۳ توسط Agrawal و همکارانش ارائه گردید، ابتدا مجموعه-اقلامِ متکرر را در پایگاه‌داده می-یابد، سپس قوانین انجمنی را بر اساس این مجموعه-اقلام تولید می‌کند. هدف متدولوژی‌های مخفی‌سازی قوانین انجمنی این‌است که پایگاه‌داده اولیه را چنان پاک‌سازی کنند که حداقل یکی از اهداف زیر محقق شود:

هیچ‌کدام از قوانین انجمنی که از دید دارنده‌داده‌ها حساس محسوب می‌شود و در پایگاه‌داده اولیه با استفاده از آستانه‌های از پیش تعیین شده‌ی درجه اطمینان و پشتیبان، قابل استخراج است، در پایگاه‌داده پاک‌سازی شده با همان آستانه‌ها و یا آستانه‌های بزرگتر قابل استخراج نباشد.

همه قوانینِ غیرحساس که در پایگاه‌داده اولیه با استفاده از آستانه‌های از پیش تعیین شده‌ی درجه اطمینان و پشتیبان، قابل استخراج هستند، در پایگاه‌داده پاک‌سازی شده نیز با استفاده از آن آستانه‌ها یا آستانه‌های بزرگتر با موفقیت استخراج شوند.

هیچ قانون جدیدی در پایگاه‌داده پاک‌سازی شده با استفاده از آستانه‌های از پیش تعیین شده‌ی درجه اطمینان و پشتیبان، استخراج نشود که در پایگاه‌داده اولیه با همان آستانه‌ها قابل استخراج نباشد [۱۵].

تکنیک تفاضل

روش حریم خصوصی تفاضلی به طور گسترده‌ای مورد بررسی قرار گرفته است تا با به حداقل رساندن احتمال شناسایی سوابق، حداکثر امنیت را به پایگاه داده‌های آماری خصوصی ارائه دهد. چندین گروه قابل اعتماد وجود دارد که دارای یک مجموعه داده از اطلاعات حساس مانند پرونده‌های پزشکی، اطلاعات ثبت‌نام رای دهندگان، استفاده از ایمیل، و گردشگری است. هدف اصلی ارائه اطلاعات آماری سراسری در مورد اطلاعات در دسترس عموم است، حین حفاظت از حریم خصوصی کاربرانی که اطلاعاتشان در مجموعه داده موجود است. مفهوم "تمایزناپذیری" که "حریم خصوصی تفاضلی" نیز نامیده می‌شود به معنای "حریم خصوصی" در زمینه پایگاه داده‌های آماری است. اطلاعات باید در انباره محافظت شوند و انتقال باید از طریق پروتکل‌های امنیت داده ایجاد گردد. علاوه بر این، در صورتی که هدف، حفظ حریم خصوصی داده باشد، آنگاه یک سری مراحل دیگر باید برای محافظت محرمانه از افراد مندرج در داده‌ها در نظر گرفته شود. معرفی برخی تعاریف برای تقویت مفاهیم حفظ حریم خصوصی در داده‌کاوی مهم است. به خصوص، یک شناسه‌ی صریح، یک ویژگی است که اجازه یک ارتباط مستقیم از یک نمونه (یک ردیف در T) به یک کاربر i را می‌دهد. به عنوان مثال، با شناسایی یک شماره تلفن همراه یا شماره گواهینامه رانندگی، ممکن است آشکارا به ردیف در T متصل شود، جایی که این شناسه‌ی صریح به فرد i در آن تعبیه شده است. در مقابل، یک شبه شناسه که مجموعه‌ای از ویژگی های غیر صریح افراد است نیز ممکن است یک ردیف در T را به یک فرد خاص متصل کند. به عنوان مثال، در ایالات متحده، سه گانه شبه شناسه: تاریخ تولد، کدپستی و جنسیت، بیشتر جمعیت کشور را شناسایی می‌کند. با ترکیب یک مجموعه داده اطلاعات بهداشت و درمان عمومی با لیستی از رای دهندگان و استفاده از شبه شناسه در دسترس عموم قرار گرفت، سوئینی متقاعد کرد که استخراج پرونده‌های مخفی سلامت تمام کارمندان دولت از یک مجموعه داده منتشر شده از فرماندار ماساچوست امکانپذیر است، که در آن تنها شناسه‌ی صریح حذف شده است. به طور کلی، روش‌های اولیه حفاظت از هویت حفظ حریم خصوصی در داده‌کاوی که براساس ایده‌های ساده بدست می‌آیند، برای مردم شناخته شده هستند چون آنها به وفور در متون و فیلم‌ها در دسترس می-باشند. این مفاهیم به عنوان "پنهان کردن در جمعیت" و "استتار" توصیف می‌شوند [۴].

تکنیک طبقه‌بندی

طبقه بندی روش‌ها و الگوریتم‌های استاندارد حفظ حریم خصوصی برای هر کلاس توسط ساداس‌ویم و همکاران مرور شد، که در آن محاسن و محدودیت‌های روش‌های مختلف نشان داده می‌شوند. در تحقیقی یونگ و همکارانش یک روش طبقه بندی همسایگی مخفی بر اساس روش‌های SMC برای حل چالش‌های حفظ حریم خصوصی در چند مرحله ارائه دادند که شامل انتخاب همسایگی مخفی حفظ حریم خصوصی و طبقه بندی حفظ حریم خصوصی است. الگوریتم پیشنهادی از نظر دقت، عملکرد، و حفاظت از حریم خصوصی، متعادل است. علاوه بر این، با تنظیمات مختلف برای برآوردن شرایط بهینه سازی، سازگار است [۷].

در سال ۲۰۱۰، سینگ و همکارانش یک طبقه بندی ساده و کارآمد از حفظ حریم خصوصی برای داده‌های ابری ارائه دادند. مقیاس شباهت برای محاسبه نزدیکترین همسایگان برای طبقه بندی مورد استفاده قرار می‌گیرد و آزمون برابری برای محاسبه آن بین دو رکورد رمزگذاری شده، ایجاد شده است. این رویکرد باعث تسهیل محاسبات همسایه محلی ایمن در هر گره در ابر شده و سوابق نهان را از طریق طرح طبقه‌بندی وزنی طبقه‌بندی کرده است. تمرکز بر روی مهیا ساختن استحکام در روش پیشنهادی از اهمیت برخوردار است به طوری

ISSN: 2745-3673
http://www.Rcsj.ir

که بتواند به وظایف داده کاوی متعدد تعمیم داد، که در آنها امنیت و حریم خصوصی مورد نیاز هستند. در سال ۲۰۱۰ بائوتو یک الگوریتم کارآمد بر اساس ماتریس اغتشاش تصادفی برای محافظت از کاوش طبقه‌بندی حریم خصوصی معرفی کردند. این بر روی داده‌های گسسته از نوع کاراکتر، نوع بولی، نوع طبقه بندی و نوع عددی اعمال می‌شود. نتایج تجربی، ویژگی‌های بهینه الگوریتم پیشنهادی را از نظر حفاظت از حریم خصوصی و دقت محاسبات کاوش نشان دادند، که در آن فرآیند محاسبات بسیار ساده شده است اما هزینه بالاتر رفته است. در سال ۲۰۰۸ وایدا و همکارانش یک رویکرد برای داده کاوی تقسیم شده به صورت عمودی ایجاد کردند. این روش می‌تواند انواع کاربردهای داده کاوی را به صورت درخت‌های تصمیم گیری، اصلاح و گسترش دهند. راه حل‌های کارآمدتر برای پیدا کردن مرز بالایی محکم بر روی پیچیدگی نیاز هستند [۱۰].

تکنیک توزیع

داده کاوی حفاظت از حریم خصوصی توزیعی با تکنولوژی های موجود به سه گروه از قبیل: ۱- محاسبه چند قسمتی ایمن-۲- اختلال-۳-سوالات محدود دسته بندی شده‌اند. این روش سود و زیان‌های هر متدی را با توسعه دادن و آنالیز کردن طرح حفاظت از حریم خصوصی توضیح داده تا از تعداد بازیابی‌ها پشتیبانی کند. یک عامل محرک مطرح شده است تا محاسبه امن را با ارائه یک سیستم معتبر در شبکه بی‌سیم مطالعه کند. این روش یک انگیزه برای گره‌های بدرفتار عرضه کرده تا بدرستی رفتار کنند. نتایج تجربی تأثیرگذاری سیستم را در تشخیص گره‌های بدرفتار و افزایش میانگین خروجی در کل شبکه را آشکار کرده است. بنابراین ریسک‌های حریم خصوصی که به بازیابی داده‌ها در سیستم ابری تأکید کرده و یک چارچوب توزیعی را ارائه کرده تا اینجنین ریسک‌هایی را دور کند. روش مطرح شده شامل طبقه بندی، تجزیه و توزیع شده است. این روش از بازیابی اطلاعات با حفاظت از سطوح حریم شخصی، تقسیم کردن اطلاعات به بخش‌هایی و ذخیره کردن آنها در تأمین کننده‌های ابری مناسب، جلوگیری کرده است. اگرچه سیستم این روش یک راه حل مناسب را پیشنهاد داده تا حریم خصوصی را از حمله‌هایی که بر اساس داده‌کاوی (mining based) انجام شده‌اند را امن نگه دارد. برای مثال مشتری باید تحلیل داده‌های جهانی را برای مجموعه داده‌های کامل انجام دهند، در شرایطی که تحلیلگر نیاز به دسترسی به اطلاعات از طریق موقعیت‌های مختلف با عملکرد ضعیف را دارد [۱۶].

در این روش یک چارچوب توزیع شده‌ی جدید برای مهیا ساختن حفظ حریم خصوصی برای ذخیره سازی برون سپاری شده‌ی داده-ها ارائه شده است. روش‌های مختلفی برای تجزیه داده‌ها استفاده می‌شود که حاکی از جستارهای بهینه هنگام پیاده‌سازی در اینگونه سیستم توزیع شده است. یک تعریف جدید برای حریم خصوصی بر اساس مجموعه‌های پنهان کننده ویژگی‌ها ایجاد شده است. این روش در مورد دستیابی حریم خصوصی امن در روش‌های تجزیه پیشنهادی بحث می‌کند و بهترین روش تجزیه حفظ حریم خصوصی را شناسایی می‌کند. یکی دیگر از کارهای آینده شامل شناسایی الگوریتم‌های بهبود یافته برای تجزیه، گسترش دامنه روش‌های موجود برای تجزیه است. برای داده‌کاوی مبتنی بر حفظ حریم در مجموعه داده‌های توزیع شده، هدف اصلی عبارتست از اجازه محاسبه آمار جمعی برای پایگاه داده کامل با اطمینان از حریم خصوصی برای اطلاعات محرمانه از پایگاه داده‌های شرکت کننده. از این رو، الگوریتم‌های حفظ حریم خصوصی، نیاز به بهبود بیشتری بر اساس موازنه بین دقت و حریم بازسازی دارند [۱۷].

تکنیک گمنامی

روش گمنامی بر پایه حفاظت از حریم خصوصی داده‌کاوی یک راه حل مقیاس پذیر برای هر تکراری است می‌تواند حداقل یک تعمیم پذیری را برای هر ویژگی که در ارتباط (Linking) شامل شده را بررسی کند. روش‌های بازیابی داده‌ها در قالب مفهوم تعمیم‌سازی داده بررسی شده در حالی که بازیابی داده با مخفی کردن اطلاعات اصلی به جای روندها و الگو اجرا شده است. بعد از پوشش دهی داده‌ها، روش‌های بازیابی داده‌های معمولی بدون هیچ اصلاحی به کار گرفته می‌شوند. در این روش دو عامل کلیدی مهم، کیفیت و مقیاس‌پذیری به طور خاصی مورد تمرکز قرار گرفته‌اند. مسئله کیفیت از طریق مبادله بین اطلاعات و حریم شخصی معین شده است. در این روش اندازه‌گیری نبود اطلاعات درست و الگوریتم رمزگذاری مؤثر معرفی شده‌اند تا فقدان اطلاعات را به حداقل برسانند. بررسی تجربی بر روی داده‌های پزشکی آشکار کرده که روش مطرح شده بیشتر جواب‌های جستجوی قابل اعتماد را پذیرفته است. این کار راه‌های امید بخشی را برای تحقیقات آینده می‌گشاید [۴].

روش‌های داده‌کاوی اجازه استخراج اطلاعات از مجموعه بزرگی از داده‌ها را می‌دهند. اطلاعات داده‌ها از روی داده‌های اصلی به صراحت استخراج می‌شود، بنابراین می‌توان استنتاج را از روی داده‌های اصلی انجام داد، احتمالاً قرار دادن محدودیت‌های تحمیل شده بر حفظ حریم خصوصی داده‌های اصلی خطرناک است. این روش همچنین برای گمنامی نیز صدق می‌کند. بنابراین تمایل به اطمینان و گمنامی داده‌ها در مجموعه داده‌ها ممکن است نیاز به اعمال محدودیت در خروجی فرآیند داده‌کاوی داشته باشد در حالی که سودمندی

ISSN: 2745-3673
http://www.Rcsj.ir

داده‌ها را با اشاره به انواع خواص تحلیلی، حفظ می‌کند. مجموعه‌ای از آزمایشات بر روی پایگاه داده‌های متوالی مختلف واقعی نشان داده است که این روش، بطور قابل ملاحظه‌ای نتایج الگوی استخراج متوالی را نه تنها از نظر الگوهای استخراج شده بلکه از نظر پشتیبانی نیز حفاظت کرده است [۱۸].

تکنیک خوشه‌بندی

روش‌های مبتنی بر خوشه‌بندی از روش‌های تعمیم برای حفظ داده در برابر حملات استفاده می‌کند. داده‌ها را به گروه‌هایی با توجه به شباهت مقادیر ویژگی‌ها یا توپولوژی همسایه‌ها خوشه‌بندی می‌کند. سپس به جای انتشار اطلاعات دقیق، این روش خلاصه‌ای از هر گروه را پست می‌کند، اشکال حیاتی این روش، از دست دادن سودمندی داده‌ها به دلیل این که اطلاعات ارتباط دقیق در دسترس نیست. تعمیم به معنای این است که داده را در سطح بالاتری از عمومیت نمایش دهیم به این شکل مقادیر کمتری سودمندی داده از دست می‌رود. به طور مثال دکتر، داروساز و پرستار همگی مصادیق خاص مشاغل درمانی هستند. می‌توانیم به جای هر یک از آنها از عنوان مشاغل درمانی استفاده کنیم. یک خوشه‌بندی چندگروهی بکار گرفته شده به یک اندازه بر روی داده‌های تقسیم شده به صورت عمودی اعمال می‌شود، در جایی که هر سایت داده، در خوشه‌بندی به طور مساوی مشارکت می‌کند. مطابق با مفهوم پایه، سایت‌های داده برای به رمز در آوردن مقدار با یک کلید عمومی مشترک در هر مرحله از خوشه‌بندی همکاری می‌کنند [۸].

تکنیک برون سپاری

این روش مسائلی را در رابطه با برون سپاری در داده‌کاوی و قوانین وابستگی برای حفاظت از حریم خصوصی توضیح داده است. مدل حمله بر اساس دانش قبلی برای حفاظت از حریم خصوصی در داده‌کاوی برون سپاری توسعه یافته است. یک طرح رمز دار مطرح شده است که بر اساس جایگزینی یک به یک رمزگذاری آیتم‌ها است که شامل انتقال جعلی است تا هر آیتم رمزدار را که با تکرار k-1 نسبت به بقیه است را به اشتراک بگذارد. خلاصه نقل وانتقالات جعلی برای پشتیبانی درست از الگوهای بازیابی شده، استفاده شده که از طریق آنها سرور می‌تواند بطور کارآمدی پوشش داده شود. این نشان می‌دهد که طرح ارائه شده در مقابل حملات قدرتمند است که بر پایه آیتم‌های واقعی و پشتیبانی دقیق است. این روش اینطور در نظر گرفته که حمله کننده از چنین اطلاعاتی ناآگاه است بنابراین هر گونه کاهش می‌تواند طرح رمزگذاری‌ها را بشکند و حریم خصوصی را آسیب برساند. راهبردهایی برای بهبود الگوریتم رمزدار بدست آمده تا تعداد الگوهای جعلی را به حداقل برسانیم. این روش مسائل مربوط به برون سپاری مکرر مجموعه آیتم‌ها را برای ساختار حفاظت از حریم خصوصی مورد بررسی قرار داده است. اولین راه حل امن برای برون سپاری داده‌کاوی اصول وابستگی با محرمانه بودن داده ها، حفظ حریم خصوصی داده کاوی و درست بودن آن را شخصی به نام لی مطرح کرده است. این روش قادر است تا متن ساده، متن رمزدار را آنچنان که خواسته شده با امنیت معنایی بدست آورد [۱۱].

در تحقیقی یک طرح رمزگذاری قابل جستجو برای برون سپاری تحلیل داده‌ها ارائه شده است. در این طرح مشتری باید داده‌ها را فقط یکبار به رمز در آورد و اطلاعات رمزدار شده را به تحلیلگر داده‌ها انتقال دهد. تحلیلگر دیتا سؤالاتی را برای مجوز موردنیاز از مشتری می‌پرسد تا محتوای داده در سؤالات را انتقال دهد. این طرح رمزگذاری مطرح شده اجازه جستجوی کلمات کلیدی و محدوده‌ای از سؤالات را داده است. تعدادی سؤال آزاد در حوزه تحقیق که قابلیت رمزگذاری را دارند در صورت تجزیه و تحلیل داده‌های برون سپاری، جالب است راندمان پیشرفت ممکن برای دامنه‌ای از سؤالات را با الزامات امنیتی لازم ازطریق رمزگذاری مبتنی بر جفت شدن ترکیب کنیم.

ورکو و همکارانش با کاهش عملیات فشرده محاسباتی مانند نگاشت دو خطی، راندمان طرح بالا را افزایش دادند. پس از تحلیلی دقیق بر عملکرد امنیت، این طرح نتایج امن و کارآمدی نشان داد. با این حال، درج بلوک داده باعث شد که طرح پیشنهادی، دینامیک (پویا) نباشد. بنابراین، ایجاد یک طرح کاملا دینامیک و حسابرسی عمومی امن، به عنوان یک چالش باز برای یک سیستم ابری باقی می‌ماند. در طی تحقیقی در سال ۲۰۱۴ به بررسی مسائل مربوط به برون سپاری مورد مجموعه آیتم‌های مکرر برای یک معماری حفظ حریم خصوصی بزرگ پرداختند. با در نظر گرفتن اینکه مهاجمان، از آیتم‌ها و پشتیبانی آیتم کاملاً آگاه هستند، یک مدل حمله معرفی شده است. علاوه بر این، حتی در احتمال، مهاجمان کاملاً از جزئیات الگوریتم رمزنگاری و برخی از جفت‌های آیتم با مقادیر رمز مربوطه آگاه هستند. این مفروضات پایه، بطور قابل ملاحظه‌ای باعث بهبود امنیت سیستم شدند و حمله مبتنی بر آیتم و مجموعه آیتم را حذف کردند و همچنین زمان پردازش را کاهش دادند [۱۲].

حفظ حریم خصوصی می‌تواند به چند طریق و در چند زمان و از جهات مختلف انجام شود. اگر اهمیت امنیت از کارایی داده‌کاوی بیشتر باشد، باید امر حفظ حریم خصوصی قبل از تحویل برای داده‌کاوی صورت بگیرد اما اگر کارایی داده‌کاوی از اهمیت بیشتری برخوردار باشد، سیاست‌های امنیتی در جهت حفظ حریم خصوصی باید به همراه داده کاوی انجام شود که شامل دستکاری در الگوریتم‌های داده‌کاوی

ISSN: 2745-3673
http://www.Rcsj.ir

برای محدود سازی داده‌کاوی و همچنین جلوگیری از خرابکاری از طرق مختلف انجام می‌شود که در این تحقیق به تعدادی از روش‌ها پرداخته شده است. با توجه به چالش‌های موجود در پژوهش‌های انجام شده اهیمت پیاده سازی حفظ حریم خصوصی درک شده است و در هر زمینه‌ای حفظ حریم خصوصی باید در نظر گرفته شود. حریم خصوصی یعنی یک فرد یا گروه بتواند خود و یا اطلاعات مربوط به خود را مجزا کند و در نتیجه بتواند خود و یا اطلاعاتش را با انتخاب خویش در برابر دیگران آشکار کند. مرزها و محتوای آنچه خصوصی قلمداد می‌شود در میان فرهنگ‌ها و اشخاص متفاوت است. ما در جدول (۱) مزایا و معایب روش‌های حفظ حریم خصوصی در داده‌کاوی را مشخص کردیم.

تکنیک اعوجاج

لی و همکاران یک روش اختلال ناشناس کم هزینه و کمتر مخاطره آمیز از طریق رمزگذاری همریخت و تبادل ناشناس ارائه دادند. روش پیشنهادی، برای پارامترهای بهینه، استحکام نشان داد. بابو و همکاران سه مدل شامل کلاینت‌ها، مراکز داده، و پایگاه داده در هر سایت معرفی کردند. مرکز داده کاملا مجهول است، به طوری که نقش کلاینت‌ها و پایگاه داده سایت قابل تعویض است. برنکویک و همکاران یک معماری شامل تکنیک‌های مختلف جدید ارائه شده است دادندکه تمام ویژگی‌ها در پایگاه داده را تحت تاثیر قرار دادند. یافته‌های تجربی نشان داد که معماری ارائه شده در حفظ الگوهای اصلی در یک مجموعه داده مختل و آشفته، بسیار مؤثر است. وانگ و لی یک روش برای جلوگیری از حملات استنتاج رو به جلو، در داده‌های پاکسازی شده (دلالت داده های اصلی) که توسط پاکسازی ایجاد شده‌اند، معرفی کردند [۱۰].

ایده شناسایی پویای ویژگی‌های حساس حفظ حریم خصوصی در داده‌کاوی در سال ۲۰۱۲ ارائه شد. شناسایی این ویژگی‌ها بستگی به حد آستانه حساسیت هر مشخصه دارد. مالک داده‌ها، با استفاده از روش مبادله برای محافظت از حریم خصوصی از اطلاعات حساس، مقدار را تحت ویژگی‌های حساس شناسایی شده اصلاح می‌کند. داده‌ها به شیوه‌ای اصلاح می‌شوند که خواص اصلی داده‌ها بدون تغییر باقی بماند. با وجود تازگی، از لحاظ زمانی، گران باقی می‌ماند. پس از آن، ژانگ و همکارانش به تازگی یک استراتژی تولید نویز مبتنی بر احتمال تاریخی پیشرفته به نام HPNGS معرفی کردند. نتایج شبیه‌سازی تایید کردند که HPNGS قادر است تعداد الزامات را در مکمل تصادفی-اش کاهش دهد. آنها بر روی حفاظت از حریم خصوصی و مبهم کردن نویز در محاسبات ابری تمرکز کردند. در نتیجه، یک استراتژی جدید تولید نویز مبتنی بر احتمال انجمنی (APNGS) ایجاد می‌شود. آنالیز تأیید کرد که APNGS پیشنهادی به طور قابل توجهی باعث بهبود حفاظت حریم خصوصی در ابهام نویز شامل احتمالات انجمن در هزینه اضافی معقول نسبت به استراتژی‌های استاندارد نماینده می‌شود لی و همکاران یک روش اختلال ناشناس کم هزینه و کمتر مخاطره آمیز از طریق رمزگذاری همریخت و تبادل ناشناس ارائه دادند. روش پیشنهادی، برای پارامترهای بهینه، استحکام نشان داد. بابو و همکاران سه مدل شامل کلاینت‌ها، مراکز داده، و پایگاه داده در هر سایت معرفی کردند. مرکز داده کاملا مجهول است، به طوری که نقش کلاینت‌ها و پایگاه داده سایت قابل تعویض است. برنکویک و همکاران یک معماری شامل تکنیک‌های مختلف جدید ارائه شده است دادندکه تمام ویژگی‌ها در پایگاه داده را تحت تاثیر قرار دادند. یافته‌های تجربی نشان داد که معماری ارائه شده در حفظ الگوهای اصلی در یک مجموعه داده مختل و آشفته، بسیار مؤثر است. وانگ و لی یک روش برای جلوگیری از حملات استنتاج رو به جلو، در داده‌های پاکسازی شده (دلالت داده های اصلی) که توسط پاکسازی ایجاد شده‌اند، معرفی کردند [۱۶].

بحث و نتیجه‌گیری

در حال حاضر، چندین تکنیک حفظ حریم خصوصی برای داده‌کاوی، موجود است که عبارتند از: گمنامی، طبقه‌بندی، خوشه بندی، قانون انجمنی، حفظ حریم خصوصی پراکنده، تنوع، تصادفی کردن، درخت طبقه‌بندی، چگالش، و رمزنگاری. روش‌های حفظ حریم خصوصی برای داده‌کاوی با تغییر داده‌ها برای پوشاندن و یا پاک کردن داده‌های حساس اصلی برای پنهان کردن آنها، از داده‌ها محافظت می‌کنند. به طور معمول، آنها براساس مفاهیم شکست حریم خصوصی، ظرفیت برای تعیین اطلاعات اصلی کاربر از اطلاعات اصلاح شده، از دست دادن اطلاعات و برآورد خسارت دقت داده‌ها می‌باشند ایده اصلی حفظ حریم خصوصی برای داده‌کاوی عبارت است از ترکیب تکنیک-های داده کاوی متداول در تبدیل داده‌ها با اطلاعات حساس به پوشش. چالش اصلی، تبدیل مؤثر داده‌ها و بازیابی نتیجه استخراج آن از داده‌های تبدیل یافته است. در نتیجه، سربار برای محاسبات داده کاوی سراسری، حفظ حریم خصوصی اطلاعات در حال رشد و کاربردپذیری داده‌ها در زمینه حفظ حریم خصوصی برای داده‌کاوی بررسی می‌شوند. در این راستا ما نقاط ضعف و قوت منابع و متون موجود را شناسایی و آنها را برای پیشرفت‌های بیشتر وحفاظت از حریم خصوصی تحلیل نمودیم. امید است که این پایان نامه به عنوان یک بررسی جامع و آموزنده جهت مرور و درک پیشرفت‌های تحقیقات جهت حفظ حریم خصوصی در داده‌کاوی عمل کند. روش‌های

ISSN: 2745-3673
http://www.Rcsj.ir

خوشه‌بندی بیشترین سهم را در بین روش‌های یادگیری ماشین در حفظ حریم خصوصی را دارند. روش‌های خوشه‌بندی به جای انتشار اطلاعات دقیق ، خلاصه‌ای از هر گروه را پست می‌کند و اطلاعات ارتباط دقیق در دسترس نیست. با توجه به یافته‌های این تحقیق، گمنامی یک روش مؤثر برای حفاظت از حریم خصوصی در داده کاوی است ولی اطلاعات پردازش شده توسط این روش اغلب مؤفق به غلبه بر برخی از حملات نمی‌شود و مستعد سوء استفاده اینترنتی می‌باشند.

حفظ حریم خصوصی می‌تواند به چند طریق و در چند زمان و از جهات مختلف انجام شود. اگر اهمیت امنیت از کارایی داده‌کاوی بیشتر باشد، باید امر حفظ حریم خصوصی قبل از تحویل برای داده کاوی صورت بگیرد اما اگر کارایی داده کاوی از اهمیت بیشتری برخوردار باشد، سیاست های امنیتی در جهت حفظ حریم خصوصی باید به همراه داده کاوی انجام شود که شامل دستکاری در الگوریتم‌های داده‌کاوی برای محدود سازی داده کاوی و همچنین جلوگیری از خرابکاری از طرق مختلف انجام می‌شود که در این تحقیق به تعدادی از تکنیک‌ها پرداخته شده است. با چالش‌های موجود در پژوهش‌های انجام شده اهمیت پیاده سازی حفظ حریم خصوصی درک شده است و در هر زمینه‌ای حفظ حریم خصوصی باید در نظر گرفته شود. حریم خصوصی یعنی یک فرد یا گروه بتواند خود و یا اطلاعات مربوط به خود را مجزا کند و در نتیجه بتواند خود و یا اطلاعاتش را با انتخاب خویش در برابر دیگران آشکار کند. مرزها و محتوای آنچه خصوصی قلمداد می‌شود در میان فرهنگ‌ها و اشخاص متفاوت است، محافظت از حریم خصوصی فرد در داده‌کاوی بسیار حائز اهمیت است.

صاحبان داده‌ها به علت ترس از افشای اطلاعات شخصی و محرمانه خود توسط دیگران، چندان تمایلی جهت انجام داده‌کاوی روی داده‌های خود نداشته ولی این مطلب را نیز می‌دانند که بدون انجام داده‌کاوی به نتایج و دانش مفید از داده‌های یکدیگر دسترسی پیدا نمی‌کنند. مالکان داده باید اطلاعات خود را برای مطالعه و تحقیق در اختیار پژوهشگران قرار دهند و اطلاعات داده شده نباید باعث نقض حریم خصوصی افراد شود به همین دلیل باید به دنبال تکنیک‌هایی باشیم تا به مالکان داده‌ها این اطمینان را بدهیم که امکان تبادل و انتشار داده‌ها وجود دارد و می‌توان با حفظ حریم خصوصی، داده‌ها را در اختیار پژوهشگران قرار داد. استفاده از اینترنت و فناوری اطلاعات در کنار مزایایی که دارد، خطر برملا شدن اسرار خصوصی را به همراه دارد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[۱] بلورکش، فهیمه، حفظ حریم مکانی برای انجام پرس‌وجوهای نزدیک‌ترین همسایه گروهی در خدمات مکان مبنا، پایان نامه کارشناسی ارشد مهندسی کامپیوتر، دانشگاه اصفهان، ایران. ۱۳۹۱.

[2] Chester et al, “Complexity of social network anonymization”, Journal of Social Network Analysis and Mining, pp.1-16, 2012.

[3] Kianmehr,"Privacy-Preserving Ranking over Vertically Partitioned Data", Computer Software and Applications Conference, 2012.

[4] Sweeney. "K-anonymity: a model for protecting privacy". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2012.

[5] Chun et al, Wensheng Gan, Philippe Fournier-Viger, Lu Yang, Qiankun Liu, Jaroslav Frnda, Lukas Sevcik, Miroslav Voznak,"High utility-itemset mining and privacy-preserving utility mining", Perspectives in Science, 2016.

[6] Belwal et al, "Hiding sensitive association rules efficiently by introducing new variable hiding counter". In: IEEE international conference on service operations and logistics, and informatics, 2008, IEEE/SOLI 2008, vol 1, pp 130–134. Doi: 10.1109/SOLI.2008.4686377.2013.

[7] Nayak et al, "A survey on privacy preserving data mining: approaches and techniques". Int J Eng Sci Tech 3(3): 2117–2133. 2011.

[8] Islam et al," Privacy preserving data mining: a noise addition framework using a novel clustering technique". Knowl Based Syst 24(8):1214–1223, 2011.

[9] Mukkamala et al, "Fuzzy-based methods for privacy-preserving data mining". In: IEEE eighth international conference on information technology: new generations (ITNG), 2011.

[10] Matwin, "Privacy-preserving data mining techniques: survey and challenges". In: Discrimination and privacy in the information society. Springer, Berlin, Heidelberg, pp 209–221. 2013.

[11] Sachan et al, "An analysis of privacy preservation techniques in data mining". In: Advances in computing and information technology, vol 3. Springer, pp 119–128, 2013.

[12] Vatsalan et al, "A taxonomy of privacy-preserving record linkage techniques". INF Syst 38(6):946–969, 2013.

[13] X. Qi and M. Zong, "An overview of privacy preserving data mining". Procedia Environ Sci 12(Icese 2011):1341–1347, 2012.

[14] Vijayarani et al, "a survey Privacy preserving data mining based on association rule". In: IEEE international conference on communication and computational intelligence (INCOCCI), 2010.

[15] Agrawal et al. Thomas," Privacy-preserving OLAP", International Conference on Management of Data,pp. 251–262, 2005.

[16] Zhang et al, "Privacy-Preserving OLAP: An Information-Theoretic Approach Knowledge and Data Engineering", IEEE, 2010.

[17] Cuzzocrea et al ."Further Theoretical Contributions to a Privacy Preserving Distributed OLAP Framework", Computer Software and Applications Conference (COMPSAC), 2013.

[18] Rolando et al,” On the privacy offered by (k, d)-anonymity”, Information Systems, VOl.38, pp.491–494, 2015.