

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



وزارت علوم، تحقیقات و فناوری

دانشگاه شهید مدنی آذربایجان

دانشکده فناوری اطلاعات و مهندسی کامپیوتر

گروه مهندسی فناوری اطلاعات

پایان نامه مقطع کارشناسی ارشد

رشته مهندسی فناوری اطلاعات

پنهان سازی قواعد وابستگی حساس در پایگاه داده های

متمرکز با استفاده از روش های مبتنی بر تحریف

استاد راهنما:

دکتر شهرام حسین زاده

استاد مشاور:

دکتر ناصر فرج زاده

پژوهشگر:

امیر حسین 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

محمد رضا و مرتضی

و برادرزاده‌های دوست‌داشتنی‌ام

مهدی، امیرمهدی و علی

که تک‌تک 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 سرفرازی و عزت می‌کنم.

امیرحسین شهسواری

خردادماه ۱۳۹۴

تبریز، ایران

فهرست مطالب

ج	فهرست جداول	۵۰
۵۰	فهرست اشکال	۵۰
ز	فهرست روابط	۵۰
ح	فهرست علائم و اختصارات	۵۰
یک	چکیده	۵۰
۱	مقدمه	۱
۲	۱-۱ داده‌کاوی و حفظ حریم خصوصی	۲
۴	۲-۱ مفهوم حریم خصوصی	۴
۵	۳-۱ حفظ حریم خصوصی در داده‌کاوی	۵
۶	۴-۱ قواعد وابستگی	۶
۱۱	۵-۱ مثالی برای تبیین لزوم پنهان‌سازی قواعد وابستگی حساس پیش از انتشار پایگاه‌داده	۱۱
۱۱	۶-۱ بیان مسأله	۱۱
۱۲	۷-۱ اهداف روش‌های پنهان‌سازی قواعد وابستگی	۱۲
۱۳	۸-۱ اثرات جانبی الگوریتم‌های پنهان‌سازی قواعد وابستگی	۱۳
۱۵	۹-۱ سایر تعاریف مورد نیاز	۱۵
۱۷	۱۰-۱ سازماندهی پژوهش	۱۷
۱۸	۲ پیشینه‌ی تحقیق	۱۸
۱۹	۱-۲ مروری بر چند روش استخراج قواعد وابستگی	۱۹
۲۲	۲-۲ دسته‌بندی روش‌های پنهان‌سازی قواعد وابستگی	۲۲
۲۲	۱-۲-۲ استراتژی پنهان‌سازی	۲۲
۲۳	۲-۲-۲ استراتژی تغییر داده‌ها	۲۳
۲۳	۳-۲-۲ تعداد پنهان‌سازی‌ها در هر تکرار از الگوریتم	۲۳
۲۴	۴-۲-۲ طبیعت الگوریتم	۲۴
۲۵	۵-۲-۲ سایر جنبه‌ها	۲۵
۲۶	۳-۲ مروری بر روش‌های پنهان‌سازی قواعد وابستگی حساس	۲۶
۳۹	۳ روش‌های پیشنهادی	۳۹
۴۱	۱-۳ استخراج آیتم‌ست‌های فراوان	۴۱
۴۱	۱-۱-۳ شمارش جزئی <i>Support</i> برای آیتم‌ست‌های غیرفراوان	۴۱
۴۲	۲-۱-۳ کاوش آیتم‌ست‌های فراوان با استفاده از PSCFII	۴۲
۵۱	۳-۱-۳ مثالی از PSCFII	۵۱
۵۵	۲-۳ استخراج قواعد وابستگی	۵۵

۵۹ ۳-۳ پنهان‌سازی قواعد وابستگی
۵۹ ۱-۳-۳ روش FMARH
۶۴ ۲-۳-۳ روش WMARH
۶۸ ۳-۳-۳ مثالی از FMARH و WMARH
۷۶ ۴ نتایج آزمایشگاهی
۷۷ ۱-۴ معیارهای کارایی
۷۷ ۱-۱-۴ نرخ شکست پنهان‌سازی
۷۷ ۲-۱-۴ نرخ قواعد گم شده
۷۸ ۳-۱-۴ نرخ قواعد غیرواقعی
۷۸ ۴-۱-۴ میزان عدم تشابه (تغییرات)
۷۹ ۲-۴ جزئیات آزمایشات
۸۱ ۳-۴ نتایج آزمایشات
۸۱ ۱-۳-۴ نرخ شکست پنهان‌سازی
۸۴ ۲-۳-۴ نرخ قواعد گم شده
۸۷ ۳-۳-۴ نرخ قواعد غیرواقعی
۹۰ ۴-۳-۴ نرخ عدم تشابه (تغییرات)
۹۳ ۵-۳-۴ زمان اجرای الگوریتم
۹۵ نتیجه‌گیری و کارهای آینده
۹۷ ضمائم
۹۸ ضمیمه‌ی الف: بررسی تأثیر تغییرات مختلف روی میزان اطمینان قانون
 ضمیمه‌ی ب: مثالی برای نمایش امکان ایجاد Ghost Rule در فرایند پنهان‌سازی، حتی در صورتی که آیتی به
۱۰۱ تراکنش‌ها افزوده نشود
۱۰۳ ضمیمه‌ی ج: اثبات لم‌های ۱ و ۲ از فصل ۳
۱۰۴ مراجع

فهرست جداول

- جدول ۱-۱: نمونه‌ای از دیتاست‌های سبد خرید ۷
- جدول ۲-۱: نمایش دودویی دیتاست جدول ۱-۱ ۸
- جدول ۱-۳: جزئیات متغیرهای استفاده شده در شبه‌کدهای این بخش ۴۴
- جدول ۲-۳: دیتاست مورد استفاده در مثال ۵۱
- جدول ۳-۳: خوشه‌های ساخته شده روی دیتاست مثال فوق ۵۲
- جدول ۴-۳: آرایه‌های Absent_Array برای آیتم‌های فراوان پس از اجرای تابع Initialize ۵۲
- جدول ۵-۳: آرایه‌های Absent_Array برای ۲-آیتم‌ست‌های کاندید پس از اجرای Generate_Candidate_Itemsets (L₁) ۵۳
- جدول ۶-۳: آرایه‌های Absent_Array به همراه مقدار SuppCount برای ۲-آیتم‌ست‌های کاندید پس از پوشش خوشه‌های مورد نیاز توسط تابع Generate_Large_Itemsets ۵۴
- جدول ۷-۳: آرایه‌های Absent_Array برای ۳-آیتم‌ست‌های کاندید پس از اجرای Generate_Candidate_Itemsets (L₂) ۵۴
- جدول ۸-۳: آرایه‌های Absent_Array به همراه مقدار SuppCount برای ۳-آیتم‌ست‌های کاندید پس از پوشش خوشه‌های مورد نیاز توسط تابع Generate_Large_Itemsets ۵۵
- جدول ۹-۳: دیتاست مورد استفاده در مثال ۶۸
- جدول ۱۰-۳: خوشه‌های ساخته شده روی دیتاست مثال فوق ۶۸
- جدول ۱۱-۳: قوانین استخراج شده از دیتاست به همراه میزان Support و Confidence ۶۹
- جدول ۱۲-۳: آرایه‌های TDSNC, TDCNC و TNNC پس از اجرای تابع genrules ۶۹
- جدول ۱۳-۳: مقادیر بروزشده‌ی آرایه‌های TDSNC, TDCNC و TNNC پس از تعیین قواعد حساس ۷۰
- جدول ۱۴-۳: اولویت آیتم‌های مختلف در روش FMARH ۷۰
- جدول ۱۵-۳: دیتاست ایمن شده توسط روش FMARH ۷۱
- جدول ۱۶-۳: پشتیبانی تراکنش‌ها از قوانین حساس ۷۱
- جدول ۱۷-۳: مقادیر درایه‌های IF و INF برای تراکنش‌ها و آیتم‌های مختلف: سلول‌های خالی، مقدار صفر دارند. ۷۲
- جدول ۱۸-۳: اولویت تراکنش‌های حساس در روش WMARH ۷۳
- جدول ۱۹-۳: تراکنش‌های حساس مرتب شده به صورت نزولی براساس اولویت ۷۴
- جدول ۲۰-۳: تراکنش‌های حساس مرتب شده به صورت نزولی براساس اولویت پس از پنهان شدن قانون دوم ۷۵
- جدول ۲۱-۳: دیتاست ایمن شده توسط روش WMARH ۷۵

- جدول ۴-۱: جزئیات دیتاست‌های مورد استفاده در آزمایشات ۷۹
- جدول ۴-۲: جزئیات تست‌های مختلف سناریوی اول ۸۰
- جدول ۴-۳: جزئیات تست‌های مختلف سناریوی دوم ۸۱
- جدول ۴-۴: س. اول. مقیاس پذیری زمانی (برحسب ثانیه) برحسب اندازه‌ی دیتاست روی دیتاست Kosarak ۹۳
- جدول ۴-۵: س. اول. مقیاس پذیری زمانی (برحسب ثانیه) برحسب اندازه‌ی دیتاست روی دیتاست Retail ۹۳
- جدول ۴-۶: س. اول. مقیاس پذیری زمانی (برحسب ثانیه) برحسب تعداد قواعد حساس روی دیتاست Kosarak ۹۴
- جدول ۴-۷: س. اول. مقیاس پذیری زمانی (برحسب ثانیه) برحسب تعداد قواعد حساس روی دیتاست Retail ۹۴
- جدول ۴-۸: س. دوم. مقیاس پذیری زمانی (برحسب ثانیه) به صورت ترکیبی ۹۴
- جدول ب-۱: پایگاه داده‌ی اصلی ضمیمه‌ی ب ۱۰۱
- جدول ب-۲: پایگاه داده‌ی ایمن شده‌ی ضمیمه‌ی ب ۱۰۲

فهرست اشکال

- شکل ۱-۱: تصاویر ماهواره‌ای و تخمین نحوه‌ی انتشار ملخ‌ها در الجزایر [۴]..... ۳
- شکل ۲-۱: بیان تصویری مسأله‌ی پنهان‌سازی قواعد وابستگی..... ۱۲
- شکل ۳-۱: نمایش تصویری اثرات جانبی الگوریتم‌های پنهان‌سازی قواعد وابستگی..... ۱۵
- شکل ۴-۱: نمونه‌ای از ساختار یک گراف آیت‌ست‌های فراوان [۲۰]..... ۱۶
- شکل ۱-۲: دسته بندی روش‌های پنهان‌سازی قواعد وابستگی از چهار بعد مختلف..... ۲۵
- شکل ۱-۳: چهارچوب کلی متداول برای پنهان‌سازی قواعد وابستگی..... ۴۰
- شکل ۲-۳: شبه‌کد تابع اصلی الگوریتم PSCFII..... ۴۵
- شکل ۳-۳: شبه‌کد تابع Initialize..... ۴۸
- شکل ۴-۳: شبه‌کد تابع Generate_Candidate_Itemsets..... ۴۹
- شکل ۵-۳: شبه‌کد تابع Generate_Large_Itemsets برای استخراج آیت‌ست‌های فراوان..... ۵۰
- شکل ۶-۳: شبه‌کد تابع Generate_Large_Itemsets برای استخراج آیت‌ست‌های فراوان جهت استفاده به عنوان فاز اول استخراج قواعد وابستگی..... ۵۷
- شکل ۷-۳: شبه‌کد توابع Generate_Association_Rules و genrules..... ۵۸
- شکل ۸-۳: شبه‌کد سفارشی شده‌ی تابع genrules برای مقداردهی متغیرهای مورد نیاز در فرایند پنهان‌سازی..... ۶۱
- شکل ۹-۳: شبه‌کد تابع FMARH..... ۶۳
- شکل ۱۰-۳: شبه‌کد تابع WMARH..... ۶۷
- شکل ۱-۴: س. اول. مقیاس پذیری نرخ شکست پنهان‌سازی برحسب اندازه‌ی دیتاست روی دیتاست Kosarak .. ۸۱
- شکل ۲-۴: س. اول. مقیاس پذیری نرخ شکست پنهان‌سازی برحسب اندازه‌ی دیتاست روی دیتاست Retail..... ۸۲
- شکل ۳-۴: س. اول. مقیاس پذیری نرخ شکست پنهان‌سازی برحسب تعداد قواعد حساس روی دیتاست Kosarak..... ۸۲
- شکل ۴-۴: س. اول. مقیاس پذیری نرخ شکست پنهان‌سازی برحسب تعداد قواعد حساس روی دیتاست Retail..... ۸۳
- شکل ۵-۴: س. دوم. مقیاس پذیری نرخ شکست پنهان‌سازی به صورت ترکیبی..... ۸۳
- شکل ۶-۴: س. اول. مقیاس پذیری نرخ قواعد گم شده از نظر اندازه‌ی دیتاست روی دیتاست Kosarak..... ۸۴
- شکل ۷-۴: س. اول. مقیاس پذیری نرخ قواعد گم شده از نظر اندازه‌ی دیتاست روی دیتاست Retail..... ۸۵
- شکل ۸-۴: س. اول. مقیاس پذیری نرخ قواعد گم شده از نظر تعداد قواعد حساس روی دیتاست Kosarak..... ۸۵
- شکل ۹-۴: س. اول. مقیاس پذیری نرخ قواعد گم شده از نظر تعداد قواعد حساس روی دیتاست Retail..... ۸۶
- شکل ۱۰-۴: س. دوم. مقیاس پذیری نرخ قواعد گم شده به صورت ترکیبی..... ۸۶

- شکل ۴-۱۱: س. اول. مقیاس پذیری نرخ قواعد غیرواقعی از نظر اندازه‌ی دیتاست روی دیتاست Kosarak ۸۷
- شکل ۴-۱۲: س. اول. مقیاس پذیری نرخ قواعد غیرواقعی از نظر اندازه‌ی دیتاست روی دیتاست Retail ۸۸
- شکل ۴-۱۳: س. اول. مقیاس پذیری نرخ قواعد غیرواقعی از نظر تعداد قواعد حساس روی دیتاست Kosarak ... ۸۸
- شکل ۴-۱۴: س. اول. مقیاس پذیری نرخ قواعد غیرواقعی از نظر تعداد قواعد حساس روی دیتاست Retail ۸۹
- شکل ۴-۱۵: س. دوم. مقیاس پذیری نرخ قواعد غیرواقعی به صورت ترکیبی ۸۹
- شکل ۴-۱۶: س. اول. مقیاس پذیری نرخ عدم تشابه از نظر اندازه‌ی دیتاست روی دیتاست Kosarak ۹۰
- شکل ۴-۱۷: س. اول. مقیاس پذیری نرخ عدم تشابه از نظر اندازه‌ی دیتاست روی دیتاست Retail ۹۱
- شکل ۴-۱۸: س. اول. مقیاس پذیری نرخ عدم تشابه از نظر تعداد قواعد حساس روی دیتاست Kosarak ۹۱
- شکل ۴-۱۹: س. اول. مقیاس پذیری نرخ عدم تشابه از نظر تعداد قواعد حساس روی دیتاست Retail ۹۲
- شکل ۴-۲۰: س. دوم. مقیاس پذیری نرخ عدم تشابه به صورت ترکیبی ۹۲

فهرست روابط

- ۹ رابطه‌ی (۱-۱) نحوه‌ی محاسبه‌ی میزان Support آیت‌مست‌ها
- ۹ رابطه‌ی (۲-۱) نحوه‌ی محاسبه‌ی minsupCount
- ۱۰ رابطه‌ی (۳-۱) نحوه‌ی محاسبه‌ی Support قوانین
- ۱۰ رابطه‌ی (۴-۱) نحوه‌ی محاسبه‌ی Confidence قوانین
- ۴۲ رابطه‌ی (۱-۳) نحوه‌ی محاسبه‌ی Total_Absents(c)
- ۴۲ رابطه‌ی (۲-۳) نحوه‌ی محاسبه‌ی k-1 عنصر اول Absent_Array(c) که c یک k-آیت‌مست است
- ۴۳ رابطه‌ی (۳-۳) نحوه‌ی محاسبه‌ی عناصر باقی مانده‌ی Absent_Array(c)
- ۶۲ رابطه‌ی (۴-۳) نحوه‌ی محاسبه‌ی اولویت آیت‌مست‌ها در روش FMARH
- ۶۵ رابطه‌ی (۵-۳) نحوه‌ی محاسبه‌ی حساسیت تراکنش‌ها در روش WMARH
- ۶۵ رابطه‌ی (۶-۳) نحوه‌ی محاسبه‌ی اولویت تراکنش‌ها در روش WMARH
- ۶۶ رابطه‌ی (۷-۳) نحوه‌ی محاسبه‌ی اولویت آیت‌مست‌ها در روش WMARH
- ۷۷ رابطه‌ی (۱-۴) نحوه‌ی محاسبه‌ی نرخ شکست پنهان‌سازی
- ۷۸ رابطه‌ی (۲-۴) نحوه‌ی محاسبه‌ی نرخ قواعد گم شده
- ۷۸ رابطه‌ی (۳-۴) نحوه‌ی محاسبه‌ی نرخ قواعد غیرواقعی
- ۷۸ رابطه‌ی (۴-۴) روش اول محاسبه‌ی نرخ عدم تشابه
- ۷۹ رابطه‌ی (۵-۴) روش دوم محاسبه‌ی نرخ عدم تشابه

فهرست علائم و اختصارات

مفهوم	نماد
دیتاست اصلی	D
تعداد تراکنش های موجود دیتاست D	$ D $
تعداد آیتم های موجود در D	I
آیتم ست X	X
تعداد آیتم ها (طول) آیتم ست X	$ X $
مجموعه ی تمام k -آیتم ست های کاندید	C_k
مجموعه ی تمام k -آیتم ست های فراوان	L_k
مجموعه ی تمام آیتم ست های فراوان	L
آستانه ی کمینه ی $Support$	$minsupp$
آستانه ی کمینه ی $Confidence$	$minconf$
میزان پشتیبانی آیتم ست X	$Support(X)$
میزان اطمینان قانون $X \Rightarrow Y$	$Confidence(X \Rightarrow Y)$
تعداد تکرار آیتم ست X در پایگاه داده	$SC(X)$ یا $SuppCount(X)$
خوشه ی i ام که تراکنش هایی را در خود نگه می دارد که طولشان برابر i می باشد	$Cluster(i)$
طول بزرگ ترین تراکنش موجود در D	$maxLength$
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	$Absent_Array$
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	$Total_Absents$
تراکنش i ام	T_i
طول تراکنش i ام	$ T_i $

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

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 پایگاه‌داده‌ی 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 پنهان‌سازی قواعد وابستگی به عنوان یکی از مهم‌ترین بخش‌های تحقیقاتی حفظ حریم خصوصی در داده‌کاوی 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 می‌شوند 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 در نظر گرفتن طول، میزان حساسیت آن‌ها نیز مورد توجه قرار می‌گیرد و آیت‌م قربانی با بهره‌گیری از پنج پارامتر به نحوی انتخاب می‌شود که ضمن پنهان‌سازی 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 MDSRRC, Algo 2b و SIF-IDF دارند.

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

فصل اول

مقدمه

۱-۱ داده‌کاوی و حفظ حریم خصوصی

روزانه میلیون‌ها داده توسط سازمان‌ها و سامانه‌های مختلف جمع‌آوری 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 نمی‌آید. داده‌کاوی^۱ فرایندی است که طی آن، دانش مفید و غیربدیهی نهفته در منابع داده‌ای 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 سازمان‌ها 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:

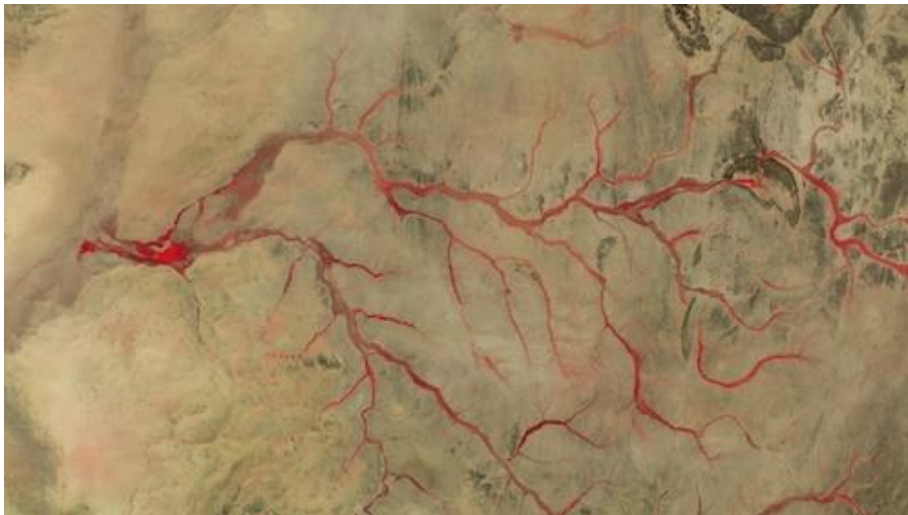
- نمایش صداقت و میزان پیشرفت خدمات: به عنوان مثال، دولت آمریکا صدها دیتاست^۲ مرتبط با خدمات در حوزه‌های مختلف را روی پرتال اینترنتی خود منتشر کرده است [۲]. دیتاست‌های مربوط به مصرف گاز و حجم نواحی تحت پوشش خدمات گاز، دیتاست مربوط به میزان و نوع مصالح ساختمانی به کار رفته در ساختمان‌سازی طی سال‌های مختلف 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 مستلزم به اشتراک‌گذاری پایگاه‌داده‌های سازمان

^۱ Data Mining

^۲ Dataset

است. به عنوان نمونه می‌توان به همکاری تجاری میان Wal-Mart (از فروشندگان معتبر ایالات متحده آمریکا) و P&G¹ (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 آیت‌هایی تمرکز کند که مشتریان به آن‌ها نیاز بیشتری دارند. هم‌چنین P&G بدون اینکه منتظر سفارشات Wal-Mart بماند می‌توانست تعداد و نوع کالاهای پر فروش 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 بین رفتن خطاهای مالی پیشین شد [۳].

- انتشار داده‌ها به دلیل مسائل انسان‌دوستانه: شرکت ماهواره‌ای DMCii² 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 ماهواره‌ای مربوط به حمله‌ی ملخ‌ها در الجزایر (سال ۲۰۱۴) اشاره کرد. آنالیز این تصاویر به همراه تصاویر مربوط به پوشش گیاهی منطقه به پیش‌بینی نحوه‌ی انتشار و مقصد حرکت ملخ‌ها و در نتیجه تعیین محل سم‌پاشی کمک فراوانی کرد. شکل (۱-۱) نمونه‌ی تصویری را نشان می‌دهد که به منظور تعیین نحوه‌ی انتشار ملخ‌ها توسط DMCii در اختیار دولت الجزایر قرار داده شده است [۴].



شکل ۱-۱: تصاویر ماهواره‌ای و تخمین نحوه‌ی انتشار ملخ‌ها در الجزایر [۴]

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 خیره 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 داده‌های 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^۱، متداول‌ترین تعاریف مربوط به این حوزه ارائه می‌شود.

- حریم خصوصی عبارتست از حق افراد برای مخفی نگاه‌داشتن امور و روابط شخصی خود [۶].
- حریم خصوصی حالتی است که فرد می‌تواند از رصد شدن توسط سایرین و یا مزاحمت آن‌ها 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^۲: 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^۳ (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 منحصر بفردی از بیمار باشد در زمره‌ی اطلاعات محافظت شده دسته‌بندی می‌شود.

۳-۱ 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^۱ و چندین کلاینت^۲ 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 سوء-استفاده نموده و با اتخاذ سیاست‌هایی سازمان همکار را از صحنه‌ی رقابت تجاری خارج کند. نمونه-ای از این نوع نقض حریم خصوصی در بخش ۱-۵ بیان خواهد شد.

حفظ حریم خصوصی در داده‌کاوی^۳، یک حوزه‌ی تحقیقاتی نسبتاً 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 [۱۴]. 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^۱: 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^۲ و تبدیل^۳ 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^۴: هدف از این رویکردها، پنهان نمودن دانش حساسی است که در اثر اعمال ابزارهای داده‌کاوی روی داده‌ها به دست می‌آید. پنهان‌سازی قواعد وابستگی^۵، پنهان‌سازی قواعد دسته‌بندی^۶، پنهان‌سازی مدل‌های خوشه‌بندی^۷ و پنهان‌سازی دنباله^۸ 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.

۴-۱ قواعد وابستگی^۹

قواعد وابستگی در قالب استنتاج‌هایی به بیان روابط و وابستگی‌های قابل توجه^{۱۰} 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 $Age \geq 40$ and $MarriageStatus = Married$

Then $NumberOfChildren \geq 1$

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

If $Age < 20$

Then $NumberOfChildren \geq 1$

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