تحلیل و کشف جرم از طریق کاوش متون در فضای مجازی

دریافت مقاله: ۹۱/۱۱/۱۰ پذیرش مقاله: ۹۲/۲/۲۰

بهزاد لک[۱]، جلال رضایی‌نور[۲]

از صفحهٔ ۱۵۹ تا ۱۸۲

چکیده

با گسترش شگرف اینترنت و استفاده روزافزون از آن در جهت ارایه و یا کسب اطلاعات، مفهومی تحت عنوان «افزونگی اطلاعاتی» مطرح است. آنچه امروزه از اهمیت بسیار زیادی برخوردار است، کمبود یا نبود اطلاعات نیست، بلکه کمبود روش‌هایی در جهت یافتن و بهره‌برداری از اطلاعات در دسترس، به نحوی بهینه است. متن کاوی، به‌عنوان روشی در استخراج دانش از متون، یکی از موضوعات مهم در گستره‌ای از اعمال مدیریت اطلاعات است. میزان زیادی از اطلاعاتی که به شکل متنی و غیر ساختاری ذخیره شده، حاوی داده‌های ضبط نشده با ارزشی هستند. داده کاوی و متن کاوی دو فناوری مهم و مناسب برای کشف الگوهای درونی در بین مجموعه عظیمی از داده‌ها است. از این‌رو بسیاری از سازمان‌های امنیتی، برای تشخیص و پیش‌بینی عمل جرم، به شیوه های داده کاوی و متن کاوی وابسته هستند.

در این مقاله پس از معرفی داده کاوی و رشته‌های علمی مرتبط با آن، روش‌های داده کاوی در قالب دو دسته توصیفی و پیش بینانه تشریح شده است. سپس کاربرد داده کاوی در حوزه‌های کاربردی پلیس قبل و بعد از وقوع جرایم مطرح شده و متن کاوی به عنوان یکی از کاربردهای مهم داده کاوی در پیشگیری و پیش‌بینی جرایم معرفی می‌شود. با توجه به اینکه بیشترین کاربرد متن کاوی در زمان قبل از وقوع جرم است، در ادامه کاربرد فناوری‌های جدید و همچنین به‌کارگیری روش‌های مبتنی بر متن کاوی، در موضوعات مرتبط با جرایم بررسی می‌شود. از این‌رو با توجه به نتایج به‌دست آمده، بومی کردن این‌گونه فناوری‌ها و روش‌های متن کاوی در ناجا، می تواند در راستای افزایش توان اطلاعاتی، پیش بینی و پیشگیری از جرایم مؤثر باشد.

کلیدواژه‌ها

داده کاوی، متن کاوی، پیشگیری جرایم، پیش بینی جرایم، شناسایی جرایم.

---

۱. عضو هیئت علمی دانشگاه علوم انتظامی امین

۲. استادیار دانشکده فنی مهندسی دانشگاه قم

مقدمه

امروزه بیش از ۸۰ درصد از دانش‌های کسب شده، به صورت متن، مستندات و دیگر صورت‌های رسانه‌ای نظیر ویدیو و صدا، نگهداری می‌شوند. از این‌رو توانایی‌های فنی بشر در برای تولید و جمع آوری داده‌ها، به سرعت افزایش یافته است. عواملی نظیر استفاده گسترده از روش‌های مکانیزه جمع‌آوری اطلاعات، به خدمت گرفتن رایانه در کسب و کار، علوم، خدمات دولتی، سیستم‌های سنجش از دور ماهواره‌ای و ... در این تغییرات نقش مهمی دارند.

به‌طور کلی استفاده همگانی از وب و اینترنت به عنوان یک سیستم اطلاع رسانی جهانی افراد را با حجم زیادی از داده و اطلاعات مواجه می‌کند. این رشد انفجاری در داده‌های ذخیره شده، نیاز مبرم وجود فناوری های جدید و ابزارهای خودکاری را ایجاد کرده که به صورت هوشمند، به انسان یاری می رسانند تا این حجم زیاد داده را به اطلاعات و دانش تبدیل کند.

در دسترس بودن حجم وسیعی از داده ها و نیاز شدید به اینکه از این داده‌ها اطلاعات و دانش سودمند استخراج شود، باعث شد تا روش‌هایی مانند داده کاوی[1]، متن کاوی[2] و وب کاوی[3]، کانون توجهات در صنعت اطلاعات قرار گیرد.

در دنیای کنونی این کمبود اطلاعات نیست که مسئله است، بلکه کمبود دانشی است که از این اطلاعات می‌توان حاصل کرد. میلیون‌ها صفحه وب، میلیون‌ها کلمه در کتابخانه‌های دیجیتال و هزاران صفحه اطلاعات در هر شرکت، تنها بخشی از این منابع اطلاعاتی هستند و به هیچ عنوان نمی‌توان به طور مشخص منبعی از دانش را در این بین معرفی کرد. دانش، نه تنها خود اطلاعات خلاصه است، بلکه نتیجه‌گیری و حاصل فکر و تحلیل بر روی اطلاعات می‌باشد.

یکی از حوزه‌هایی که در سال های اخیر به عنوان داده کاوی مورد توجه قرار گرفته

1. Data mining
2. Text mining
3. Web Mining

است، مسایل مربوط به تشخیص و کشف جرایم است. شناسایی جرایم، پیش‌بینی و پیشگیری از آنها و همچنین عوامل تأثیرگذار در ارتکاب جرایم، مسایلی هستند که با استفاده از مفاهیم جرم شناسی و فنون داده کاوی، مورد تحلیل قرار گرفته‌اند.

سرعت در حال افزایش میزان اطلاعات متنی ذخیره شده روی رایانه‌های شخصی یا روی وب، و شبکه‌های اجتماعی که در فضای سایبر فعالیت می‌کنند، باعث شده‌اند که اندازه اطلاعات، سیر صعودی را دنبال کند. یکی از مخاطراتی که در این میان مطرح است، دشوار بودن بازیابی اطلاعات مهمی است که بعضی اوقات منجر به ارتکاب جرم توسط اعضای این‌گونه شبکه‌ها می‌شود. از این‌رو روش هوشمند متن کاوی یکی از کاربردی ترین روش‌ها جهت استخراج الگوها از رسانه‌های متنی به شمار می آید.

در این تحقیق سعی شده است؛ پس از معرفی داده کاوی و اجزای آن، متن کاوی به عنوان یکی از روش‌های هوشمند بازیابی اطلاعات، معرفی و در ادامه، کاربرد این روش در تشخیص و تحلیل جرایم بیان شود. نتایج نشان می‌دهد کاربردی کردن این‌گونه از روش‌های هوشمند در ساختار پلیس، نه تنها الگوهای ارزشمندی از نحوه ارتکاب جرایم در فضای سایبر استخراج می‌کند، بلکه از طریق یافته های موجود می‌توان راهبردهای منطقی در جهت پیشگیری و مقابله به جرایم سایبری اتخاذ کرد.

داده‌کاوی

داده کاوی به معنای استخراج دانش از حجم عظیم داده‌ها است و به عنوان مهم‌ترین مرحله در فرایند کشف دانش معرفی شده است. شکل (۱) رشته های علمی مرتبط با داده کاوی را نشان می‌دهد. روش‌های داده کاوی اعم از ابزارهای توصیفی و پیش بینانه در حوزه های مختلفی وارد شده و حجم عظیمی از تحقیقات را به خود اختصاص داده است. از جمله حوزه‌های کاربردی داده کاوی، می‌توان به کابردهای تجاری، مدیریتی، پزشکی، ورزشی، اقتصاد سنجی، مدیریت مالی، وب کاوی و متن کاوی اشاره کرد (هان و کامبر، ۲۰۰۶). یکی از کاربردهایی که در داده کاوی مورد توجه قرار گرفته است،

مسایل مرتبط با شناسایی، پیش بینی و پیشگیری از جرایم است.

شکل ۱: رشته های علمی مرتبط با داده کاوی

داده کاوی به عنوان مهم‌ترین مرحله فرایند کشف دانش معرفی شده است. در یک دید کلی، داده کاوی را می‌توان به عنوان یک فرآیند چهار مرحله‌ای تعریف کرد:

۱. جمع آوری یک مجموعه از داده ها برای تحلیل؛

۲. ارائه این داده‌ها به برنامه نرم افزاری داده کاوی؛

۳. تفسیر نتایج؛

۴. به‌کارگیری نتایج برای مسئله یا موقعیت‌های جدید.

الف) روش‌های داده کاوی

دو دسته توصیفی[1] و پیش بینانه[2] برای روش‌های داده کاوی مطرح است. در روش های

1. Descriptive
2. Predictive

توصیفی، هدف توصیف یک رویداد با یک واقعیت است؛ اما در روش‌های پیش‌بینی، هدف پیش‌بینی متغیر ناشناخته از داده های آتی است (تان و همکاران، ۲۰۰۶).

شکل ۲: روش های داده کاوی

روش رگرسیون: متغییرهای خروجی را با متغییرهای ورودی متعدد ارتباط می‌دهد. در حقیقت ارتباط بین متغیرهای ورودی و خروجی را به‌صورت خطی برقرار می کند:

$$y = a + X_1 * b_1 + X_2 * b_2 + \cdots + X_n * b_n$$

که در آن y متغیر خروجی و وابسته، x ها متغیرهای ورودی و a,b نیز ضرایب رگرسیون هستند.

روش سری‌های زمانی: یک سری زمانی دنباله‌ای مرتب شده از مشاهده‌ها است. سری زمانی بر حسب زمان(فواصل زمانی مساوی) و بر اساس ابعاد دیگری مانند فاصله نیز مرتب می‌شود. یک سری زمانی مانند پالس‌های الکتریکی، یک سری پیوسته است. هر سری زمانی ارزش یک شئ را به عنوان تابعی از زمان در مجموعه داده‌های جمع‌آوری‌شده توصیف می‌کند (غضنفری و دیگران، ۱۳۸۶).

روش پیش بینی: روش های دسته بندی برای پیش‌بینی مشخصه‌های گسسته مورد استفاده قرار می‌گیرند؛ در حالی که در روش‌های پیش‌بینی توابع پیوسته را به عنوان نمونه قرار می‌دهند. این روش‌ها شامل رگرسیون خطی و غیرخطی، شبکه عصبی و ماشین‌های بردار پشتیبان هستند (همان).

روش دسته‌بندی: فرایند یافتن الگویی که با تشخیص طبقات یا یافتن مفاهیم داده، می‌تواند طبقه ناشناخته اشیاء دیگر را پیش‌بینی کند. برخی از روش‌های متداول دسته بندی شامل: درخت تصمیم، طبقه‌بندی بیز(بیز ساده و شبکه‌های بیزی)، شبکه‌های عصبی، ماشین‌های بردار پشتیبان و... .

خوشه‌بندی: یکی از روش‌های توصیفی است که داده‌های بدون برچسب را در قالب گروه‌هایی تحلیل می کند. این تقسیم‌بندی به شکلی است که داده‌های داخل هر خوشه بیشترین شباهت نسبت به یکدیگر و بیشترین اختلاف را با خوشه‌های دیگر دارند (همان).

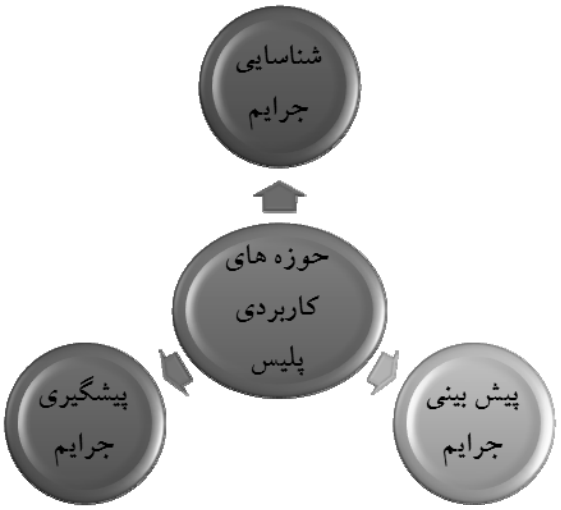
خلاصه‌سازی: در برگیرنده سایر روش‌هایی است که برای دستیابی به یک توصیف فشرده از داده‌ها مورد استفاده قرار می‌گیرند و معمولاً در تولید گزارش استفاده می‌شوند. مانند به‌دست آوردن میانگین و انحراف معیار برای موضوع‌های موردنظر فنون مصورسازی چند متغیر و کشف روابط تابعی بین متغیرها.

روش قوانین با هم آیی: استخراج قوانین با هم آیی یک حالت غیر نظارتی داده کاوی است که به جستجو برای یافتن ارتباط میان ویژگی‌ها در مجموعه داده‌ها می‌پردازد. در حقیقت تحلیل وابستگی‌ها، مطالعه ویژگی‌ها یا خصوصیاتی است که با یکدیگر همراه هستند. به عبارت دیگر، این روش‌ها به دنبال استخراج قوانین به منظور کمی کردن ارتباط میان دو یا چند ویژگی هستند.

کشف توالی: کاوش اگوهای متوالی به معنای کشف حوادثی است که پی در پی اتفاق می‌افتند. به‌طور مثال احتمال خرید چاپگر رنگی توسط فردی که در همان ماه دوربین دیجیتال خریداری کرده است، بسیار زیاد است. این روش در بازاریابی، حفظ مشتری، پیش‌بینی هوا و بسیاری از صنایع دیگر کاربردهای زیادی دارد (هان و کامبر، ۲۰۰۶).

ب) کاربرد داده کاوی در حوزه های کاربردی پلیس

امروزه یکی از مهم‌ترین و اثربخش‌ترین ابزارها در رابطه با تحلیل و کشف دانش از اطلاعات و داده‌های پلیس، داده کاوی است. در رابطه با حوزه‌های مختلف پلیس، سه فعالیت مهم، شناسایی، پیش بینی و پیشگیری مطرح است:



شکل ۳: حوزه های کاربردی پلیس در رابطه با جرایم

برخی از این اقدامات قبل از وقوع جرم و برخی از آنها بعد از وقوع جرم صورت می‌گیرد. پیش‌بینی و پیش گیری جزء اقدامات قبل از وقوع جرم هستند؛ در حالی‌که شناسایی و کشف شواهد جرم پس از ارتکاب آن، در گروه اقدامات بعد از وقوع جرم به حساب می‌آیند. تحقیقات نشان می‌دهد، در سال‌های اخیر یکی از موضوعات مهم پژوهشگران بحث داده کاوی و روش‌های مختلف آن در نمونه‌سازی جرایم است. جدول (۱) برخی از تحقیقات صورت گرفته در این زمینه را نشان می‌دهد.

جدول ۱: برخی از تحقیقات مهم صورت گرفته در رابطه با داده کاوی و رابطه آن با جرایم

ردیف	موضوع پژوهش	منبع	نتایج
۱	الگویی مبتنی بر روش خوشه بندی برای شناسایی و گروه بندی انواع جرایم	(کارلیس و همکاران، ۲۰۰۷)	با توجه به این روش داده کاوی، جرایم را می توان در سطوح مختلفی شناسایی و تقسیم بندی کرد.
۲	پیشگیری جرایم با استفاده از روش رگرسیون	(مون و همکاران، ۲۰۱۰)	میان ساعات استفاده از رایانه و عضویت در گروه ها و شبکه های رایانه ای میزان جرایم رایانه ای را افزایش می دهد.
۳	معرفی داده کاوی به عنوان یکی از کاراترین ابزارها در جرایم رایانه ای (مروری بر روش های داده کاوی)	(چانگ و همکاران، ۲۰۰۶)	در مورد جرایم رایانه ای و مشکلات مربوطه بحث شده و پیشنهادهایی برای مقابله با این‌گونه جرایم مطرح شده است.
۴	الگوی پشتیبان تصمیم بر اساس روش «فازی سام[1]» برای تشخیص و تحلیل الگو ها و روندهای موجود در وقوع جرایم	(لی و همکاران، ۲۰۰۲)	نتایج به‌دست آمده برای مدیران نیروی پلیس در تدوین راهبردهای جلوگیری و پیشگیری از جرم و جنایت مفید است.
۵	الگویی برای پیش بینی محل جرم هفته آینده با توجه به داده های هفته فعلی	(لیو و همکاران، ۲۰۰۳)	در پیشگیری از جرم بسیار مفید است.
۶	ویژگی‌های جمعیت شناختی[2] و اخلاقی مجرمینی که دوباره مرتکب جرم شده اند	(کوراپسوگلو و همکاران، ۲۰۰۴)	هدف این تحقیق کشف ویژگی هایی بود که منجر به ارتکاب مجدد جرم می شود. بر اساس نتایج به‌دست آمده عصبانیت و خشمگینی، مصرف الکل، بیکاری، سطح تحصیلات و ... در ارتکاب جرم افراد مؤثر است.
۷	روش‌های داده کاوی برای پیش بینی و جلوگیری از جرایم شبکه‌های اجتماعی در محیط اینترنت	(خان و همکاران، ۲۰۰۸)	در این تحقیق با توجه به ماهیت اینترنت و شبکه های اجتماعی، روش های کاربردی داده کاوی برای پیش بینی و جلوگیری از جرایم در فضای مجازی ارائه شده است.
۸	روش‌های داده کاوی در حوزه کاربردی نیروی پلیس	(اوتلی و همکاران، ۲۰۰۳)	هدف اصلی این پژوهش کمک به نیروی پلیس در رسیدگی به جرایمی بود که با نرخ بالایی صورت می‌گرفت.
۹	پیش‌بینی تاریخ وقوع دزدی هایی که در یک روز تکرار می‌شود(با استفاده از روش سری های زمانی)	(دیدمن و همکاران، ۲۰۰۳)	با این روش می توان وقوع دزدی مشابهی در یک روز را پیش بینی کرد.
۱۰	آخرین گزارش سال ۲۰۱۰ در رابطه با جرایم سایبری(توسط HTCIA)	(تاد و کارول، ۲۰۱۰)	این گزارش شامل آخرین اطلاعات در خصوص انواع جرایم و روش های برخورد با آن است.
۱۱	ده الگوریتم کاربردی در داده کاوی	(ژین دانگ و همکاران، ۲۰۰۷)	شامل الگوریتم های *SVM ،k-Means ،C4.5 ،Apriori ،EM ،PageRank ،AdaBoost ،kNN* و *Naive Bayes* ،*CART*

1. Fussy Som (FSOM)
2. Demographic

کاظمی و حسین پور(۱۳۸۸)، در تحقیق خود، داده کاوی را در برخی سازمان‌های پلیسی و قضایی، به منظور پیش بینی و پیشگیری از وقوع جرم مطرح کرده اند:

۱. داده کاوی و تحلیل حوادث ویژه؛

۲. داده کاوی و بررسی صحنه وقوع جرم؛

۳. داده کاوی و دوباره قربانی شدن؛

۴. داده کاوی و حوادث تیر اندازی؛

۵. داده کاوی و سرقت های مسلحانه؛

۶. داده کاوی و جرایم خشونت آمیز؛

۷. داده کاوی و حملات تروریستی؛

۸. داده کاوی و سرقت از منازل؛

۹. داده کاوی و جرایم مجازی.

لازم به توضیح است علاوه بر روش‌های داده کاوی، یکی از فناوری‌های سیستم‌های اطلاعاتی که در این حوزه مورد توجه قرار گرفته است، سیستم‌های اطلاعات جغرافیایی است. نتایج نشان می‌دهد روش‌های پیش‌بینی بیشتر از سایر ابزارهای داده کاوی، به منظور پیش بینی ارتکاب جرم و عوامل و پارامترهای تأثیرگذار در آن، مورد استفاده قرار گرفته است.

متن کاوی

بهترین توصیف از داده کاوی به‌وسیله اجتماع آمار، هوش مصنوعی و یادگیری ماشین به‌دست می‌آید. این روش‌ها با کمک یکدیگر، برای مطالعه داده و پیدا کردن الگوهای نهفته در آنها استفاده می‌شوند. از این‌رو مهم‌ترین کاربردهای داده کاوی شامل موارد مندرج در شکل (۴) است.

شکل ۴: کاربردهای داده کاوی

همان‌طور که در شکل(۴) مشخص است، یکی از کاربردهای داده کاوی، متن کاوی است که در این روش پالایش متن بر روی نامه های الکترونیکی، گروه‌های خبری و ... انجام می گیرد.

امروزه حجم وسیعی از دانش ما به صورت متن، مستندات و دیگر صورت‌های رسانه‌ای نگهداری می شوند که همه آنها به‌صورت غیرساختاریافته می‌باشند. برای دریافت دانش از اطلاعات یک متن، لازم است ابتدا آن را درک کرد، سپس پردازش کرد تا فهمید چه معانی و مفاهیمی در آن موجود است؛ چه ارتباطی میان مفاهیم وجود دارد و از میان این مفاهیم کدام جدید است و کدام قدیمی؛ از این‌رو در عصر فناوری، هر چیزی باید بتواند به‌صورت خودکار، انجام شود. «درک معنی متون» نیز از این جمله کارها محسوب می‌شود. «متن کاوی»، «کاوش داده های متنی (گریور و همکاران، ۲۰۰۴)» و یا «کشف دانش در متن» یا KDT، از نام‌های مورد قبول دراین زمینه هستند.

اینترنت به‌عنوان بزرگ‌ترین منبع اطلاعاتی همگانی، تشکیل یافته از صدها میلیون صفحه اطلاعات است که به جهت همگانی بودن آن و نبود آینده‌نگری کافی در زمان تشکیل و رشد آن، متحمل نگهداری اطلاعات نویسندگان، محققان، اندیشمندان و... به همان نحوی که آنها می‌نوشتند شد. نبود یک استاندارد همه جانبه و دقیق در تنظیم متون و قرار گیری این مجموعه عظیم به صورتی غیرساختاریافته و یا بعضاً نیمه ساختاریافته،

جامعه اطلاعاتی را دچار نوعی سردر گمی و مشکل در دستیابی به اطلاعات مورد نیاز کرده؛ به‌طوری‌که برای یافتن مطالب مورد نظر خود متحمل هزینه‌های زمانی بسیاری می‌شوند. محققان به ارایه راه کارهایی برای ساختار یافته کردن اطلاعات پرداختند و با ارایه زبان‌های نشانه گذاری استاندارد نظیر XML تا حد زیادی جلوی این از هم پاشیدگی اطلاعاتی را گرفتند؛ اما آنچه همچنان باقی است، وجود بسیاری از متون غیرساختاریافته است. در همین راستا ارایه ابزارهایی که با بررسی متون بتوانند تحلیلی روی آنها انجام دهند، منجر به شکل‌گیری زمینه‌ای جدید در هوش مصنوعی و فناوری اطلاعات شده که به یادگیری متن معروف است.

این حوزه، تمام فعالیت‌هایی که به نوعی به دنبال کسب دانش از متن هستند را شامل می‌شود. آنالیز داده‌های متنی توسط فنون یادگیری ماشین، بازیابی اطلاعات هوشمند، پردازش زبان طبیعی یا روش‌های مرتبط دیگر همگی در زمره مقوله یادگیری متن قرار می‌گیرند. یکی از روش‌هایی که ذکر شد، استفاده از فنون یادگیری ماشین در زمینه پردازش متن است. مسئله قابل تأمل این است که این روش‌ها، در ابتدا در مورد داده‌های ساختاریافته به کار گرفته شدند و علمی به نام داده کاوی را به‌وجود آوردند. داده های ساختاریافته به داده‌هایی اطلاق می‌شود که به‌طور کاملاً مستقل از همدیگر ولی یکسان از لحاظ ساختاری در یک محل گردآوری شده‌اند. انواع بانک‌های اطلاعاتی را می‌توان به عنوان نمونه‌هایی از این دسته اطلاعات نام برد. در این‌صورت مسئله داده کاوی عبارت از کسب اطلاعات و دانش از این مجموعه ساخت یافته؛ اما در مورد متون که عمدتاً غیرساختاریافته یا نیمه ساخت‌یافته هستند؛ ابتدا باید توسط روش‌هایی، آنها را ساختارمند کرد و سپس از این روش‌ها برای استخراج اطلاعات و دانش از آنها استفاده کرد. به هر حال استفاده از داده کاوی در مورد متن خود شاخه‌ای دیگر را در علوم هوش مصنوعی به‌وجود آورد به نام متن کاوی. از جمله فعالیت‌های بسیار مهم در این زمینه، طبقه‌بندی (دسته بندی) متن است.

طبقه‌بندی متن؛ یعنی انتساب اسناد متنی بر اساس محتوی، به یک یا چند طبقه از قبل تعیین شده می‌باشد. یکی از مهم‌ترین مسایل در متن کاوی؛ مرتب کردن بلادرنگ نامه‌های الکترونیکی یا فایل‌ها در سلسله مراتبی از پوشه‌ها، تشخیص موضوع متن، جستجوی ساختاریافته و یا پیدا کردن اسنادی در راستای علایق کاربر است، از جمله کاربردهای مبحث طبقه‌بندی (دسته‌بندی-کلاسه بندی) متن است. در بسیاری از موارد، افراد حرفه ای آموزش دیده، برای طبقه بندی متون جدید به کار گرفته می‌شوند. این فرآیند بسیار زمان بر و پر هزینه است؛ لذا کاربرد خود را محدود می‌سازد، به همین منظور علاقه روزافزونی به توسعه فناوری‌هایی در دسته بندی خودکار متن ابراز می‌شود.

رشد تصاعدی در اطلاعات دیجیتالی منتشر شده در سطح اینترنت و فضای سایبر، سازماندهی اطلاعات متنی را به امری مهم تبدیل کرده است. با توجه به اینکه در ماهیت اطلاعات از نظر میزان حجم، دسترسی پذیری و اهمیت، تغییر ایجاد شده است، پیشرفت فناوری اطلاعات نیز به موازات رشد اطلاعات، نوید تصمیم گیری پیشرفته را داده است.

با توجه به اینکه با استفاده از متن کاوی می‌توان از مقادیر عظیمی از داده‌های متنی خام که در پایگاه داده‌های بزرگ ذخیره شده‌اند، اطلاعات ارزشمند و نهفته‌ای را استخراج کرد؛ متن کاوی نیز مانند داده کاوی رشد مستمر و پویایی داشته است (شرما، ۲۰۰۵).

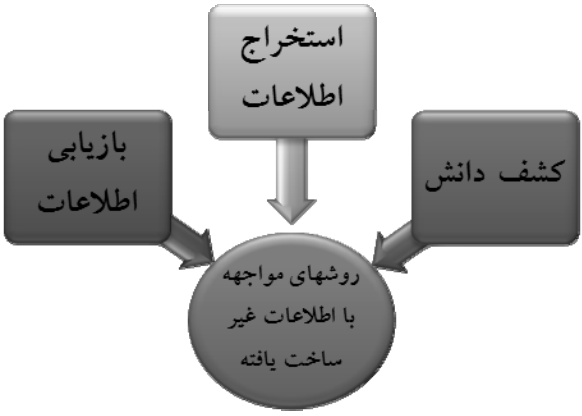
در حقیقت متن کاوی، به تحلیل هوشمند متن، داده کاوی متنی یا کشف دانش در متن معروف است و به فرایند استخراج دانش و اطلاعات مورد علاقه و مهم از مجموعه متنی غیرساختاریافته اشاره دارد. به عبارت بهتر از طریق متن کاوی می‌توان، فرایند تحلیل متون را به منظور کشف و ثبت اطلاعات معنا دار در یک ساختار سازمان دانش، به شکلی بهینه انجام داد.

متن کاوی، در فناوری‌های متفاوتی ریشه دارد و از این‌رو این تعجب انگیز نیست

که تعاریف زیادی نیز برای آن وجود دارد. داده کاوی، یک روش بسیار کاراست برای کشف اطلاعات از داده های ساخت یافته ای که در جداول نگهداری می‌شوند. داده کاوی، الگوها را از تراکنش‌ها، استخراج می‌کند (کاراتوف، ۱۹۹۹)، داده را گروه بندی، و آن‌را دسته بندی می‌کند. به‌وسیله داده کاوی می‌توان به وجود روابطی میان اقلام داده‌ای که پایگاه داده را پر کرده اند، پی برد. کتابخانه‌های دیجیتال، اخبار، کتاب‌های الکترونیکی، بسیاری از مدارک مالی، مقالات علمی و تقریباً هر چیزی که بتوان در داخل فضای سایبر و شبکه‌های اجتماعی یافت، ساختاریافته نیستند. در نتیجه نمی‌توان آموزه‌های داده کاوی را در مورد آنها به طور مستقیم استفاده کرد. با این حال، سه روش اساسی در مواجهه با این حجم وسیع از اطلاعات غیرساختاریافته گسترده شده در جهان وجود دارد.

- بازیابی اطلاعات (راجمن، ۱۹۹۷)؛
- استخراج اطلاعات (اون یانگ نام، ۲۰۰۱)؛
- کشف دانش در متن.

این سه روش برخورد با این مسئله هستند.



شکل ۵: روش‌های مواجهه با اطلاعات غیر ساختاریافته

در حقیقت متن کاوی، کشف حقایق و الگوها به‌وسیله اطلاعات ناشناخته قبلی و جدید شامل استخراج خودکار اطلاعات از منابع مختلف و سایر ادله های الکترونیکی است.

متن کاوی، متفاوت از مفاهیم وب کاوی است. در وب کاوی، به جستجوی چیزی پرداخته می شود که شناخته شده است و به‌وسیله کس دیگری نوشته شده است؛ اما متن کاوی، هدف کشف اطلاعات ناشناخته است، چیزی که کسی نمی داند و تا کنون ثبت نشده است.

متن کاوی به سازمان ها کمک می‌کند: محتوای «پنهان» اسناد را بیابند؛ از جمله روابط مفید اضافی.

شکل ۶: متن کاوی

در متن کاوی برای ساختاردهی به داده‌های بدون ساختار، ابتدا شاخص‌های عددی معناداری را از متون بدون ساختار استخراج می‌کنند و سپس این شاخص‌ها را با استفاده از الگوریتم‌های داده کاوی پردازش می‌کنند. با استفاده از این شاخص‌های عددی می‌توان وظایف زیر را انجام داد:

- جمع بندی و خلاصه‌سازی مستندات بر اساس مفاهیم کلیدی؛
- خوشه بندی مستندات بر اساس مفاهیم مشابه و موضوعات مشترک؛
- تعیین روابط بین مستندات و استخراج محتوای پنهان مستندات.

الف) کاربردهای متن‌کاوی

با توجه به اینکه تعاریف گسترده ای از متن‌کاوی وجود دارد، در نتیجه کاربردهای متن کاوی نیز متنوع است (راجمن، ۱۹۹۷):

- جستجو و بازیابی؛
- گروه بندی[1] (دسته بندی بدون نظارت[2]) و طبقه بندی [3](دسته بندی بانظارت)؛

1. Clustering
2. Unsupervised Classification
3. Categorization

- خلاصه سازی؛
- استخراج روابط؛
- یافتن و تحلیل روند[1]ها؛
- برچسب زدن نحوی[2]؛
- ساخت اتوماتیک هستی شناسی[3] و فرهنگ جامع[4]؛
- شناسایی و تشخیص جرایم؛
- و ... .

همان‌طور که مشخص است از روش متن کاوی می‌توان در شناسایی و تشخیص جرایم استفاده کرد. بیشترین کاربرد متن کاوی در زمان قبل از وقوع جرم به منظور پیش‌بینی و پیش‌گیری از جرایم است.

ب) متن‌کاوی و شیوه‌های تشخیص و جلوگیری از جرم

در چند سال اخیر، نگرانی برای امنیت ملی و ممانعت از وقوع جرم به شدت افزایش یافته؛ به‌طوری که خیلی از کشورها شبکه‌ها و طرح‌های تحقیقاتی را با هدف مبارزه با جرایم سازماندهی شده، فعال کرده‌اند. آژانس طرح تحقیق پیشرفته دفاعی[5] در ایالات متحده آمریکا، یک برنامه امنیت داخلی با نام «آگاهی کامل اطلاعاتی» را ارایه کرد که با فناوری‌هایی مثل ادغام داده‌ها، تحقیقات داده‌ای، بیومتریک و شناسایی الگوها سر و کار دارد. این برنامه در جستجوی گسترش شبکه‌ای از فناوری‌هاست که به افسران امنیت در پیش‌بینی و ممانعت از اقدام تروریستی کمک می‌کند (کنیون، ۲۰۰۳). شورای تحقیقات مهندسی و علوم جسمی در بریتانیا که ادعا می‌کند «جرم» حدود پنجاه بیلیون یورو را در سال به خودش اختصاص می‌دهد، با انجام طرح‌های تحقیقی مرتبط با فناوری‌های ممانعت و تشخیص جرم به اجرای برنامهٔ فناوری جرم اقدام کرده است.

---
1. Trend
2. Part of Speech tagging
3. Ontology
4. Thesaurus
5. DARPA

امروزه بسیاری از سازمان‌های امنیتی، برای تشخیص و پیش بینی عمل جرم، به شیوه‌های استخراج داده و متن وابسته هستند. اگرچه استخراج داده برای ارایهٔ یک الگوی با مفهوم و قوانین (بری و لینوف، ۱۹۹۷) به کشف و تحلیل تعداد زیادی از داده‌ها اشاره دارد؛ اما استخراج متن، فرآیند تحلیلی است که به طور طبیعی با هدف استخراج الگوهای مورد نظر و با اهمیت، یا اطلاعاتی از متون غیر ساختاری اتفاق می‌افتد (هرست، ۱۹۹۷). هم استخراج داده و هم استخراج متن، اغلب به عنوان زیر فرآیند عرصه کشف آگاهی و دانش در نظر گرفته می‌شود. تا زمان‌های اخیر، مهم‌ترین کاربرد استخراج داده در پیش‌بینی ارجاعات مصرف کننده و طرح دورنما برای فرآورده‌ها و خدمات بوده است. پس از حملات تروریستی اخیر، استخراج داده و متن، یکی از شایع‌ترین شیوه‌ها، در تعداد رو به افزایش، طرح‌های تحقیقی مرتبط با جرایم سازماندهی شده و مخصوصاً فعالیت های مرتبط با ضد تروریست شده است. هدف بسیاری از طرح های تحلیل داده، استفاده از استخراج داده برای پیدا کردن همدستان و کشف در بین ماهیت‌های مشکوک بر اساس داده‌های تاریخی(قدیمی) است. در هنگامی که استخراج داده، داده ها را بر اساس ساختار داده‌ای تحلیل می‌کند، حجم عظیمی از اطلاعات متنی از قبیل پست الکترونیکی، مکالمات تلفنی، پیام‌های متنی و ... وجود دارد که بازرسان جرم باید آنها را بررسی کنند. پاپ[1]، آرمور[2]، سناتور[3] و نومریک[4] (۲۰۰۴) بیان کرده‌اند که تحلیل‌گران جاسوسی، برای تحقیق و فرآیندسازی داده برای تحلیل، نتیجه گیری برای ارایهٔ گزارش، زمان زیادی را صرف می کنند؛ بنابراین برای تحلیل داده‌های متنی زمان کمی در اختیار آنها خواهد بود. شیوه‌های پیشرفته شناسایی فعالیت های مشکوک، روابط بین عناصر، انسان‌ها، سازمان‌ها و ماجراها را کشف کرده و الگوهای رفتاری را که در تشخیص فعالیت‌های جرایم سازماندهی شده از اسناد و شواهد موجود می‌تواند به تحلیل گر کمک کند، در اختیار او

1. Popp
2. Armour
3. Senator
4. Numrych

قرار می‌دهد. در سال‌های اخیر، سیستم های نرم افزاری استخراج داده و متن از رشد مناسبی برخوردار بوده است که فهرستی از آنها در جدول (۲) نشان داده شده است.

جدول ۲: داده و متن کاوی

نام شرکت	فناوری	روش	هدف	کشور مورد استفاده
HNC Falcon System	شبکه های عصبی	داده کاوی	A	آمریکا، بریتانیا، کانادا، آفریقای جنوبی
DARPA	فن آوری عامل و حسگر ها[1]	داده کاوی، متن کاوی	B	آمریکا
ATAC	قوانین مرتبط[3]	تحلیل متصل[2]	A	آمریکا
Autonomy	الگوی مفهومی سازگار احتمالی[4]، شبکه های عصبی، و عوامل مفهومی[5]	متن کاوی	C	بریتانیا، آمریکا
Dolphin Search	شبکه های عصبی	داده کاوی	A	آمریکا
Wolverhampton University and W. Midlands Police(UK)	شبکه های عصبی-نقشه های کوهونن[6]- رهبری نظارت نشده[7]	داده کاوی	S	بریتانیا (پلیس)، آمریکا
ChoicePoint (AutoTrackXP)	زبان پرس و جوی ساخت یافته (SQL)	تحلیل متصل	D	آمریکا
ALTA Analytics(NETMAP)	جهت دهی[8]	تحلیل متصل	A	آمریکا
Bair Software Inc.	شبکه های عصبی	داده کاوی	G	آمریکا
I2 Ltd.	قوانین مرتبط	تحلیل متصل	G,P	بریتانیا ، امریکا
Crime Link	قوانین مرتبط	تحلیل متصل	G	بریتانیا، آمریکا
Memex(Crime Workbench)	سامانه مدیریت هوشمند پایگاه داده[10]	موتور جستجوی پیشرفته[9]	C	بریتانیا (پلیس)، آمریکا (بخش دفاع)
Clearforest	طبقه بندی	طبقه بندی	G	آمریکا(FBI)

1. Agent Technology and Sensors
2. Link Analysis
3. Association Rules
4. Adaptive Probabilistic Concept Modelling(APCM)
5. Concept Agent
6. Kohonen Maps(SOMs)
7. Unsupervised learning
8. Vectorization
9. Enhanced Search engine
10. Database Intelligent management system

بریتانیا	F	عاملان و متن کاوی	بازیافت اطلاعات/ مفاهیم[1]	Copernic
آمریکا (بخش امنیت داخلی)	T,P	ابزار متن کاوی، تحلیل متصل	شناسایی الگوها و روابط[2]	Quenza(Xanalyst)
آمریکا (پلیس)	C	تحلیل متصل، متن کاوی	هماهنگی الگو / داده[3]	COPLINK
بریتانیا (پلیس)	E,G	تحلیل متصل	نظریه های احتمالی پایگاه داده دانش قدیمی[4]	FLINTS

A- کلاهبرداری و سرقت پول؛

B – تروریسم زیستی؛

C – دسته بندی و طبقه بندی طریقه اجرا و دیگر اطلاعات مرتبط بـه جـرم در پایگـاه داده؛

D – افزایش مواد مخدر، مهاجرت و طبیعت گرایی؛

E – استخراج داده و پیش بینی حملات آینده؛

F– بازرسی قانونی؛

G – تحلیل عمومی متصل برای تصویرسازی جرم و شبکه جرم و جنایت؛

P – جرایم مرتبط با نقاشی های جنسی؛

S – جرایم مرتبط با مسائل جنسی و قتل؛

T – جستجوی گروه های مجرم؛

O – سایر (که می تواند شامل استخراج ، تحقیق و تصویرسازی آینده باشد).

متن کاوی، روش‌های داده کاوی مثل دسته بندی، قوانین مرتبط، و شبکه عصـبی را بـا شیوه‌هایی که در استخراج اطلاعات و روش‌های طبیعی زبان از آنها استفاده می‌شود، بـا هم ترکیب می‌کند. یک توافق عمومی وجود دارد که بر اساس آن چـارچوب اکتشـاف

1. Information/concept retrieval
2. Pattern recognition and relationships
3. Pattern / data matching
4. Probability theories Previous database-Knowledge

دانش عمومی، هم می‌تواند در استخراج داده و هم در استخراج متن به کار گرفته شود، که شامل سه مرحلهٔ اصلی است: پیش پردازی، اکتشاف، و پس پردازی (فیاد و همکاران، ۱۹۹۶؛ آهونن و همکاران، ۱۹۹۷).

مرحلهٔ پیش پردازی، بر اساس هدف فعالیت استخراج متن می‌تواند با سطوح متفاوتی از تحلیل ها سر و کار داشته باشد: ممکن است برخی از طرح‌ها به تحلیل واژه و معنا اهمیت دهند و برخی دیگر شامل تحلیل گرامری و دانش حاکم باشند. در برخی از طرح‌های اخیر، دو مرحلهٔ آخر با هم ترکیب شده‌اند. در حقیقت مرحله تقویت متن را که در آن متن خام به شکل ابتدایی اش تبدیل می‌شود و مرحلهٔ عصاره کشی متن که متن تبدیل شده را به منظور کشف الگوها و روابط معنایی در بین ماهیت‌ها، تحلیل می‌کند، پیشنهاد می‌کند. در حالی‌که، دسته‌بندی، فهرست‌بندی و تصویرسازی به طور خاصی به تقویت اجزای متن وابسته هستند، عصاره کشی متن با استفاده از شیوه‌های الگوسازی آشنا مثل دسته بندی، قوانین مربوطه و طبقه بندی، الگوهایی را از شکل ابتدایی بیرون می‌کشد. در زمان ایجاد یک شکل ابتدایی، ممکن است کاربردهای متفاوت به سطوح طبقه‌بندی شده متفاوت نیاز داشته باشند؛ بنابراین برای دستیابی به روابط بین ماهیت‌ها یا مفاهیم توصیف شده در متن، ممکن است گاهی به تحلیل معنایی عمیقی نیاز شود. در تقویت متن و مراحل عصاره کشی، حوزه آگاهی نقش مهمی را بازی می‌کند.

اگرچه در سال‌های اخیر علاقه به استخراج داده در کاربردهایی مثل جلوگیری از جرم و تشخیص آن افزایش یافته است؛ اما استخراج متن، دانش نسبتاً جوانی است که در عرصه جلوگیری از جرم و شناسایی آن، جنبش باشتابی پیدا کرده است.

استخراج اطلاعات معمولاً در مرحلهٔ پیش پردازش اتفاق می‌افتد. هدف آن استخراج واژه‌های با مفهوم، از متن است که بعدها می توان به آن بیشتر توجه کرد. با این حال، بسیاری از شیوه‌هایی که از استخراج اطلاعات استفاده می‌کنند به سمت پیدا کردن دسته خاصی از وقایع گرایش دارند. از شیوه‌های استخراج اطلاعات برای

شناسایی خودکار ماهیت‌هایی مثل آدرس، وسیلهٔ نقلیه، شماره تلفن مشکوک و موارد مشکوک دیگری که پلیس آنها را گزارش داده یا رایانامه فرستاده است، استفاده می‌شود. چن[1] و دیگران (۲۰۰۴) سامانه‌ای را برای استخراج ماهیت‌های اسمی از گزارش پلیس ارایه کردند که قوائد زبانی را بر اساس تطبیق الگوها و واژگان با شبکهٔ عصبی ترکیب می‌کرد.

برای تحلیل داده‌ها و متون مرتبط با جرم از فناوری‌های متفاوتی استفاده می‌شود که هدف برخی از آنها اشاعه نمونه توصیفی است؛ در حالی‌که برخی دیگر بر روی الگوی پیشگویانه تمرکز می‌کنند. با وجود این، بیشتر کار بر روی تحلیل و الگوبرداری از داده‌هاست و تأکید کمی به تحلیل جرم مرتبط در تصمیم‌گیری می‌شود. پروژه تحقیقی کارولاین و دیگران[2](۲۰۰۶) با نام «اسکری[3]» برای تسهیل استفاده از داده‌های ساختاری و غیرساختاری متنی مرتبط با جرم مثل نامه‌های الکترونیکی، پیام‌های متنی، گزارش مکالمات تلفنی و منابع متنی مرتبط دیگر طراحی شده است. شیوه پیشنهادی فناوری عامل را با شیوه استخراج متن ترکیب می کند تا الگوهای جرم را به شکل پویایی استخراج کرده، ارتباط بین اعمال مشکوک را کشف کرده و آنها را در منابع اسنادی چندگانه جستجو می‌کند. هدف پروژه «اسکری» پشتیبانی از بازرسان جرم و جنایت و تحلیل گران در تصمیم گیری شان، به‌وسیلهٔ پیش‌بینی و جلوگیری از وقوع جرایم آینده است.

نتیجه‌گیری

داده کاوی فرایند استخراج اطلاعات معتبر، ناشناخته و قابل درک از پایگاه‌های بزرگ به منظور بهبود و ارتقای تصمیمات سازمان است. اصطلاح کشف دانش در چایگاه‌های اطلاعاتی بیانگر کل فرایند تبدیل داده‌های سطح پایین به دانش سطح بالا است. در

1. Chen et al,2004
2. Caroline et al.,2006
3. ASKARI

حقیقت داده کاوی به عنوان مهم‌ترین مرحله فرایند کشف دانش معرفی شده است که روش‌های آن، شامل دو دسته توصیفی و پیش‌بینانه است. در روش‌های توصیفی، هدف توصیف یک رویداد با یک واقعیت است؛ اما در روش‌های پیش‌بینی، هدف پیش‌بینی متغیر ناشناخته از داده‌های آتی است.

با توجه به اینکه یکی از مهم‌ترین و اثربخش‌ترین ابزارها در رابطه با تحلیل و کشف دانش از اطلاعات و داده‌های پلیس، داده کاوی است؛ از این‌رو در رابطه با حوزه‌های مختلف پلیس، سه فعالیت مهم، شناسایی، پیش بینی و پیشگیری مطرح است که برخی از این اقدامات قبل از وقوع جرم و برخی از آنها بعد از وقوع جرم هستند. پیش‌بینی و پیش‌گیری، جزء اقدامات قبل از وقوع جرم هستند؛ در حالی‌که شناسایی و کشف شواهد جرم پس از ارتکاب آن، در گروه اقدامات بعد از وقوع جرم به حساب می‌آیند. تحقیقات نشان می‌دهد، در سال‌های اخیر یکی از موضوعات مهم پژوهشگران بحث داده کاوی و روش‌های مختلف آن در موضوعاتی از قبیل بررسی صحنه وقوع جرم، دوباره قربانی شدن، حوادث تیراندازی، سرقت‌های مسلحانه، جرایم خشونت آمیز، حملات تروریستی، سرقت از منازل، جرایم مجازی و ... است.

یکی از کاربردهای داده کاوی، متن کاوی است که در این روش پالایش متن بر روی نامه‌های الکترونیکی، گروه‌های خبری و ... انجام می‌گیرد. در حقیقت متن کاوی، به تحلیل هوشمند متن، داده کاوی متنی یا کشف دانش در متن معروف است و به فرایند استخراج دانش و اطلاعات مورد علاقه و مهم از مجموعه متنی غیرساختاریافته اشاره دارد. سه روش اساسی در مواجهه با این حجم وسیع از اطلاعات غیرساخت‌یافته گسترده شده در جهان وجود دارد: بازیابی اطلاعات، استخراج اطلاعات و کشف دانش.

کاربردهای متن کاوی متنوع است؛ اما در این پژوهش شناسایی و تشخیص جرایم به عنوان یکی از مهم‌ترین کاربردهای متن کاوی مطرح شده است. بیشترین کاربرد متن کاوی در زمان قبل از وقوع جرم به منظور پیش بینی و پیش‌گیری از جرایم است.

امروزه محققان با استفاده از فناوری‌های جدید و همچنین به‌کارگیری روش‌های

مبتنی بر متن کاوی، توانسته‌اند به موضوعاتی مرتبط با جرایم از قبیل کلاهبرداری و سرقت پول، تروریسم زیستی، دسته بندی و طبقه بندی طریقه اجرا و دیگر اطلاعات مرتبط به جرم در پایگاه داده، استخراج داده و پیش بینی حملات آینده، بازرسی قانونی، تحلیل عمومی متصل برای تصویرسازی جرم و شبکه جرم و جنایت، جرایم مرتبط با نقاشی‌های جنسی، جرایم مرتبط با مسائل جنسی و قتل، جستجوی گروه‌های مجرم و ... دست یابند تا از طریق الگوهای استخراج شده به پیش‌بینی و پیشگیری از جرایم بپردازند. از این‌رو با بومی کردن اینگونه فناوری‌ها و روش‌های متن کاوی در ناجا، می‌توان با وجود بانک‌های اطلاعاتی فراوان و با بهره‌گیری از یک نرم افزار مناسب متن کاوی در جهت کاوش دانش دلخواه، در راستای افزایش توان اطلاعاتی و پیشگیری از جرایم استفاده کرد.

منابع

- کاظمی. پ، حسین پور.ج، (۱۳۸۸)، **کاربرد داده کاوی در سازمان پلیسی و قضایی به منظور شناسایی الگوهای جرم و کشف جرایم**، مجله علمی ترویجی کارگاه، سال دوم، شماره ۸، ص ۶۳-۳۲.

*- Han,j. & Kamber,M.(2006), **Data Mining: Concepts and Techniques**, Second Edition, Morgan Kaufman Publisher.*

*- Tan, P.N., Steinbach, M. & Kumar, V. (2006), **Introduction to Data Mining**, Boston, pearson.*

*- Karlis, D. & Meligkotsidou, L. (2007), **Finite Mixtures of Multivariate Poisson distributions With Application**, Journal of Statistical Planning and Inference, No137, PP. 1942-1960.*

*- Moon, B., McCluskey, J.B. & McCluskey, C.P. (2010), **General Theory of Crime and Computer Crime: an Empirical Test**, Journal o Crimial Justice, No. 38, PP.767-772.*

*- Chung, W., Chen, H., Ch., Chang, W. & Chou, SH. (2006),**Fighting Cybercrime:A Review and the Taiwan Experience**, Decision Support Systems, No. 41, PP. 669-682.*

*- Corapcıoglu, A. & Erdogan, S. (2004), **A Cross-Sectional Study on Expression of Anger and Factors Associated With Criminal Recidivism in Prisoners With Prior Offences**, Forensic Science International, No.*

140, PP167-174 .

- Cutts, B.B., Darby, K.J., Boone, CH.G. & Brewis, A. (2009), ***City Structure , Obesity, and Environmental Justice: An Integrated Analysis of Physical and Social Barriers to Walkable Streets and Park Access****, Social Science & Medicine, No. 69 , ,PP. 1314-1322.*

- Khan, J.I. & Shaikh, S.S. (2008), ***Computing in Social Networks With Relationship Algebra****, Journal of Network and Computer Applications, No. 31, PP. 862-878.*

- Li, SH.T., Kuo, SH.CH. & Tsai, F.CH. (2010), ***An Intelligent Decision-Support Model Using FSOM and Rule Extraction for Crime Prevention****, Expert Systems with Applications, No. ۳۷, PP. 7108-7119 .*

–Liu, H. & Brown, Donald E. (2003), ***Criminal Incident Prediction Using a Point–Pattern-Based Density Model****, International Journal of Forecasting, No. 19, PP603-622.*

- Oatley, G.C. & Ewart, B.W. (2003), ***Crimes Analysis Software: 'Pins in Maps , Clustering and Bayes Net Prediction****, Expert Systems with Applications, No. 25 „ PP. 569-588.*

- Ozkan, K. (2004), ***Managing Data Mining at Digital Crime Investigation,*** *Forensic Science International, No. 146, PP. S37–S38 .*

- Deadman, D. (2003), ***Forecasting Residential Burglary****, International Journal of Forecasting, No. 19, PP. 567-578.*

- Claire Grover, Harry Halpin, Ewan Klein, Jochen L. Leidner, Stephen Potter, Sebastian Riedel, Sally Scrutchin, and Richard Tobin.(2004), ***A framework for text mining services****. In Proceedings of the Third UK e-Science Programme All Hands Meeting.*

*- Kodratoff Y.,(1999), "****Knowledge Discovery in Texts****: A Definition, and Applications," in Foundation of Intelligent Systems, Ras & Skowron (Eds.) LNAI 1609, Springer.*

- M. Rajman.(1997), ***Text Mining, knowledge extraction from unstructured textual data****. Proc. of EUROSTAT Conference, Francfort (Deutchland), may.*

- Un Yang Nahm(2004), ***Text Mining with Information Extraction****, PhD Proposal, The University of Texas at Austin.*

- Xindong. Wu, Vipin Kumar. J, Ross Quinlan. Joydeep ghosh, (2007), ***Top 10 algorithms in data mining****, Springer.*