



دانشگاه فردوسی مشهد
دانشکده مهندسی - گروه مهندسی کامپیوتر

پایان نامه 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 داشته‌است. لذا در سال‌های اخیر حفظ حریم خصوصی در داده‌کاوی به‌عنوان مبحثی حیاتی، مورد توجه جامعه علمی قرار گرفته‌است. 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 شده‌است. 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 شهودی 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 چشم‌گیر 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

ii.....	چکیده
iii.....	فهرست مطالب
vii.....	فهرست شکل‌ها
ix.....	فهرست جداول
۱۱.....	فصل ۱- مقدمه
۱۲.....	۱-۱- داده‌کاوی
۱۳.....	۱-۱-۱- کاوش قوانین انجمنی
۱۳.....	۱-۲- حفظ حریم خصوصی در داده‌کاوی
۱۴.....	۱-۲-۱- شاخه‌های تحقیقاتی در <i>PPDM</i>
۱۶.....	۲-۲-۱- الگوریتم‌ها و روش‌های <i>PPDM</i>
۱۸.....	۳-۱- مخفی‌سازی قوانین انجمنی
۱۹.....	۴-۱- حفظ حریم خصوصی در کاوش سودمندی
۲۰.....	۵-۱- نمای کلی پایان‌نامه
۲۱.....	فصل ۲- مخفی‌سازی قوانین انجمنی
۲۲.....	۱-۲- پیش‌نیاز
۲۳.....	۱-۱-۲- مقدمات و تعاریف
۲۵.....	۲-۱-۲- تئوری مرزها
۲۶.....	۳-۱-۲- تعریف رسمی مسأله
۲۶.....	۱-۳-۱-۲- اهداف متدولوژی‌های مخفی‌سازی قوانین انجمنی
۲۸.....	۲-۳-۱-۲- بیان مسأله
۲۹.....	۳-۳-۱-۲- گونه ۱: مخفی‌سازی قوانین انجمنی حساس
۲۹.....	۴-۳-۱-۲- گونه ۲: مخفی‌سازی مجموعه-اقلام حساس

۳۰	۲-۲- دسته‌بندی متدولوژی‌های مخفی‌سازی قوانین انجمنی
۳۱	۲-۳- الگوریتم‌های مکاشفه‌ای
۳۲	۲-۳-۱- الگوریتم‌های مبتنی بر درهم‌سازی
۳۴	۲-۳-۲- الگوریتم‌های مبتنی بر مسدودسازی
۳۵	۲-۴- الگوریتم‌های مبتنی بر اصلاح مرز
۳۵	۲-۴-۲- الگوریتم <i>BBA</i>
۳۶	۲-۴-۳- الگوریتم <i>Max-Min</i>
۳۶	۲-۵- الگوریتم‌های دقیق
۳۷	۲-۵-۱- الگوریتم <i>Menon</i>
۳۹	۲-۶- خلاصه
۴۰	فصل ۳- حفظ حریم خصوصی در کاوش سودمندی
۴۱	۳-۱- کاوش سودمندی
۴۴	۳-۲- حفظ حریم خصوصی در کاوش سودمندی (<i>PPUM</i>)
۴۵	۳-۲-۱- فرآیند پاک‌سازی
۴۶	۳-۲-۲- محاسبه کارایی
۴۷	۳-۲-۳- الگوریتم‌های <i>PPUM</i>
۴۷	۳-۲-۳-۱- الگوریتم مخفی‌سازی قلم با سودمندی بالا، اول (<i>HHUIF</i>)
	۳-۲-۳-۲- الگوریتم مخفی‌سازی قلم با بیش‌ترین تداخل با مجموعه-اقلام حساس، اول (<i>MSICF</i>)
۴۹	اول (<i>MSICF</i>)
۵۰	۳-۳- خلاصه
۵۱	فصل ۴- روش‌های پیشنهادی
۵۶	۴-۱- مشکلات و چالش‌های موجود در حفظ حریم خصوصی در کاوش سودمندی
۵۹	۴-۲- روش‌های پیشنهادی

- ۶۰ ۱-۲-۴- رهیافت مکاشفه‌ای
- ۶۱ ۱-۱-۲-۴- الگوریتم مخفی‌سازی تراکنشِ بابیش‌ترین تداخل با مجموعه-اقلام حساس،
اول (HTMSCF) ۶۲
- ۶۲ ۲-۱-۲-۴- الگوریتم مخفی‌سازی تراکنش با کم‌ترین تداخل با مجموعه-اقلام مجاز،
اول (HTMLCF) ۶۴
- ۶۳ ۳-۱-۲-۴- الگوریتم مخفی‌سازی تراکنشِ بابیش‌ترین تداخل با مجموعه-اقلام حساس
و کم‌ترین تداخل با مجموعه-اقلام مجاز، اول (HTMMCf) ۶۵
- ۶۴ ۴-۱-۲-۴- الگوریتم مخفی‌سازی قلمِ با کم‌ترین تداخل با مجموعه-اقلام مجاز،
اول (HIMLCF) ۶۶
- ۶۵ ۵-۱-۲-۴- الگوریتم مخفی‌سازی قلمِ بابیش‌ترین تداخل با مجموعه-اقلام حساس و کم‌ترین
تداخل با مجموعه-اقلام مجاز، اول (HIMMCF) ۶۷
- ۶۶ ۶-۱-۲-۵- جمع‌بندی الگوریتم‌های مکاشفه‌ای ۶۷
- ۶۷ ۲-۲-۴- رهیافت تکاملی ۶۹
- ۶۸ ۳-۲-۴- رهیافت دقیق ۷۱
- ۶۹ ۳-۴- پیچیدگی زمانی الگوریتم‌های پیشنهادی ۷۶
- ۷۰ ۱-۳-۴- افزایش سرعت جستجو با استفاده از فایل وارونه ۷۷
- ۷۱ ۲-۳-۴- پیچیدگی زمانی بخش‌های مختلف الگوریتم‌ها ۷۷
- ۷۲ ۴-۲-۴- پیچیدگی زمانی الگوریتم‌های مکاشفه‌ای ۷۸
- ۷۳ ۳-۴- پیاده‌سازی روش‌های پیشنهادی ۸۰
- ۷۴ ۴-۴- مجموعه داده‌ها ۸۱
- ۷۵ ۵-۴- معیارهای ارزیابی ۸۳
- ۷۶ ۶-۴- ارزیابی و بررسی نتایج روش‌های پیشنهادی ۸۵
- ۷۷ ۱-۶-۴- بهبود تابع ارزیابی الگوریتم ژنتیک ۸۵
- ۷۸ ۲-۶-۴- تعیین مقادیر پارمترهای الگوریتم ژنتیک ۸۶

۹۰ تعیین مقادیر حداقل آستانه سودمندی مناسب جهت ارزیابی الگوریتم‌ها
۹۲ بررسی الگوریتم‌ها از نظر <i>HF</i>
۹۲ بررسی الگوریتم‌ها از نظر <i>MC</i>
۹۲ پایگاه داده‌های کوچک
۹۵ پایگاه داده‌های بزرگ
۹۷ بررسی الگوریتم‌ها از نظر <i>DBDR</i>
۹۷ پایگاه داده‌های کوچک
۱۰۰ پایگاه داده‌های بزرگ
۱۰۱ بررسی الگوریتم‌ها از نظر <i>TA</i>
۱۰۱ پایگاه داده‌های کوچک
۱۰۴ پایگاه داده‌های بزرگ
۱۰۵ بررسی الگوریتم‌ها از دیدگاه زمان اجرا
۱۰۵ پایگاه داده‌های کوچک
۱۰۸ پایگاه داده‌های بزرگ
۱۱۰ بررسی کلی الگوریتم‌ها
۱۱۲ خلاصه
۱۱۴	فصل ۵- نتیجه‌گیری و توصیه‌های آتی
۱۱۷ ۱-۵- کارهای آتی
۱۱۷ ۱-۱-۵- استفاده از روش مسدودسازی به جای درهم‌سازی در الگوریتم‌های پیشنهادی
۱۱۸ ۲-۱-۵- افزاز پایگاه داده به زیرمجموعه‌های مستقل
۱۱۹	مراجع

فهرست شکل‌ها

- شکل ۱-۲: پایگاه داده D به همراه مجموعه-اقدام وقوانین انجمنی آن ۲۵
- شکل ۲-۲: مثالی از دوشبکه برای پایگاه داده‌هایی با $I=\{a,b,c\}$ (a) و $I=\{a,b,c,d\}$ (b) ۲۶
- شکل ۳-۲: طبقه‌بندی رهیافت‌های مخفی‌سازی قوانین انجمنی از چهار دیدگاه مختلف ۳۰
- شکل ۱-۳: فرآیند پاک‌سازی در $PPUM$ ۴۶
- شکل ۱-۴: شبکه مجموعه-اقدام پایگاه داده مثال ۱ ۵۳
- شکل ۲-۴: دسته‌بندی الگوریتم‌های مکاشفه‌ای ۶۹
- شکل ۳-۴: فرآیند پاک‌سازی در الگوریتم $PPUM_GA$ ۷۱
- شکل ۴-۴: درصد کاهش به صفر در پایگاه داده آزمایشی ۷۲
- شکل ۵-۴: فرآیند پاک‌سازی در رهیافت دقیق ۷۶
- شکل ۶-۴: توابع توزیع $lognormal$ با $\delta = 1$ و $\mu = 0$ ۸۱
- شکل ۷-۴: نمودار تغییرات MC برای مقادیر مختلف نرخ جهش و تقاطع ۸۷
- شکل ۸-۴: نمودار تغییرات $DBDR$ برای مقادیر مختلف نرخ جهش و تقاطع ۸۷
- شکل ۹-۴: نمودار تغییرات ترکیب MC و $DBDR$ برای مقادیر مختلف نرخ جهش و تقاطع ۸۸
- شکل ۱۰-۴: نمودار تغییرات MC برای مقادیر مختلف نرخ جهش و تقاطع ۸۸
- شکل ۱۱-۴: نمودار تغییرات $DBDR$ برای مقادیر مختلف نرخ جهش و تقاطع ۸۹
- شکل ۱۲-۴: نمودار تغییرات ترکیب MC و $DBDR$ برای مقادیر مختلف نرخ جهش و تقاطع ۸۹
- شکل ۱۳-۴: نمودار مقایسه الگوریتم‌ها از نظر MC در پایگاه داده $db1$ ۹۳
- شکل ۱۴-۴: نمودار مقایسه الگوریتم‌ها از نظر MC در پایگاه داده $db2$ ۹۳
- شکل ۱۵-۴: نمودار مقایسه الگوریتم‌ها از نظر MC در پایگاه داده $db3$ ۹۴
- شکل ۱۶-۴: نمودار مقایسه الگوریتم‌ها از نظر MC در پایگاه داده $db4$ ۹۴
- شکل ۱۷-۴: نمودار مقایسه الگوریتم‌ها از نظر MC در پایگاه داده $db5$ ۹۴
- شکل ۱۸-۴: نمودار مقایسه الگوریتم‌ها از نظر MC در پایگاه داده $db6$ ۹۶
- شکل ۱۹-۴: نمودار مقایسه الگوریتم‌ها از نظر MC در پایگاه داده $db7$ ۹۶
- شکل ۲۰-۴: نمودار مقایسه الگوریتم‌ها از نظر $DBDR$ در پایگاه داده $db1$ ۹۷
- شکل ۲۱-۴: نمودار مقایسه الگوریتم‌ها از نظر $DBDR$ در پایگاه داده $db2$ ۹۷
- شکل ۲۲-۴: نمودار مقایسه الگوریتم‌ها از نظر $DBDR$ در پایگاه داده $db3$ ۹۸

- شکل ۴-۲۳: نمودار مقایسه الگوریتم‌ها از نظر *DBDR* در پایگاه داده *db4* ۹۸
- شکل ۴-۲۴: نمودار مقایسه الگوریتم‌ها از نظر *DBDR* در پایگاه داده *db5* ۹۸
- شکل ۴-۲۵: نمودار مقایسه الگوریتم‌ها از نظر *DBDR* در پایگاه داده *db6* ۱۰۰
- شکل ۴-۲۶: نمودار مقایسه الگوریتم‌ها از نظر *DBDR* در پایگاه داده *db7* ۱۰۰
- شکل ۴-۲۷: نمودار مقایسه الگوریتم‌ها از نظر *TA* در پایگاه داده *db1* ۱۰۲
- شکل ۴-۲۸: نمودار مقایسه الگوریتم‌ها از نظر *TA* در پایگاه داده *db2* ۱۰۲
- شکل ۴-۲۹: نمودار مقایسه الگوریتم‌ها از نظر *TA* در پایگاه داده *db3* ۱۰۲
- شکل ۴-۳۰: نمودار مقایسه الگوریتم‌ها از نظر *TA* در پایگاه داده *db4* ۱۰۲
- شکل ۴-۳۱: نمودار مقایسه الگوریتم‌ها از نظر *TA* در پایگاه داده *db5* ۱۰۳
- شکل ۴-۳۲: نمودار مقایسه الگوریتم‌ها از نظر *TA* در پایگاه داده *db6* ۱۰۴
- شکل ۴-۳۳: نمودار مقایسه الگوریتم‌ها از نظر *TA* در پایگاه داده *db7* ۱۰۴
- شکل ۴-۳۴: نمودار مقایسه الگوریتم‌ها از نظر زمان اجرا در پایگاه داده *db1* ۱۰۶
- شکل ۴-۳۵: نمودار مقایسه الگوریتم‌ها از نظر زمان اجرا در پایگاه داده *db2* ۱۰۶
- شکل ۴-۳۶: نمودار مقایسه الگوریتم‌ها از نظر زمان اجرا در پایگاه داده *db3* ۱۰۶
- شکل ۴-۳۷: نمودار مقایسه الگوریتم‌ها از نظر زمان اجرا در پایگاه داده *db4* ۱۰۶
- شکل ۴-۳۸: نمودار مقایسه الگوریتم‌ها از نظر زمان اجرا در پایگاه داده *db5* ۱۰۷
- شکل ۴-۳۹: نمودار مقایسه الگوریتم‌ها از نظر زمان اجرا در پایگاه داده *db6* ۱۰۹
- شکل ۴-۴۰: نمودار مقایسه الگوریتم‌ها از نظر زمان اجرا در پایگاه داده *db7* ۱۰۹

فهرست جداول

- جدول ۱-۳: (a) یک مثال از پایگاه داده (b) جدول ارزش خارجی اقلام ۴۴
- جدول ۲-۳: الگوریتم مخفی سازی قلم با سودمندی بالا، اول ۴۸
- جدول ۳-۳: الگوریتم مخفی سازی قلم بابیش ترین تداخل با مجموعه-اقلام حساس، اول ۴۹
- جدول ۱-۴: جدول (الف) تراکنش ها و (ب) ارزش خارجی یک پایگاه داده فرضی ۵۳
- جدول ۲-۴: جداول تراکنش های پایگاه داده های پاک سازی شده ۵۴
- جدول ۳-۴: کارایی جواب های ارائه شده در مثال ۱ ۵۵
- جدول ۴-۴: شبه کد الگوریتم *HTMSCF* ۶۳
- جدول ۵-۴: شبه کد الگوریتم *HTMLCF* ۶۴
- جدول ۶-۴: شبه کد الگوریتم *HTMMCFCF* ۶۵
- جدول ۷-۴: شبه کد الگوریتم *HIMLCF* ۶۶
- جدول ۸-۴: شبه کد الگوریتم *HIMMCF* ۶۸
- جدول ۹-۴: مشخصات پایگاه داده های آزمایشی ۸۳
- جدول ۱۰-۴: تعداد مجموعه-اقلام چندتایی در پایگاه داده های آزمایشی ۹۱
- جدول ۱۱-۴: جدول مقایسه الگوریتم ها از نظر $MC(\%)$ در پایگاه داده های *db1*، *db2* و *db3* ۹۴
- جدول ۱۲-۴: جدول مقایسه الگوریتم ها از نظر $MC(\%)$ در پایگاه داده های *db4* و *db5* ۹۵
- جدول ۱۳-۴: جدول تعداد رتبه های نخست تا سوم کسب شده توسط الگوریتم ها در پایگاه داده های کوچک از نظر MC ۹۵
- جدول ۱۴-۴: جدول مقایسه الگوریتم ها از نظر $MC(\%)$ در پایگاه داده های *db6* و *db7* ۹۶
- جدول ۱۵-۴: جدول تعداد رتبه های نخست تا سوم کسب شده توسط الگوریتم ها در پایگاه داده های بزرگ از نظر $MC(\%)$ ۹۷
- جدول ۱۶-۴: جدول مقایسه الگوریتم ها از نظر $DBDR$ در پایگاه داده های *db1*، *db2* و *db3* ۹۹
- جدول ۱۷-۴: جدول مقایسه الگوریتم ها از نظر $DBDR$ در پایگاه داده های *db4* و *db5* ۹۹
- جدول ۱۸-۴: جدول تعداد رتبه های نخست تا سوم کسب شده توسط الگوریتم ها در پایگاه داده های کوچک از نظر $DBDR$ ۹۹
- جدول ۱۹-۴: جدول مقایسه الگوریتم ها از نظر $DBDR$ در پایگاه داده های *db6* و *db7* ۱۰۰

جدول ۴-۲۰: جدول تعداد رتبه‌های نخست تا سوم کسب شده توسط الگوریتم‌ها در پایگاه‌داده‌های بزرگ از نظر *DBDR* ۱۰۱

جدول ۴-۲۱: جدول مقایسه‌ی الگوریتم‌ها از نظر $TA(\%)$ در پایگاه‌داده‌های *db1*، *db2* و *db3* ۱۰۳

جدول ۴-۲۲: جدول مقایسه‌ی الگوریتم‌ها از نظر $TA(\%)$ در پایگاه‌داده‌های *db4* و *db5* ۱۰۳

جدول ۴-۲۳: جدول تعداد رتبه‌های نخست تا سوم کسب شده توسط الگوریتم‌ها در پایگاه‌داده‌های کوچک از نظر *TA* ۱۰۴

جدول ۴-۲۴: جدول مقایسه‌ی الگوریتم‌ها از نظر $TA(\%)$ در پایگاه‌داده‌های *db6* و *db7* ۱۰۵

جدول ۴-۲۵: جدول تعداد رتبه‌های نخست تا سوم کسب شده توسط الگوریتم‌ها در پایگاه‌داده‌های بزرگ از نظر *TA* ۱۰۵

جدول ۴-۲۶: جدول مقایسه‌ی الگوریتم‌ها از نظر زمان در پایگاه‌داده‌های *db1*، *db2* و *db3* ۱۰۷

جدول ۴-۲۷: جدول مقایسه‌ی الگوریتم‌ها از نظر زمان در پایگاه‌داده‌های *db4* و *db5* ۱۰۸

جدول ۴-۲۸: جدول تعداد رتبه‌های نخست تا سوم کسب شده توسط الگوریتم‌ها در پایگاه‌داده‌های کوچک از نظر زمان ۱۰۸

جدول ۴-۲۹: جدول مقایسه‌ی الگوریتم‌ها از نظر زمان در پایگاه‌داده‌های *db6* و *db7* ۱۰۹

جدول ۴-۳۰: جدول تعداد رتبه‌های نخست تا سوم کسب شده توسط الگوریتم‌ها در پایگاه‌داده‌های بزرگ از نظر زمان ۱۱۰

جدول ۴-۳۱: جدول مقایسه‌ی کلی الگوریتم‌ها ۱۱۲

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 تجارت 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

^۱Data Warehouse

^۲Data Mining

^۳Information Overload

^۴Classification

^۵Clustering

^۶Association 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 برده شده‌است.

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-اقدام مکرر^۲ کشف و استخراج می‌شوند و در مرحله دوم با استفاده از این مجموعه-اقدام، قوانین انجمنی تولید می‌شوند. تاکنون الگوریتم‌های بسیاری برای یافتن مجموعه-اقدام مکرر 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

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

^۱Association Rule Mining, ARM

^۲Frequent Itemsets

^۳Utility Mining

^۴Privacy Preserving Data Mining, PPDM

به صورت عمومی منتشر می‌کنند تا متخصصین و محققین بتوانند این اطلاعات را تحلیل و ارزیابی نمایند و سازمان‌ها و مؤسسات دیگر نیز 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^۱ 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]:

^۱Privacy

- **حفظ حریم خصوصی در انتشار داده‌ها^۱**: در این شاخه روش‌های مختلف تبدیل، برای تغییر داده‌ها جهت حفظ حریم خصوصی مورد بررسی قرار گرفته است. این روش‌ها شامل روش‌های تصادفی سازی^۲ [۶]، گمنامی درجه- k [۱۱، ۱۲]^۳ و تنوع درجه- l [۴۷]^۴ 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^۵**: این شاخه از روش‌ها مشابه روش‌های شاخه قبلی هستند با این تفاوت که در این‌جا نتایج پرسش‌ها محدود یا دست‌کاری می‌شوند. روش‌های آشفته-سازی^۶ 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^۷ 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

¹Privacy Preserving Data Publishing, PPDP

²Randomization

³k-Anonymity

⁴l-Diversity

⁵Query Auditing

⁶Perturbing

⁷Common 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)^۱:** 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 بر این، روش‌ها و الگوریتم‌ها هنگامی که ابعاد داده‌ها 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 را معرفی می‌کنیم [۲].

- **تصادفی‌سازی^۲:** 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^۳ حریم خصوصی در این روش در [۲۲، ۲۰، ۱۳، ۶] مورد بررسی قرار گرفته است. انواع حمله‌ها به روش تصادفی- روش در [۲۲، ۲۰، ۱۳، ۶] مورد بررسی قرار گرفته است.

¹Curse of Dimensionality

²Randomization

³Privacy Quantification

سازی نیز در [۳-۱] مورد بحث قرار گرفته‌اند. تصادفی‌سازی جریان داده‌ها^۱ در [۲۱،۲۸] و درهم‌سازی ضربی^۲ در کارهای [۱۶،۳۵،۴۴،۵۳،۵۵،۵۸] ارائه شده است. تعویض داده^۳ در [۲۳] و دیگر روش‌های تصادفی‌سازی در [۷،۲۱،۶۶،۸۰] مورد بحث قرار گرفته‌اند.

- **گمنام‌سازی گروهی^۴:** 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 شده است 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 بیندازد. 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 شده است که گروه‌هایی گمنام از رکوردها ساخته شود تا حمله‌کننده نتواند به رکورد مورد نظر با احتمال بالا دستیابی پیدا کند. روش گمنامی درجه k - [۷۰]، تنوع درجه t - [۴۷] و نزدیکی درجه t^5 - [۳۸] در این دسته قرار می‌گیرند.

- **حفظ حریم خصوصی توزیع شده:** هدف اصلی در بسیاری از الگوریتم‌های حفظ حریم خصوصی توزیع شده محاسبه توابع تجمعی آماری بر روی همه داده‌ها است به نحوی که داده‌های یک شریک برای شرکای دیگر افشا نشود. در این حالت شرکا علاقه‌مند هستند که در محاسبه توابع تجمعی آماری بر روی همه داده‌ها همکاری کنند اما به‌طور کامل به هم‌دیگر اعتماد ندارند. داده‌های به اشتراک گذاشته شده می‌تواند دو حالت داشته باشد: (۱)

¹Data Streams

²Multiplicative Perturbation

³Data Swapping

⁴Group Based Anonymization

⁵t-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 هر یک 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 پایگاه داده 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 پایگاه داده، محدود نمودن دسترسی به داده‌ها و همچنین انتشار نمونه‌هایی از داده‌ها به جای انتشار

^۱Horizontally Partitioned Data

^۲Vertically Partitioned Data

^۳Association Rule Hiding

^۴Downgrading Classifier Effectiveness

^۵Query 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 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 اقلام متکرر^۸) کاوش قوانین انجمنی [۵] به خوبی مورد بررسی قرار گرفته است اما حفظ حریم خصوصی در مدل‌های جدیدتر کاوش قوانین انجمنی مانند مدل کاوش مجموعه-اقلام-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.

^۱Heuristic

^۲Computational Complexity

^۳Scalability

^۴Local Optima

^۵Border Based Algorithms

^۶Exact Algorithms

^۷Privacy Preserving Utility Mining, PPUM

^۸Frequent Itemset Mining Model

^۹Share Frequent Itemset Mining Model

^{۱۰}Utility Mining Model

۱-۵- نمای کلی پایان نامه

در فصل اول مقدمه‌ای بر حفظ حریم خصوصی در کاوش سودمندی ارائه شد و دسته‌بندی‌ها و روش‌های 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.