دانشگاه فردوسی مشهد
دانشکده مهندسی - گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد

حفظ حریم خصوصی در کاوش سودمندی

نگارنده:

طاهر ره‌گوی

استاد راهنما:

دکتر رضا منصفی

بهار ۹۰

چکیده

یکی از مسائل مهم دنیای امروز استخراج دانش از پایگاه‌داده‌های بسیار بزرگ است. داده‌کاوی این امکان را فراهم کرده‌است که بتوان به‌صورت خودکار اطلاعات مفیدی را از پایگاه‌داده‌های بسیار بزرگ استخراج نمود. اطلاعات استخراج شده توسط داده‌کاوی ممکن است موجب نقض حریم خصوصی افراد و سازمان‌ها گردد. با افزایش موارد نقض حریم خصوصی توسط داده‌کاوی، نگرانی‌ها در میان شهروندان برای حریم خصوصی نیز افزایش روزافزون داشته است. لذا در سال‌های اخیر حفظ حریم خصوصی در داده‌کاوی به‌عنوان مبحثی حیاتی، مورد توجه جامعه علمی قرار گرفته است. یکی از روش‌های بسیار مهم در داده‌کاوی، کاوش قوانین انجمنی می‌باشد. مبحث حفظ حریم خصوصی در کاوش سودمندی که یکی از مدل‌های کاوش قوانین انجمنی است، اخیراً در محافل علمی مطرح گردیده و تاکنون دو الگوریتم مکاشفه‌ای برای آن ارائه شده است. اگرچه این الگوریتم‌ها به خوبی می‌توانند حریم خصوصی داده‌ها را حفظ کنند اما داده‌های تولید شده توسط این الگوریتم‌ها از کیفیت مناسبی برخوردار نیستند.

از این رو هدف اصلی این تحقیق بر روی طراحی و پیاده‌سازی روش‌ها و الگوریتم‌های جدید و کارآمد برای حفظ حریم خصوصی در کاوش سودمندی تمرکز دارد. مسأله یافتن راه‌حل بهینه برای پاک‌سازی پایگاه‌داده از مجموعه-اقلام حساس، یک مسأله برنامه‌ریزی غیرخطی عدد صحیح است و راه‌حل کارآمد ریاضی برای حل آن وجود ندارد. روش‌های پیشنهادی برای حل مسأله فوق در سه دسته‌ی مکاشفه‌ای، تکاملی و دقیق ارائه شده‌اند. الگوریتم‌های مکاشفه‌ای بر اساس قوانین شهودیِ کلی می‌کوشند که فرآیند پاک‌سازی را با وارد نمودن کم‌ترین آسیب به کیفیت داده‌ها، به اتمام برسانند. الگوریتم‌های تکاملی پیشنهادی تلاش می‌کنند که با بهینه‌سازی معیارهای کیفیت داده عملیات پاک‌سازی را انجام دهند. در نهایت الگوریتم‌های دقیق، مدل ساده شده‌ای از مسأله فوق که با روش‌های ریاضی قابل حل باشد را به-عنوان تقریبی از مسأله اصلی در نظر گرفته و با حل آن امید دارند که عملیات پاک‌سازی را با تولید داده‌های با کیفیت بالا به انجام برسانند. نتایج آزمایشات ارائه شده در این تحقیق، برتری چشم‌گیرِ روش‌های پیشنهادی بر روش‌های موجود کنونی را نشان می‌دهد.

کلمات کلیدی: حفظ حریم خصوصی، کاوش سودمندی، کاوش قوانین انجمنی، داده‌کاوی، برنامه‌ریزی عدد صحیح، الگوریتم‌های مکاشفه‌ای، الگوریتم‌های تکاملی

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل ۱

مقدمه

پیشرفت‌های شایانی که در دهه‌های پیشین در زمینه سخت‌افزار و ذخیره و بازیابی اطلاعات صورت گرفته، بستری فراهم آورده‌است که بتوان حجم عظیمی از اطلاعات را با هزینه‌ای بسیار کم ذخیره و نگهداری کرد. بر همین اساس سازمان‌ها و شرکت‌ها توانسته‌اند حجم بسیار زیادی از تراکنش‌های مربوط به اطلاعات شخصی افراد و شرکت‌های دیگر را در انبارداده‌های[۱] خود ذخیره‌سازی و نگهداری کنند تا بتوانند از این اطلاعات در جهت تصمیم‌گیری، برنامه‌ریزی، ارائه خدمات بهتر و کسب سود بیش‌تر استفاده نمایند.

۱-۱- داده‌کاوی[۲]

حجم انبوه داده‌های جمع‌آوری شده، چالش جدیدی برای پردازش و استفاده مؤثر از داده‌ها پیش روی سازمان‌ها و مؤسسات می‌گذارد که از این چالش با نام «انفجار اطلاعات[۳]» یاد می‌شود. الگوریتم‌ها و روش‌های بسیار زیادی توسط محققین برای حل این معضل ارائه گردیده و بسیاری از آن‌ها با موفقیت در صنعت و تجـارت به کارگرفته شده‌اند. الگوریتم‌ها و روش‌های فوق در محافل علمی به‌عنوان شاخه‌ای جدید به نام «داده‌کاوی» شناخته می‌شوند.

روش‌های داده‌کاوی همانند طبقه‌بندی[۴]، خوشه‌بندی[۵]، کاوش قوانین انجمنی[۶] و ...توانسته‌اند به خوبی الگوها و نظام‌های موجود در داده‌های بسیار عظیم را استخراج کرده و آن‌ها را به‌صورت مدل‌ها و قوانین کلی بیان نمایند و به این ترتیب حجم عظیم داده‌ها را به مجموعه‌ای از مدل‌ها و قوانین کلی خلاصه کنند. قوانین و مدل‌های استخراج شده در اختیار مدیران و تصمیم‌گیرندگان قرار می‌گیرد تا بتوانند بر اساس آن‌ها برنامه‌ریزی و تصمیم‌سازی بهتری داشته باشند. روش‌های داده‌کاوی به مؤسسات و شرکت‌های تجاری کمک کرده است تا بتوانند خدمات مؤثرتری به مشتریان خود ارائه

[1] *Data Warehouse*
[2] *Data Mining*
[3] *Information Overload*
[4] *Classification*
[5] *Clustering*
[6] *Association Rule Mining*

نموده و در نتیجه سوددهی بیشتری داشته باشند. استفاده از روش‌های داده‌کاوی تنها به تجارت محدود نمی‌شود؛ برای مثال هر دو کاندیدای اصلی در انتخابات ریاست جمهوری آمریکا در سال ۲۰۰۴ از داده‌کاوی جهت ارزیابی و پیش‌بینی آرای خود استفاده نمودند[۳۰]. همچنین داده‌کاوی به صورت وسیع در بخش بهداشت و سلامت[۶۷]، بانک‌داری[۷۸]، بیمه[۶۵]، مخابرات و ارتباطات[۴۳]، کنترل کیفیت[۱۷]و ... نیز به کار برده‌شده‌است.

۱-۱-۱- کاوش قوانین انجمنی

یکی از روش‌های بسیار مهم در داده‌کاوی، کاوش قوانین انجمنی[۱](ARM)[۵]است.کاوش قوانین انجمنی یک فرآیند دو مرحله‌ای است. در مرحله نخست مجموعه-اقلامِ متکرر[۲] کشف و استخراج می شوند و در مرحله دوم با استفاده از این مجموعه-اقلام، قوانین انجمنی تولید می‌شوند. تاکنون الگوریتم‌های بسیاری برای یافتن مجموعه-اقلامِ متکرر توسط محققان ارائه شده است. الگوریتم *Apriori*[۷,۹] معروف‌ترین الگوریتم در این زمینه است. یکی از مدل‌های استفاده شده در کاوش قوانین انجمنی، مدل کاوش سودمندی[۳][۹]است. کاوش سودمندی کمبودهای مدل سنتی که سوددهی و تعداد کالاها را در تراکنش‌ها در نظر نمی‌گرفت را برطرف می‌سازد و قوانین مفیدتری را تولید می‌کند.

۱-۲- حفظ حریم خصوصی در داده‌کاوی[۴]

بسیاری از شرکت‌ها برای دستیابی به سود بیش‌تر و همچنین ارائه خدمات بهتر به کاربران خود، بخشی از اطلاعات مربوط به کاربران را با شرکت‌های همتای خود به اشتراک می‌گذارند. برخی مؤسسات دولتی نیز اطلاعات استخراج شده همانند آمارگیری‌ها و اطلاعات مربوط به بهداشت و ... را

[1] *Association Rule Mining, ARM*
[2] *Frequent Itemsets*
[3] *Utility Mining*
[4] *Privacy Preserving Data Mining, PPDM*

به‌صورت عمومی منتشر می‌کنند تا متخصصین و محققین بتوانند این اطلاعات را تحلیل و ارزیابی نمایند و سازمان‌ها و مؤسسات دیگرنیز بتوانند از این اطلاعات برای تصمیم‌سازی و برنامه‌ریزی بهتر استفاده کنند. انتشار این نوع داده‌ها به صورت عمومی اگرچه اطلاعات خصوصی افراد و مؤسسات را به صورت مستقیم افشا نمی‌کند، اما این امکان را فراهم می‌سازد که حمله‌کننده با تحلیل دقیق این داده‌ها و ترکیب آن‌ها با داده‌های عمومی دیگر، بتواند به الگوها و استنتاج‌هایی دستیابی پیدا کند که حریم خصوصی[1] افراد و مؤسسات را به خطر اندازد.

افزایش نگرانی‌ها در بین شهروندان و بسیاری از مؤسسات برای به خطر افتادن حریم خصوصی آن‌ها موجب گردید که دولت‌ها ادامه استفاده از روش‌های داده‌کاوی را از لحاظ قانونی مورد بازبینی قرار دهند و در مواردی استفاده از روش‌های داده‌کاوی را ناقض قوانین حفظ حریم خصوصی دانسته و آن‌ها را غیرقانونی معرفی کنند. افزایش نگرانی‌ها از یک طرف و نیاز مبرم مؤسسات و دولت‌ها برای استفاده از روش‌های داده‌کاوی از طرف دیگر موجب گردید که شاخه جدیدی از تحقیقات با نام حفظ حریم خصوصی در داده‌کاوی(*PPDM*) به‌عنوان شاخه‌ای مهم و حیاتی برای ادامه امکان استفاده از روش‌های داده‌کاوی، مورد توجه محافل علمی قرار گیرد. این مبحث نخستین بار توسط *Agrawal* و *Srikant* در سال ۲۰۰۰ به محافل علمی معرفی شد[۶] و تا کنون روش‌های زیادی برای آن ابداع گردیده و مورد استفاده قرار گرفته است.

۱-۲-۱- شاخه‌های تحقیقاتی در *PPDM*

شاخه‌های اصلی تحقیقات در حفظ حریم خصوصی در داده‌کاوی عبارتند از[۱]:

[1] *Privacy*

- **حفظ حریم خصوصی در انتشار داده‌ها**[1]:در این شاخه روش‌های مختلف تبدیل، برای تغییر داده‌ها جهت حفظ حریم خصوصی مورد بررسی قرار گرفته است. این روش‌ها شامل روش‌های تصادفی سازی[2] [۶]، گمنامی درجه-k[3] [۱۱,۱۲] و تنوع درجه-l[4] [۴۷] می‌باشد.
- **تغییر نتایج الگوریتم‌های داده‌کاوی برای حفظ حریم خصوصی:** در بسیاری از موارد نتیجه حاصل از الگوریتم‌های داده‌کاوی مانند کاوش قوانین انجمنی و طبقه‌بندی موجب به خطر افتادن حریم خصوصی می‌شود و نه خود داده‌ها. این امر شاخه‌ای جدید را در *PPDM* به خود اختصاص داده است که هدف آن تغییر الگوریتم‌های داده‌کاوی است به نحوی که نتایج آن‌ها حریم خصوصی را نقض نکند. به‌عنوان مثال در مخفی‌سازی قوانین انجمنی، قوانینی که حریم خصوصی را به خطر می‌اندازند حذف می‌گردند.
- **بازبینی پرسش‌ها**[5]:این شاخه از روش‌ها مشابه روش‌های شاخه قبلی هستند با این تفاوت که در این‌جا نتایج پرسش‌ها محدود یا دست‌کاری می‌شوند. روش‌های آشفته-سازی[6] نتایج پرسش‌ها در[۱۴] و روش‌های محدودسازی پرسش‌ها در[۱۵,۱۶]بررسی شده‌اند.
- **روش‌های مبتنی بر رمزنگاری برای حفظ حریم خصوصی در سیستم‌های توزیع-شده:**در بسیاری از حالات ممکن است که اطلاعات به صورت توزیع‌شده بر روی چندین سایت قرار گرفته باشد. مالکان این داده‌ها ممکن است بخواهند که یک تابع مشترک[7] را بر روی همه داده‌ها اعمال کنند، به‌طوری‌که اطلاعات هر سایت برای دیگری افشا نشود و تنها حاصل تابع برای همه آن‌ها قابل مشاهده باشد. برای انجام این کار روش‌های مختلف

[1] *Privacy Preserving Data Publishing, PPDP*
[2] *Randomization*
[3] *k-Anonymity*
[4] *l-Diversity*
[5] *Query Auditing*
[6] *Perturbing*
[7] *Common Function*

رمزنگاری به کار رفته است، نمونه‌هایی از کارهای انجام شده در این زمینه را می‌توان در[۶۰] یافت.

- **چالش‌های نظری در مسائل با ابعاد بالا(مشکل افزایش ابعاد[1])**: مجموعه داده‌های واقعی معمولاً دارای ابعاد بسیار بالایی هستند، این امر الگوریتم‌های *PPDM* را چه از لحاظ زمان محاسباتی و چه از لحاظ کارایی با مشکلاتی مواجه کرده است. برای مثال در[۵۰] نشان داده شده است که یافتن جواب بهینه برای مسئله گمنامی درجه-k ، یک مسئله *NP-Hard* است. علاوه براین، روش‌ها و الگوریتم‌ها هنگامی‌که ابعاددادەها خیلی بزرگ باشد، کارایی مناسبی از خود نشان نمی‌دهند، زیرا در این حالات ترکیب ابعاد مختلف با داده‌های عمومی دیگر ممکن است سبب کشف اطلاعات بیش‌تری شود. مجموعه‌ای از حمله‌ها به مجموعه داده‌های با ابعاد بالا را می‌توانید در[۱۹, ۲۰]بیابید.

۱-۲-۲- الگوریتم‌ها و روش‌های *PPDM*

در این بخش برخی از روش‌ها و الگوریتم‌هایی که در هر کدام از شاخه‌های *PPDM* استفاده گردیده-است رامعرفی می‌کنیم[۲].

- **تصادفی‌سازی[2]**: روش تصادفی‌سازی به صورت سنتی برای حفظ حریم خصوصی در انتشار داده‌هایی مانند نظرسنجی‌ها استفاده می‌شود. در این روش با استفاده از یک تابع توزیع احتمال، به داده‌ها نویز اضافه می‌گردد و به این ترتیب احتمال دسترسی به اطلاعات حساس کاهش می‌یابد. این روش در[۶] معرفی گردیده است. کمی‌سازی[3] حریم خصوصی در این روشدر[۶,۱۳,۲۰,۲۲]مورد بررسی قرار گرفته است. انواع حمله‌ها به روش تصادفی-

[1] *Curse of Dimensionality*
[2] *Randomization*
[3] *Privacy Quantification*

سـازی نیز در[۱-۳] مورد بحث قرار گرفته‌اند. تصادفی‌سازی جریان داده‌ها[1] در[۲۱,۲۸] ودرهم‌سازی ضربی[2] در کارهای*[۱۶,۳۵,۴۴,۵۳,۵۵,۵۸]* ارائه شده است. تعویض داده[3]در[۲۳] و دیگر روش‌های تصادفی‌سازی در[۷,۲۱,۶۶,۸۰] مورد بحث قرار گرفته‌اند.

- **گمنام‌سازی گروهی**[4]:روش تصادفی‌سازی، روشی بسیار ساده است که به راحتی در زمان جمع‌آوری داده‌ها قابل پیاده‌سازی است زیرا نویز اضافه شده به هر رکورد مستقل از نویز اضافه شده به دیگر رکوردها است. یکی از ضعف‌های روش تصادفی‌سازی این است که احتمال استفاده از داده‌های خارجی و ترکیب آن‌ها با داده‌های موجود را در نظر نمی‌گیرد. در[۴] نشان داده شده‌است که استفاده از رکوردهایی که به صورت عمومی در دسترس هستند و ترکیب آن‌ها با داده‌هایی که به روش تصادفی‌سازی پاک‌سازی شده‌اند، می‌تواند در موارد با ابعاد بالا به شدت حریم خصوصی را به خطر بیاندازد. این امر در حالتی که رکورد مورد نظر به راحتی از رکوردهای محلی قابل تمییز باشد، شدیدتر است. به همین دلیل در بسیاری از رهیافت‌ها برای حفظ حریم خصوصی تلاش شده‌است که گروه‌هایی گمنام از رکوردها ساخته شود تا حمله کننده نتواند به رکورد مورد نظر با احتمال بالا دستیابی پیدا کند. روش گمنامی درجه-k[۷۰] ، تنوع درجه-l[۴۷] و نزدیکی درجه-t[5][۳۸] در این دسته قرار می‌گیرند.
- **حفظ حریم خصوصی توزیع‌شده:**هدف اصلی در بسیاری از الگوریتم‌های حفظ حریم خصوصی توزیع‌شده محاسبه توابع تجمعی آماری بر روی همه داده‌ها است به نحـوی که داده‌های یک شریک برای شرکای دیگر افشا نشود. در این حالت شرکا علاقه‌مند هستند که در محاسبـه توابع تجمعی آمـاری بر روی همه داده‌ها همکاری کنند اما به‌طور کامل بـه هم‌دیـگر اعتـماد ندارند. داده‌های بـه اشتراک گذاشته‌شده می‌تواند دو حالت داشته باشد:۱)

[1] *Data Streams*
[2] *Multiplicative Perturbation*
[3] *Data Swapping*
[4] *Group Based Anonymization*
[5] *t-Closeness*

داده‌ها به صورت افقی[۱]۲) به صورت عمودی[۲] افراز شده باشند. در افراز افقی، رکوردها بین شرکای مختلف پخش شده‌است؛ همه شرکا دارای رکوردهایی با خصیصه‌های مشابه هستند اما هر یک از شرکا بخشی از رکوردها را دارد. در افراز عمودی، خصیصه‌ها در بین شرکا پخش شده است و هریک از شرکا بخشی از خصیصه‌های هر کدام از رکوردها را در اختیار دارند. تعداد رکوردها برای همه شرکا یکسان است. اکثر روش‌های استفاده شده در این زمینه مبتنی بر رمزنگاری می‌باشند. برای اطلاعات بیشتر به[۶۰] مراجعه کنید.

- **حفظ حریم خصوصی در نتایج الگوریتم‌های داده‌کاوی:** در بسیاری از موارد، خروجی یک الگوریتم می‌تواند توسط دشمن برای استنتاج رفتار و خصوصیات داده‌های اصلی مورد استفاده قرار گیرد. به همین دلیل بسیاری از تلاش‌ها و تحقیقات صرف محدودسازی الگوریتم‌های داده‌کاوی به منظور حفظ حریم خصوصی شده‌است. از روش‌هایی که در این دسته قرار می‌گیرند می‌توان از مخفی‌سازی قوانین انجمنی[۳][۱۱]، کاهش کارایی و دقت طبقه‌بندها[۴][۱۵,۵۱]و بازبینی پرسش‌ها و کنترل استنتاج‌ها[۵][۱۴,۱۹]نام برد.

۱-۳- مخفی‌سازی قوانین انجمنی

عبارت «مخفی‌سازی قوانین انجمنی» برای نخستین بار در سال ۱۹۹۹ در[۱۱] مطرح گردید. نویسندگان این مقاله مسأله تغییر یک پایگاه‌داده به نحوی که برخی از قوانین انجمنی(قوانین حساس) قابل استخراج نباشند را مورد بررسی قرار دادند. آن‌ها تلاش کردند که تغییرات به نحوی باشد که کم‌ترین تأثیر را بر روی قوانین مجاز داشته باشد. آن‌ها راه‌حل‌های مختلفی از قبیل فازی‌سازی پایگاه‌داده، محدود نمودن دسترسی به داده‌ها و همچنین انتشار نمونه‌هایی از داده‌ها به جای انتشار

[1] *Horizontally Partitioned Data*
[2] *Vertically Partitioned Data*
[3] *Association Rule Hiding*
[4] *Downgrading Classifier Effectiveness*
[5] *Query Auditing and Inference Control*

کل داده‌ها را پیشنهاد کردند. به دلیل پیچیدگی زیاد این مسئله، راه‌حل‌های ارائه شده در آن مقاله و بسیاری از راه‌حل‌هایی که بعداً ارائه شدند به صورت مکاشفه‌ای بودند[۱]. راه‌حل‌های مکاشفه‌ای اگرچه به لحاظ سرعت و پیچیدگی محاسباتی[۲] و مقیاس‌پذیری[۳] از کارایی بسیار خوبی برخوردار هستند اما از مشکل نقطه بهینه محلی[۴] رنج می‌برند. بر همین اساس اخیراً الگوریتم‌های دقیقی برای مخفی‌سازی قوانین انجمنی معرفی شده است. این الگوریتم‌ها به ۳ دسته اساسی تقسیم‌بندی می‌شوند[۲۵]: ۱)الگوریتم‌های مکاشفه‌ای، ۲) الگوریتم‌های مبتنی بر مرز[۵]۳)الگوریتم‌های دقیق[۶].

۱-۴- حفظ حریم خصوصی در کاوش سودمندی[۷]

اگرچه در سال‌های اخیر مبحث مخفی‌سازی قوانین انجمنی برای مدل سنتی(مدل کاوش اقلامِ متکرر[۸]) کاوش قوانین انجمنی[۵] به خوبی مورد بررسی قرار گرفته است اما حفظ حریم خصوصی در مدل‌های جدیدترکاوش قوانین انجمنی مانند مدل کاوش مجموعه-اقلامِ سهم-متکرر[۹][۴۰] و مدل کاوش سودمندی[۱۰][۴۶] به دلیل پیچیدگی بیشتر نسبت به مدل سنتی، کم‌تر مورد توجه قرار گرفته‌اند. در سال ۲۰۱۰ مبحث حفظ حریم خصوصی در کاوش سودمندی(*PPUM*) توسط *Yeh* و *Hsu* مطرح گردید[۷۹]. آنها مسأله مطرح در *PPUM* را به صورت رسمی بیان نموده و برای آن دو الگوریتم مکاشفه‌ای جدید معرفی کردند.

[1] *Heuristic*
[2] *Computational Complexity*
[3] *Scalability*
[4] *Local Optima*
[5] *Border Based Algorithms*
[6] *Exact Algorithms*
[7] *Privacy Preserving Utility Mining, PPUM*
[8] *Frequent Itemset Mining Model*
[9] *Share Frequent Itemset Mining Model*
[10] *Utility Mining Model*

۱–۵– نمای کلی پایان‌نامه

در فصل اول مقدمه‌ای بر حفظ حریم خصوصی در کاوش سودمندی ارائه شد و دسته‌بندی‌ها و روش-های مختلف آن به‌طور مختصر معرفی گردید. تحلیل ادبیات در قالب فصل‌های دوم و سوم ارائه شده است. در فصل دوم روش‌ها و الگوریتم‌های ارائه شده برای مخفی‌سازی قوانین انجمنی مورد بررسی قرار گرفته‌اند و دسته‌بندی‌های مختلف این روش‌ها ارائه گردیده است. در فصل سوم مسأله حفظ حریم خصوصی در کاوش سودمندی به صورت رسمی بیان شده‌است و دو الگوریتم موجود برای حل این مسأله بررسی شده است. در فصل چهارم به‌طور مفصل روش‌های پیشنهادی و همچنین نتایج ارزیابی هر کدام از آن‌ها ارائه شده است. در فصل پنجم نیز نتیجه‌گیری و توصیه‌های آتی بیان شده است.