
بسم الله الرحمن الرحيم

وزارت علوم، تحقیقات و فناوری

دانشگاه شهید مدنی آذربایجان

دانشکده فناوری اطلاعات و مهندسی کامپیوتر

گروه مهندسی فناوری اطلاعات

پایان‌نامه مقطع کارشناسی ارشد

رشته مهندسی فناوری اطلاعات

پنهان‌سازی قواعد وابستگی حساس در پایگاه داده‌های متمرکز با استفاده از روش‌های مبتنی بر تحریف

استاد راهنما:

دکتر شهرام حسین زاده

استاد مشاور:

دکتر ناصر فرج زاده

پژوهشگر:

امیرحسین شهسواری

خردادماه ۱۳۹۴

تبریز / ایران

تقدیم به

پدر و مادر مهربانم

و تشکر به خاطر زحمات و حمایت‌های بی دریغ‌شان

و تقدیم به

برادران عزیزم

محمدرضا و مرتضی

و برادرزاده‌های دوست‌داشتنی‌ام

مهدی، امیرمهدی و علی

که تک‌تکِ بیت‌های وجودم مملو از عشق تمامی آن‌هاست

تشکر و قدردانی

بیش از هرچیز از پروردگار منان به خاطر الطاف بی‌کرانش شاکرم که باری دیگر توفیق کسب دانش در سایه‌سار میهنم را به من عطا کرد. از استاد راهنمای محترم جناب آقای دکتر شهرام حسین-زاده و استاد مشاور ارجمندم جناب آقای دکتر ناصر فرج‌زاده به خاطر راهنمایی‌های مفیدشان در مسیر پژوهش و نگارش این پایان‌نامه و مقالات مستخرج از آن کمال تشکر و قدردانی را دارم.

هم‌چنین از اساتید محترم جناب آقای دکتر عسگرعلی بویر و جناب آقای دکتر محمد نادری دهکردی که در سایه‌ی دانش ایشان با مفاهیم این حوزه‌ی علمی آشنا شدم و علاقه به این موضوع را در من شکوفا کردند بسیار سپاسگزارم و برای تک‌تک آنان آرزوی سرافرازی و عزّت می‌کنم.

امیرحسین شهسواری

خردادماه ۱۳۹٤

تبریز، ایران

فهرست مطالب

فهرست جداول

فهرست اشکال

فهرست روابط

فهرست علائم و اختصارات

نماد	مفهوم
D	دیتاست اصلی
$\|D\|$	تعداد تراکنش های موجود دیتاست D
I	تعداد آیتم های موجود در D
X	آیتم ست X
$\|X\|$	تعداد آیتم ها (طول) آیتم ست X
C_k	مجموعه‌ی تمام k–آیتم‌ست‌های کاندید
L_k	مجموعه‌ی تمام k–آیتم‌ست‌های فراوان
L	مجموعه‌ی تمام آیتم‌ست‌های فراوان
minsupp	آستانه‌ی کمینه‌ی *Support*
minconf	آستانه‌ی کمینه‌ی *Confidence*
Support (X)	میزان پشتیبانی آیتم ست X
Confidence ($X \Rightarrow Y$)	میزان اطمینان قانون $X \Rightarrow Y$
SC (X) یا *SuppCount (X)*	تعداد تکرار آیتم ست X در پایگاه داده
Cluster (i)	خوشه ی i ام که تراکنش‌هایی را در خود نگه می‌دارد که طولشان برابر i می‌باشد
maxLength	طول بزرگ‌ترین تراکنش موجود در D
Absent_Array	آرایه ای دو بعدی است که برای هر آیتم ست کاندید، تعداد تراکنش هایی در هر خوشه که شامل آن آیتم ست نیستند را نگه داری می کند
Total_Absents	آرایه ای است که برای هر آیتم ست کاندید تعداد تراکنش هایی از خوشه ی اول تا محل جاری پویش که شامل آن آیتم ست نیستند را نگه داری می کند
T_i	تراکنش i ام
$\|T_i\|$	طول تراکنش i ام

TID	شناسه‌ی تراکنش
$minSuppCount$	حداقل تعداد تکرار یک آیتم‌ست برای آنکه فراوان شود
$minTotal\_Absents$	حداقل تعداد تراکنش‌هایی که باید از یک آیتم‌ست پشتیبانی نکنند تا آن آیتم‌ست غیرفراوان باشد
D'	دیتاست ایمن شده
R	قواعد وابستگی قابل استخراج از D
R'	قواعد وابستگی قابل استخراج از D'
R_s	قوانین حساس
$\sim R_s$	قوانین غیرحساس
$\lfloor . \rfloor$	تابع جزء صحیح
$\lceil . \rceil$	تابع سقف
$TDSNC\ (I_j)$	مجموع تفاضلات میان $minsupp$ و $Support$ قوانین غیرحساسی که شامل آیتم I_j هستند.
$TDCNC\ (I_j)$	مجموع تفاضلات میان $minconf$ و $Confidence$ قوانین غیرحساسی که شامل آیتم I_j هستند.
$TNNC\ (I_j)$	مجموع تعداد قوانین غیرحساسی که شامل آیتم I_j هستند.
IP	اولویت آیتم‌ها
TS	حساسیت تراکنش‌ها
TP	اولویت تراکنش‌ها
IF_i^j	تعداد تکرار آیتم I_j در سمت راست قوانین حساسی که توسط T_i پشتیبانی می‌شوند
INF_{VT}^j	تعداد تکرار آیتم I_j در سمت چپ قوانین حساسی که تراکنش VT تنها از سمت چپ آن‌ها پشتیبانی می‌کند

چکیده

پیشرفت‌های اخیر در فناوری اطلاعات و رسانه‌های ذخیره‌سازی دیجیتال، نگه‌داری حجم عظیمی از داده‌ها را با هزینه‌ی اندک امکان‌پذیر نموده است. بسیاری از صاحبان داده‌ها از این فرصت استفاده می‌کنند و به دلایل مختلفی (نظیر همکاری یک سازمان با سازمان‌های دیگر)، دست به انتشار داده‌های دیجیتال خود می‌زنند. از سوی دیگر، استفاده‌ی سودجویانه‌ی افراد و یا سازمان‌های رقیب از ابزارهای داده‌کاوی جهت استخراج دانش حساس از پایگاه‌داده-ی منتشر شده می‌تواند منافع و حریم خصوصی صاحبان داده‌ها را با خطر مواجه سازد. به این ترتیب، فیلد تحقیقاتی جدیدی تحت عنوان حفظ حریم خصوصی در داده‌کاوی از اهمیت خاصی برخوردار شده است.

در این پایان‌نامه روی پنهان سازی قواعد وابستگی به عنوان یکی از مهم‌ترین بخش‌های تحقیقاتی حفظ حریم خصوصی در داده کاوی تمرکز شده است و دو روش مبتنی بر تحریف به نام‌های FMARH و WMARH برای پنهان‌سازی قواعد وابستگی حساس در پایگاه‌داده‌های متمرکز ارائه شده است. در این روش‌ها برای اولین بار از مرحله‌ی استخراج قواعد وابستگی برای مقداردهی برخی از متغیرهای مورد نیاز در بخش پنهان‌سازی کمک گرفته شده است تا زمان اجرای الگوریتم‌ها کاهش یابد. در روش اول، تراکنش‌های حساسی برای ایمن‌سازی انتخاب می-شوند که طول کوتاه‌تری دارند. برای انتخاب آیتم قربانی در این روش از سه پارامتر برای کاهش میزان قواعد گم شده استفاده می‌شود. در روش دوم که از جمله روش‌های پنهان‌سازی چندقانونی به شمار می‌رود، برای انتخاب تراکنش‌ها علاوه بر درنظرگرفتن طول، میزان حساسیت آن‌ها نیز مورد توجه قرار می‌گیرد و آیتم قربانی با بهره‌گیری از پنج پارامتر به نحوی انتخاب می‌شود که ضمن پنهان‌سازیِ هم‌زمان قوانین مختلف، تأثیر کمتری روی قواعد غیرحساس بگذارد و از آشکار شدن مجدد قوانین حساسی که در مراحل گذشته پنهان شده بودند نیز جلوگیری به عمل آید. نتایج آزمایشات نشان می‌دهد ضمن آن که روش‌های FMARH و WMARH از سرعت بالایی برخوردار هستند از لحاظ اثرات جانبی نیز کارایی مناسبی در برابر روش‌های شناخته شده‌ی Algo 2b، MDSRRC و SIF-IDF دارند.

کلمات کلیدی: حفظ حریم خصوصی در داده‌کاوی، ایمن‌سازی پایگاه داده‌ها، پنهان‌سازی قواعد وابستگی، روش‌های مبتنی بر تحریف.

فصل اول

مقدمه

۱ مقدمه

۱-۱ داده‌کاوی و حفظ حریم خصوصی

روزانه میلیون‌ها داده توسط سازمان‌ها و سامانه‌های مختلف جمع آوری می‌شود. پیشرفت‌های گسترده در رسانه‌های ذخیره‌سازی دیجیتال و هزینه‌های بسیار اندک آن‌ها، سازمان‌ها را به نگه‌داری دیجیتال این حجم عظیم از داده‌ها ترغیب نموده است. بهره‌برداری کامل از این داده‌های خام نیازمند تحلیلی جامع است که بتواند دانش مفید پنهان در داده‌ها را استخراج نماید. اما حجم بالای داده‌ها این مسأله را به امری چالش برانگیز تبدیل می‌کند که از عهده‌ی روش‌های سنتی بر نمی آید. داده‌کاوی[1] فرایندی است که طی آن، دانش مفید و غیربدیهی نهفته در منابع داده‌ایِ حجیم به صورت خودکار استخراج می‌شود [۱].

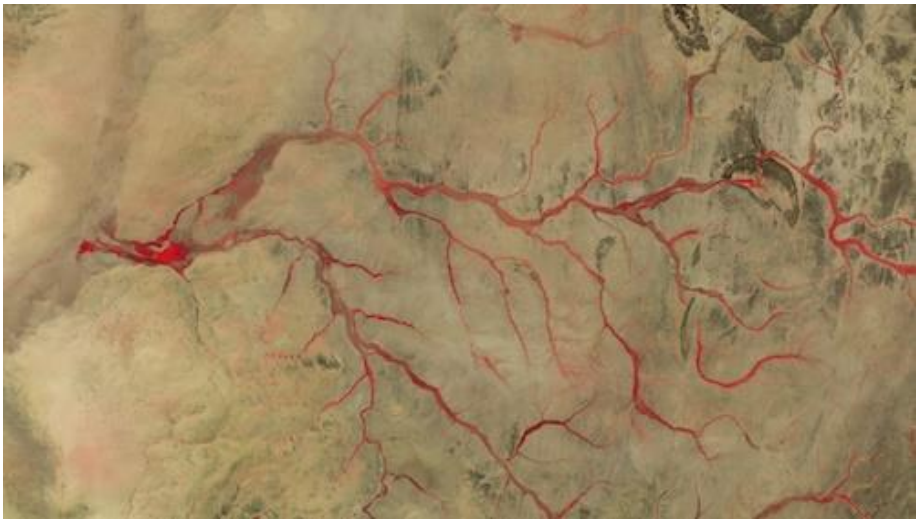
امروزه داده‌کاوی نقش مهمی در تصمیم‌گیری‌های استراتژیک سازمانی ایفا می‌کند. اما زمانی که سخن از انتشار داده‌ها به میان می‌آید نگرانی‌هایی نسبت به فاش شدن اطلاعات و دانش مرتبط با حریم خصوصی افراد و سازمان ها مطرح می‌شود. در دنیای امروزی دلایل مختلفی برای انتشار داده‌ها توسط صاحبان آن‌ها وجود دارد. در ادامه برای روشن‌تر شدن بحث، به برخی از دلایل انتشار خود خواسته‌ی داده‌ها اشاره می‌کنیم:

- نمایش صداقت و میزان پیشرفت خدمات: به عنوان مثال، دولت آمریکا صدها دیتاست[2] مرتبط با خدمات در حوزه‌های مختلف را روی پرتال اینترنتی خود منتشر کرده است [۲]. دیتاست‌های مربوط به مصرف گاز و حجم نواحی تحت پوشش خدمات گاز، دیتاست مربوط به میزان و نوع مصالح ساختمانی به کار رفته در ساختمان‌سازی طی سال‌های مختلف براساس ناحیه و ... از جمله دیتاست‌هایی است که در این پرتال منتشر شده است. این امر نه تنها صداقت و میزان پیشرفت در ارائه‌ی خدمات را به نحوی مناسب نمایش می‌دهد بلکه بستری مناسب برای انجام مطالعات تحقیقاتی روی داده‌ها را نیز فراهم می‌آورد.
- همکاری با سایر سازمان‌ها: بسیاری از شرکت‌ها تمایل دارند با سازمان‌های دیگری که دارای فعالیت‌های مشابه تجاری هستند همکاری نمایند و به این ترتیب منافع تجاری خود را تقویت نمایند. این همکاری در اکثر موارد مستلزمِ به اشتراک‌گذاری پایگاه‌داده‌های سازمان

[1] Data Mining
[2] Dataset

است. به عنوان نمونه می‌توان به همکاری تجاری میان Wal-Mart (از فروشندگان معتبر ایالات متحده‌ی آمریکا) و P&G[1] (از تولیدکنندگان نامدار جهانی) اشاره نمود. در سال ۱۹۸۸، Wal-Mart داده‌های مربوط به خرید مشتریان خود را از طریق یک شبکه‌ی ارتباطی با P&G به اشتراک گذاشت. P&G نیز با کاوش این داده‌ها، از نظر تجاری به Wal-Mart کمک کرد تا روی فروش آیتم‌هایی تمرکز کند که مشتریانش به آن‌ها نیاز بیشتری دارند. هم-چنین P&G بدون اینکه منتظر سفارشات Wal-Mart بماند می‌توانست تعداد و نوع کالاهای پرفروش را جهت ارسال به Wal-Mart آماده کند و مطابق با موجودی جاری کالا در Wal-Mart زمان انتقال کالاها را مدیریت نماید. این کار باعث افزایش سود هر دو طرف تجاری شد. هم‌چنین اشتراک داده‌ها باعث از بین‌رفتن خطاهای مالی پیشین شد [۳].

- انتشار داده‌ها به دلیل مسائل انسان‌دوستانه: شرکت ماهواره‌ای DMCii[2] بارها در مواقع بحرانی، عکس‌های ماهواره‌ای خود را به منظور کمک به مدیریت بحران در اختیار نهادهای مسئول کشورهای مختلف قرار داده است. به عنوان نمونه می‌توان به ارائه‌ی تصاویر ماهواره-ای مربوط به حمله‌ی ملخ‌ها در الجزایر (سال ۲۰۱٤) اشاره کرد. آنالیز این تصاویر به همراه تصاویر مربوط به پوشش گیاهی منطقه به پیش‌بینی نحوه‌ی انتشار و مقصد حرکت ملخ‌ها و درنتیجه تعیین محل سم‌پاشی کمک فراوانی کرد. شکل (۱-۱) نمونه ی تصویری را نشان می‌دهد که به منظور تعیین نحوه‌ی انتشار ملخ‌ها توسط DMCii در اختیار دولت الجزایر قرار داده شده است [٤].



شکل ۱-۱: تصاویر ماهواره‌ای و تخمین نحوه‌ی انتشار ملخ‌ها در الجزایر [٤]

[1] Procter & Gamble
[2] DMC International Imaging

- برون‌سپاری داده‌کاوی [۵]: برخی از سازمان‌ها به دلیل صرفه‌جویی در منابع مالی از استخدام متخصصین داده‌کاو خودداری می‌کنند و در عوض داده‌های خود را در اختیار داده‌کاوان خارجی قرار می‌دهند. عامل دیگر برون‌سپاری داده‌کاوی، بالابردن کیفیت تحلیل توسط شرکت‌های خبره در این امر است.

این موارد، تنها برخی از مهم‌ترین دلایل انتشار داده‌ها در دنیای امروزی است. در کنار فوایدی که انتشار داده‌ها برای صاحبان آن‌ها به همراه دارد در صورتی که داده‌های به اشتراک گذاشته شده با مقاصد سودجویانه مورد کاوش و آنالیز قرار بگیرد می‌تواند منافع مادی و معنوی صاحبان آن‌ها و حتی افراد مرتبط با آن‌ها (نظیر مشتریان سازمان) را با خطر جدی روبرو سازد. بنابراین انتشار داده-های حاوی دانش حساس، امری متضاد با منافع و حریم خصوصی تلقی می‌شود.

۲-۱ مفهوم حریم خصوصی

در این بخش، برای آشنایی بیشتر با مفهوم حریم خصوصی[1]، متداول‌ترین تعاریف مربوط به این حوزه ارائه می‌شود.

- حریم خصوصی عبارتست از حق افراد برای مخفی نگاه‌داشتن امور و روابط شخصی خود [٦].
- حریم خصوصی حالتی است که فرد می‌تواند از رصد شدن توسط سایرین و یا مزاحمت آن-ها در امان بماند. به عبارتی دیگر از توجه عموم آزاد باشد [۷].
- حریم خصوصی عبارتست از علاقه‌ی افراد به داشتن محیطی شخصی بدون دخالت و مزاحمت سایر افراد و سازمان‌ها [۸].
- حریم خصوصی داده‌ها[2]: عبارتست از حق یک موجودیت برای در امان ماندن از فاش شدن غیرمجاز اطلاعات حساسی که در یک مخزن الکترونیکی ذخیره شده است و یا اطلاعات حساسی که از داده‌های موجود در آن قابل استنباط است [۹].

با توجه به رشد فناوری و افزایش نگرانی افراد و سازمان‌ها نسبت به مسائل مرتبط با حریم خصوصی، قوانین مختلفی در این زمینه وضع شده است که از میان آن‌ها می‌توان به قوانین حفاظت از داده‌های شخصی اروپا [۱۰] و قانون حریم خصوصی HIPAA[3] (در رابطه با اطلاعات مربوط به

[1] Privacy
[2] Data/Information Privacy
[3] Health Insurance Portability and Accountability Act

سلامتی افراد) [۱۱، ۱۲] اشاره کرد. به عنوان مثال از دید قانون HIPAA، حریم خصوصی عبارتست از توانایی اشخاص برای کنترل و تعیین افرادی که می‌توانند به اطلاعات پزشکی آن‌ها دسترسی داشته باشند. مطابق این قانون، تمامی اطلاعاتی که شامل نام بیمار، شماره‌ی تلفن، آدرس و هر شماره‌ی شناسایی منحصربفردی از بیمار باشد در زمره‌ی اطلاعات محافظت شده دسته‌بندی می‌شود.

۱-۳ حفظ حریم خصوصی در داده‌کاوی

برای روشن‌تر شدن مفهوم حریم خصوصی در حوزه‌ی داده‌کاوی، نمونه‌ای بسط‌یافته و تکمیل شده از مثال موجود در [۱۳] را در نظر بگیرید: فرض کنید یک سرور[1] و چندین کلاینت[2] داریم (کلاینت‌های این سیستم فروشنده‌های دنیای واقعی هستند). هر کلاینت مجموعه‌ای از آیتم‌های خریداری شده توسط مشتریان را در نزد خود نگه می‌دارد. هم‌چنین فرض کنید کلاینت‌ها تمایل دارند با به اشتراک‌گذاری داده‌های خود امکانی فراهم بیاورند که سرور بتواند وابستگی‌های میان آیتم‌ها در مجموع دیتاست‌ها را به دست بیاورد. با استفاده از این روابط کلاینت‌ها خواهند توانست عملکرد خود را بهبود داده و پیشنهادات مناسب‌تری برای مشتریان خود ارائه نمایند. طبیعی است که کلاینت‌ها به منظور حفظ حریم خصوصی مشتریان خود علاقه‌ای ندارند سرور از هویت خریداران آیتم‌ها مطلع شود. بنابراین پیش از ارسال داده‌های خود به سرور، ابتدا آن را براساس سیاستی معین تغییر می‌دهند و یا شناسه‌ی خریداران را حذف می‌کنند.

حالات پیچیده‌تری از تجاوز به حریم خصوصی در مورد همکاری‌های سازمانی مشاهده می‌شود. در این حالت ممکن است سازمان رقیب از دانش موجود در دیتاست سازمان همکار خود سوء‌استفاده نموده و با اتخاذ سیاست‌هایی سازمان همکار را از صحنه‌ی رقابت تجاری خارج کند. نمونه‌ای از این نوع نقض حریم خصوصی در بخش ۱-۵ بیان خواهد شد.

حفظ حریم خصوصی در داده‌کاوی[3]، یک حوزه‌ی تحقیقاتی نسبتا جدید است که هدفش ارائه‌ی الگوریتم‌هایی برای تغییر داده‌های اصلی است به طوری که محرمانگی داده‌ها و دانش خصوصی حتی پس از فرایند داده‌کاوی نیز حفظ شود [۱٤]. بنابراین رویکردهای موجود در این حوزه به دو دسته‌ی اصلی تقسیم می‌شوند [۱۵]:

[1] Server
[2] Client
[3] Privacy Preserving Data Mining (PPDM)

- پنهان‌سازی داده‌ها[1]: در این رویکردها، داده‌های خام مورد توجه قرار می‌گیرند و سعی می‌شود پیش از انتشار داده‌ها، اطلاعات محرمانه و خصوصی موجود در آن‌ها حذف شود. ایجاد اختلال[2] و تبدیل[3] از جمله روش‌های این دسته می‌باشند. برای مثال باید پیش از انتشار یک دیتاست سلامت، کد شناسایی بیماران حذف شود تا هویت بیمار فاش نشود.
- پنهان‌سازی دانش[4]: هدف از این رویکردها، پنهان نمودن دانش حساسی است که در اثر اِعمال ابزارهای داده‌کاوی روی داده‌ها به دست می‌آید. پنهان‌سازی قواعد وابستگی[5]، پنهان‌سازی قواعد دسته‌بندی[6]، پنهان‌سازی مدل‌های خوشه‌بندی[7] و پنهان‌سازی دنباله[8] از جمله انواع مختلف رویکردهایی هستند که در این دسته قرار می‌گیرند.

در این پایان‌نامه روی پنهان‌سازی قواعد وابستگی حساس تمرکز می‌شود. با توجه به وابستگی این مبحث به مفاهیم قواعد وابستگی، لازم است ابتدا با اصول و تعاریف این حوزه بیشتر آشنا شویم.

۱-٤ قواعد وابستگی[9]

قواعد وابستگی در قالب استنتاج‌هایی به بیان روابط و وابستگی‌های قابل توجه[10] موجود در میان داده‌ها می‌پردازد [۱]. این استنتاج‌ها به صورت " اگر A آن‌گاه B " (یا $A \Rightarrow B$) بیان می‌شود. به عنوان مثال، می‌توان قانون زیر را در یک پایگاه‌داده‌ی فرضی ثبت احوال در نظر گرفت:

If *Age* ≥ *40 and MarriageStatus = Married*

Then *NumberOfChildren* ≥ *1*

در همین دیتاست، قانون زیر را نیز در نظر بگیرید:

If *Age* < *20*

Then *NumberOfChildren* ≥ *1*

[1] Data Hiding in Databases (DHD)
[2] Perturbation
[3] Transformation
[4] Knowledge Hiding in Databases (KHD)
[5] Association Rule Hiding
[6] Classification Rule Hiding
[7] Clustering Model Hiding
[8] Sequence Hiding
[9] Association Rules
[10] Interesting