APJ. 4(2):1-10, Summer 2023

Arman Process Journal (APJ)

Homepage: armanprocessjournal.ir

Data Mining Operationalizing Process, Applications and Tools

***M. Rezaei*[*,1]**

[1] Department of Computer Science, Shahid Beheshti University, Tehran, Iran

Received: 19 June 2023
Accepted: 29 August 2023

KEYWORDS:
Data Analysis,
Data Management,
Analysis,
Data Repository,
Pattern Recognition,

[1] Corresponding author
mo.rezaei@yahoo.com

ABSTRACT

In today's competitive world, information has emerged as one of the important production factors. As a result, the effort to extract information from data has attracted the attention of many people involved in the information industry and related fields. The large volume of data is constantly growing in all fields and the vast difference in data production process has increased the complexity of information management and extraction. Recently, several strategies and techniques have been used to collect, store, organize and efficiently manage existing data and achieve meaningful results, and data mining is one of the recent developments in the direction of data management technologies. The term data mining refers to the semi-automatic process of analyzing large databases and data warehouses in order to find useful and applicabale patterns. In this research, we are going to examine the operationalization process of data mining and do a practical analysis of this issue. In addition, we will research about the important applications and tools of this field.

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
27	1	0

نشریه تخصصی آرمان پردازش، دوره ۴، شماره ۲، تابستان ۱۴۰۲

فصلنامه تخصصی آرمان پردازش
(APJ)
Homepage: www.armanprocessjournal.ir

فرایند عملیاتی سازی داده کاوی، کاربردها و ابزارها

مرتضی رضائی[*،1]

[1] دانشکده مهندسی و علوم کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران

چکیده

در دنیای رقابت محور امروز، اطلاعات بعنوان یکی از فاکتورهای تولیدی مهم پدیدار شده ا ست. در نتیجه تلاش برای استخراج اطلاعات از داده ها توجه بسیاری از افراد دخیل در صنعت اطلاعات و حوزه های وابسته را به خود جلب نموده اســت. حجم بالای داده های دائما در حال رشــد در همه حوزه ها می باشــد و تفاوت وســیع در فرآیندهای تولید داده پیچیدگی مدیریت و اســتخراج اطلاعات را افزایش داده اســت. اخیرا اســتراتژیها و فنون متعددی برای گردآوری، ذخیره، ســازماندهی و مدیریت کارآمد داده های موجود و رســیدن به نتایج معنی دار بکار گرفته شده انداده کاوی یکی از پی شرفتهای اخیر در را ستای فن آوریهای مدیریت داده ها ست. ا صطلاح داده کاوی به فرایند نیم خودکار تجزیه و تحلیل پایگاه داده های بزرگ به منظور یافتن الگوهای مفید اطلاق می شود. در این تحقیق قصد داریم فرایند عملیاتی سازی داده کاوی را برر سی نموده و به تحلیل دقیق و کاربردی این موضوع بپردازیم. بعلاوه درخصوص کاربردها و ابزارهای مهم این حوزه تحقیق نمائیم.

واژگان کلیدی:
داده کاوی،
مدیریت داده،
تحلیل،
مخزن داده،
تشخیص الگو،

تعداد جداول	تعداد شکل ها	تعداد مراجع
۰	۱	۲۷

مقدمه

امروزه با پیشرفت فناوری و حضور گسترده‌ی آن در زندگی روزمره‌مان شاهد کاربرد پررنگ داده و اطلاعات هستیم. یکی از روش‌های استخراج اطلاعات از داده‌های خام داده کاوی[1] می‌باشد، که در ادامه این مقاله ابعاد مهم و کاربردی ان را بررسی می نمائیم. اساسا به فرایند استخراج و کشف همبستگی‌ها و الگوهای مفید از میان حجم زیادی از داده‌های خام که با استفاده از الگوریتم و سازوکارهای هوشمند انجام می‌گیرد دیتاماینینگ یا داده کاوی می‌گویند. به زبان ساده‌تر، استخراج دانش از میان مجموعه‌ای از داده‌ها را داده‌کاوی می‌نامند [۱]. البته لازم به ذکر است، برای اینکه این الگوریتم بتواند دانش را به خوبی استخراج کند نیاز به یک‌سری پیش‌پردازش بر روی داده‌های اولیه و همچنین یک‌سری پس‌پردازش بر روی اطلاعات خروجی خواهد داشت. با پیدایش کامپیوتر ، پایگاه داده بزرگ و اینترنت ، آسانتر می توان میلیون ، بیلیون و حتی تریلیون قسمت از داده ها را جمع کرد که می تواند از روی قاعده آنالیز انجام داد و به جستجو ارتباط ها وپیدا کردن راه حل در مشکلات مختلف کمک کرد [۲] . به علاوه دولتمندان در بیشتر خرید و فروش از داده کاوی در پیدا کردن الگو ها و ارتباط های توان مشتری استفاده می کنند. سازمان های بزرگ و سازمان آموزشی نیز از داده کاوی برای فهمیدن همبستگی پر معنا که می تواند در جامعه ما توسعه یابد ، استفاده می شود. داده کاوی پل ارتباطی میان علم آمار، علم کامپیوتر، هوش مصنوعی، الگوشناسی، فراگیری ماشین و بازنمایی بصری داده می باشد. داده کاوی فرآیندی پیچیده جهت شناسایی الگوها و مدل های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده ها می باشد، به طریقی که این الگوها و مدلها برای انسانها قابل درک باشند [۳]. عموما در این حوزه، داده ها اغلب حجیم می باشند و به تنهایی قابل استفاده نیستند، بلکه دانش نهفته در داده ها قابل استفاده می باشد. بنابراین بهره گیری از قدرت فرآیند داده کاوی جهت شناسایی الگوها و مدلها و نیز ارتباط عناصر مختلف در پایگاه داده جهت کشف دانش نهفته در داده ها و نهایتا تبدیل داده به اطلاعات، روز به روز ضروری تر می شود. در مراجع مهم تعاریف مختلفی از داده سازی ارائه شده است. به عنوان نمونه، داده کاوی استخراج اطلاعات مفهومی، ناشناخته و به صورت بالقوه مفید از پایگاه داده تعریف شده است. همچنین بیان شده است که داده کاوی علم استخراج اطلاعات مفید از پایگاه های داده یا مجموعه داده ای می باشد.

داده کاوی استخراج نیمه اتوماتیک الگوها، تغییرات، وابستگی ها، نابهنجاری ها و دیگر ساختارهای معنی دار آماری از پایگاه های بزرگ داده می باشد. داده کاوی علاوه بر کاربرد در پایگاههای عظیم، در پایگاه های داده کوچک نیز بسیار پرکاربرد است و از نتایج و الگوهای تولید شده بوسیله آن در تصمیم گیری های استراتژیک تجاری شرکتهای کوچک نیز می توان بهره های فراوان برد. کاربرد داده کاوی در یک جمله را این گونه می توان بیان کرد : " داده کاوی اطلاعاتی می دهد، که شما برای گرفتن تصمیم هوشمندانه ای درباره مشکلات سخت شغلتان به آنها نیاز دارید" [۶-۴].

در حوزه داده کاوی معمولا به کشف الگوهای مفید از میان ابر داده ها اشاره می شود. منظور از الگوی مفید، مدلی در داده ها است که ارتباط میان یک زیر مجموعه از داده ها را توصیف می کند و معتبر ، ساده ، قابل فهم و جدید است [۷]. در متون آکادمیک نیز تعاریف گوناگونی برای داده کاوی ارائه شده اند . در برخی از این تعاریف داده کاوی در حد ابزاری که کاربران را قادر به ارتباط مستقیم با حجم عظیم داده ها می سازد معرفی گردیده است و در برخی دیگر، تعاریف دقیقتر که درآنها به کاوش در داده ها توجه می شود موجود است. برخی از این تعاریف عبارتند از [۱۱-۸]:

- داده کاوی عبارت است از فرایند استخراج اطلاعات معتبر، از پیش ناشناخته، قابل فهم و قابل اعتماد از پایگاه داده های توزیعی و بزرگ و استفاده از آن در تصمیم گیری در فعالیت های تجاری مهم
- اصطلاح داده کاوی به فرایند نیم خودکار تجزیه و تحلیل داده های پایگاه داده های بزرگ به منظور یافتن الگوهای مفید اطلاق می شود
- داده کاوی یعنی جستجو در یک پایگاه داده ها برای یافتن الگوهایی میان داده ها
- داده کاوی یعنی استخراج دانش کلان ، قابل استناد و جدید از پایگاه داده ها ی بزرگ
- داده کاوی یعنی تجزیه و تحلیل مجموعه داده های قابل مشاهده برای یافتن روابط مطمئن بین داده ها

همانگونه که در تعاریف گوناگون داده کاوی مشاهده می شود، تقریبا در تمامی تعاریف به مفاهیمی چون استخراج دانش، تحلیل و یافتن الگوی بین داده ها اشاره شده است. داده کاوی فرآیندی است که طی آن با استفاده از ابزار های تحلیل داده به دنبال کشف الگوها و ارتباطات میان داده های موجود که ممکن است منجر به استخراج اطلاعات جدیدی از پایگاه داده گردند، می باشد. در داده کاوی از بخشی از به نام تحلیل اکتشافی داده ها استفاده می شود که در آن بر کشف اطلاعات نهفته و ناشناخته از درون حجم انبوه داده ها تاکید می شود. بنابراین می توان گفت در داده کاوی تئوریهای پایگاه داده ها، هوش مصنوعی، یادگیری ماشین وعلم آمار را در هم می آمیزند تا زمینه کاربردی فراهم شود. باید توجه داشت که اصطلاح داده کاوی درمعنای عام زمانی به کار برده می شود که با حجم بزرگی از داده ها در حد گیگابایت یا ترابایت، مواجه باشیم که از این نظر یکی از بزرگترین بازارهای هدف، انبارجامع داده ها، مراکز داده وسیستم های پشتیبانی تصمیم برای بدست آوردن تخصص هایی در صنایعی مثل شبکه های توزیع مویرگی، تولید،

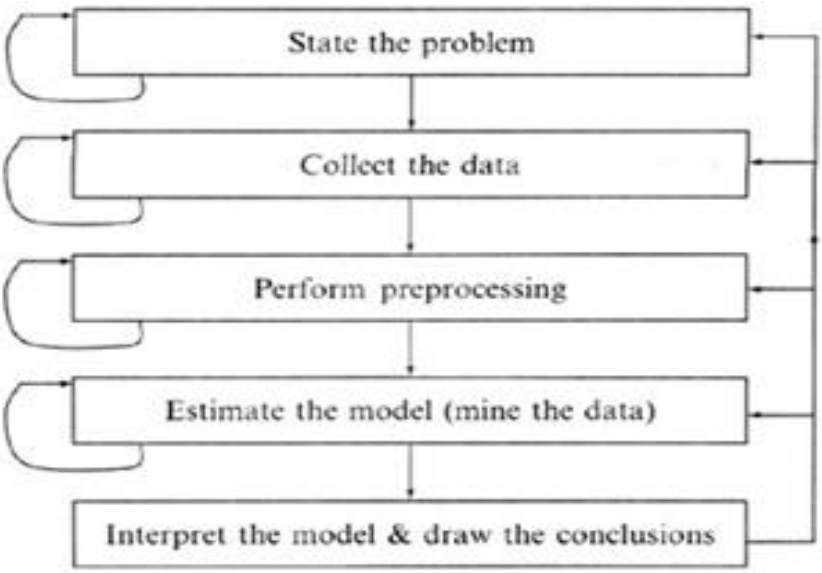
[1] Data Mining

مخابرات، بیمه و ...می باشد. در قسمت های بعدی مقاله به سایر ابعاد مهم داده کاوی می پردازیم.

فرایند عملیاتی سازی داده کاوی

در دنیای بشدت رقابتی امروز، اطلاعات بعنوان یکی از فاکتورهای تولیدی مهم پدیدار شده است. در نتیجه تلاش برای استخراج اطلاعات از داده ها توجه بسیاری از افراد دخیل در صنعت اطلاعات و حوزه های وابسته را به خود جلب نموده است. حجم بالای داده های دائما در حال رشد در همه حوزه ها و نیز تنوع آنها به شکل داده متنی، اعداد، گرافیکها، نقشه ها، عکسها، تصاویر ماهواره ای و عکسهای گرفته شده با اشعه ایکس نمایانگر پیچیدگی کار تبدیل داده ها به اطلاعات است. علاوه بر این، تفاوت وسیع در فرآیندهای تولید داده مثل روش آنالوگ مبتنی بر کاغذ و روش دیجیتالی مبتنی بر کامپیوتر، مزید بر علت شده است. استراتژیها و فنون متعددی برای گردآوری، ذخیره، سازماندهی و مدیریت کارآمد داده های موجود و رسیدن به نتایج معنی دار بکار گرفته شده اند. بعلاوه، عملکرد مناسب ابرداده[1] که داده ای درباره داده است در عمل عالی بنظر میرسد. داده کاوی یکی از پیشرفتهای اخیر در راستای فن آوریهای مدیریت داده هاست. داده کاوی مجموعه ای از فنون است که به شخص امکان میدهد تا ورای داده پردازی معمولی حرکت کند و به استخراج اطلاعاتی که در انبوه داده ها مخفی و یا پنهان است کمک می کند. داده کاوی فرآیند مرتب سازی از طریق مجموعه ای از داده های بزرگ هستند که به شناسایی الگوها و روابطی می پردازد واز طریق تجزیه و تحلیل داده ها به حل مشکلات تجاری کمک می کند. داده کاوی بخش کلیدی تجزیه و تحلیل داده به طور کلی و یکی از رشته های اصلی در علم داده است که از تکنیک های تجزیه و تحلیل پیشرفته برای یافتن اطلاعات مفید در مجموعه داده ها استفاده می کند. داده کاوی معمولا با نوشتن مقدار زیادی گزارش وتحقیق و استعلام در آنها اشتباه گرفته می شود. اما در واقع داده کاوی هیچ کدام ازاینها را شامل نمی شود. داده کاوی توسط تجهیزات خاصی صورت می پذیرد، که عملیات کاوش را بر اساس تجزیه و تحلیل مکرر داده ها انجام می دهد. داده کاوی با آنالیزهای متداول آماری نیز متفاوت است. در تعاریفی که از داده کاوی ارائه شد به اصطلاح فرایند اشاره شد. حتی در بعضی محیط های حرفه ای این نظر وجود دادرد که داده کاوی شامل انتخاب و بکارگیری ابزارهای مبتنی بر کامپیوتر برای حل مسائل فعلی و بدست آوردن یک راه حل بطور اتوماتیک و خودکار میباشد. برای آموزش داده کاوی، باید بر مفاهیم و روش های اعمال شده برخلاف همه جاذبه های ابزارهای مبتنی بر کامپیوتر که امور رابا جزئیات ودستورات با فرمت های خاصی باید به خیلی از سوالات از جمله چگونگی طراحی واستفاده از فرایندها را پاسخ داد به جای بیان جزئیات عملی ابزار مختلف داده کاوی تکیه نمود. فرایند عملیاتی سازی داده کاوی درواقع همان فرایند کشف دانش از پایگاه داده ها در حوزه داده کاوی می باشد و شامل پنج مرحله اصلی است که عبارتند از [۱۲-۱۴]:

- درک قلمرو یا بیان مسئله و فرموله کردن فرضیه
- انتخاب و جمع آوری داده ها
- تبدیل داده ها
- کاوش در داده ها
- تفسیر مدل، نتیجه گیری و گزارش



شکل ۱. فرایند عملیاتی سازی داده کاوی

[1] Metadata

بیان مسئله و فرموله کردن فرضیه :

در ابتدای امر پیش زمینه کشف دانش، فهم درست داده و مساله می باشد. بدون این فهم درست هیچ الگوریتمی صرف نظر از خبره بودن آن نمی تواند نتیجه مطمئنی برای شما حاصل نماید و داده را جهت کاوش آماده نموده یا نتایج را به طور صحیح تفسیر نمود. برای استفاده بهتر از داده کاوی باید یک بیان واضح از هدف داشت. در این مرحله انچه نیاز است ترکیبی از تخصص یک زمینه کاربردی و یک مدل داده کاوی است و شاید بتوان گفت یک تقابل نزدیک سر یک مسئله واحد و چندین فرضیه فرموله شده بین متخصصین داده کاوی و متخصصین کاربردی میباشد [۱۵].

انتخاب و جمع آوری داده ها:

این مرحله درارتباط با چگونگی تولید و جمع آوری داده ها است. بطور کلی، دو امکان وجود دارد:

- روش آزمون طراحی: زمانی است که فرایند تولید و توسعه داده ها تحت کنترل یک متخصص کاربردی (مدل ساز سیستم) باشد.
- روش دیداری: امکان دوم زمانی مطرح است که متخصص قادر به تولید فرآیند نیست یعنی تولید داده بصورت تصادفی در نظر گرفته شود.

پس از اینکه داده ها جمع اوری شدند یا در فرایند جمع اوری داده ها تا اندازه ای قرار گرفتند، توزیع نمونه گیری کاملا نامعلوم است. یعنی داده هایی که بعدا برای تست و بکارگیری آن مدل بکار می روند از چند نمونه مشابه استفاده می شوند. برای فرایند داده کاوی داده ها ی مورد نیاز موجود در انبار داده ها باید انتخاب شوند. درک این مطلب که برای ارزیابی یک مدل که بعدا برای تست و بکارگیری آن مدل بکار می رود، موفقیت آمیز باشد، بسیار مهم است در غیر اینصورت نتایج درستی حاصل نمی گردد. مثلا انبار داده ها شامل انواع مختلف و گوناگونی از داده ها است به عنوان مثال در یک پایگاه داده های مربوط به سیستم فروشگاهی، اطلاعاتی در مورد خرید مشتریان، خصوصیات آماری آنها، dispatcherها (توزیع کنندگان)، مشتریان، حسابداری و ...وجود دارند که همه آنها در داده کاوی مورد نیاز نیستند [۱۶].

پیش پردازش ها یا تبدیل داده ها

زمانی که که داده های مورد نیاز از پایگاه داده های موجود در انبار داده ها "جمع اوری "شدند و داده های مورد کاوش مشخص گردیدند، معمولا به تبدیلات خاصی روی داده ها نیاز است که شامل حداقل دو مرحله متداول می باشد [۱۷]:

- آشکارسازی (حذف) داده های غیرعادی: داده های غیرعادی یا غیر معمول درحقیقت داده های نتیجه سنجش خطاها، کدنویسی و ثبت خطاها است. دراینجا باید یا .1داده های غیرعادی را تشخیص داد و خذف کرد ویا باید روش های قوی مدل سازی رابگونه ای توسعه داد که نسبت به این نوع داده ها غیر حساس باشند.
- ویژگی های مقیاس بندی، رمزگذاری و انتخاب: در تبدیل داده ها توصیه میشود که داده ها را جهت تحلیل و بررسی مقیاس بندی و ورمزگذاری کرد .مثلا یک مشخصه با دامنه [0,1] ودیگری با دامنه [100,1000-] دارای ارزش مشابهی در تکنیک های اعلام شده نیستند، که در صورت نادیده گرفتن همین تفاوت در دامنه داده ها، روی نتایج نهایی داده کاوی تاثیر خواهند گذاشت.

برآورد مدل یا کاوش در داده ها

در این مرحله داده های تبدیل شده با استفاده از تکنیکها و عملیاتهای داده کاوی مورد کاوش قرار می گیرند تا الگوهای مورد نظر کشف شوند. یا به عبارتی دیگه، انتخاب و پیاده سازی تکنیک های داده کاوی در این مرحله صورت میگیرد. لبته این فرایند خیلی روشن و واضح نیست زیرا هنگام پیاده سازی ممکن است که مبتنی بر چندین مدل در یک فرآیند تکراریباشد. این مدل ها بطور کامل تر در مباحث مربوط به مفاهیم انواع دسته بندی،درختان تصمیم و قوانین تصمیم، شبکه های عصبی، انواع الگوریتم ها و... پیاده سازی می شوند [۱۸و۱۵].

تفسیر مدل، نتیجه گیری و گزارش

اطلاعات استخراج شده با توجه به هدف کاربر تجزیه و تحلیل شده و بهترین نتایج باید در تصمیم گیری کاربر موثر می باشند. هدف از این مرحله تنها ارائه نتیجه بصورت منطقی و یا نموداری نیست، بلکه پالایش اطلاعات ارایه شده به کاربر نیز از اهداف مهم این مرحله است. اگرچه تاکید بر دو مرحله آخر فرایند عملیاتی سازی داده کاوی بیشتر است اما باید به این نکته توجه داشت که اینها فقط دو مرحله از یک فرایند پیچیده هستند. همه مراحل در فرایند عملیاتی سازی داده کاوی و تک تک مراحل بطور مجزا بسیار تکرار پذیر هستند. همچنین باید توجه داشت که بدون توجه به صحت و درستی مراحل پنج گانه داده کاوی، ممکن است که مدل و داده حاصل آنچنان معتبر نباشد.

باید توجه داشت که داده کاوی یک ابزار جادویی نیست که بتواند در پایگاه داده به دنبال الگوهای جالب بگردد و اگر به الگویی جدیدی برخورد کرد آن را اعلام نماید بلکه صرفا الگوها و روابط بین داده ها را اعلام می کند، بدون توجه به ارزش آنها. بنابراین الگوهایی که به این وسیله کشف می شوند باید با جهان واقع تطابق داشته باشند. برای تضمین بدست آمدن نتایج با معنی لازم است بتوانید داده های خود را تحلیل کنید کیفیت خروجی شما به اطلاعات خارج از پایگاه داده (به عنوان مثال داده ای با ارزشی که متفاوت از داده های نوعی در پایگاه داده است) ستونهای ظاهرا بی ارتباط یا با ارتباط نزدیک به بقیه پایگاه داده (مانند تاریخ تولید یا انقضای کالا) بستگی نزدیکی دارند و الگوریتم های مختلف بر اساس حساسیتشان به داده ها روشهای متفاوتی دارند.

در قسمت بعدی از مقاله حاضر قصد داریم به کاربردهای مهم و رایج حوزه داده کاوی بپردازیم.

کاربردها

انگیزه جهت گسترش داده کاوی بطور عمده از دنیای تجارت در دهه ۱۹۹۰ پدید آمد. اساسا اغلب حوزه صنعت و تجارت به تصمیم گیریهای استراتژیک و یا اتخاذ خط مشی های جدید برای خدمت رسانی بهتر به مشتریان نیاز دارند. لذا، به داده هایی در مورد رفتار مصرفی گذشته مشتریان برای تعیین الگوهایی به وسیله داده کاوی، نیاز دارند. در واقع ابزار داده کـــــاوی، داده را می گیرد و یک تصویر از واقعیت به شکل مدل می سازد، این مدل روابط موجود در داده ها را شرح می دهد. برای بهبود بهره وری از یک فروشگاه داده کاوی از داده های انبار داده ، مدل هایی را ارائه میدهد که بیانگر این هستند که چه محصولات یا خدماتی، به چه مشتریانی، در چه زمانی و از طریق چه کانالی عرضه شود. بیشتر شرکتها، بانکهای داده ای عظیمی شامل داده های بازاریابی، منابع انسانی و مالی را دارا هستند. بنابراین، سرمایه گذاری در زمینه انبار داده، یکی از اجزای حیاتی در استراتژی مدیریت ارتباط با مشتری است. داده کاوی می تواند سودآوری مشتری های بالقوه را که می توانند به مشتریان بالفعل تبدیل شوند، پیش بینی کند و اینکه تا چه مدت به صورت مشتریان وفادار خواهند ماند و چگونه احتمالاً ما را ترک خواهند کرد. بنابراین می توان از طریق کاربردهای داده کاوی ارزش مشتریان را تعیین، رفتار آینده آنها را پیش بینی و تصمیمات آگاهانه ای را در این رابطه اتخاذ کرد [۲۰-۱۹].

امروزه عملیات داده کاوی به صورت گسترده توسط تمامی شرکت هایی که مشتریان در کانون توجه آنها قرار دارند، استفاده می شود، از جمله فروشگاه ها، شرکت های مالی، ارتباطاتی، بازاریابی و غیره. استفاده ازداده کاوی به این شرکتها کمک می کند تا ارتباط عوامل داخلی از جمله قیمت، محل قرارگیری محصولات، مهارت کارمندان را با عوامل خارجی از جمله وضعیت اقتصادی، رقابت در بازار و محل جغرافیایی مشتریان کشف نمایند. از آنجـائیـکه هـوش مصنوعی یکی از اصلی ترین عنـــاصـر داده کـــاوی می باشد و با توجه به اینکه به کمک سیستمهای کامپیوتری و پایگاه های داده، روزانه به میزان داده ها افزوده می شود، بنابراین استفاده هوشمندانه از دانش بالقوه ای که در این داده نهفته است در دنیای رقابتی امروز برای شرکت ها حیاتی می باشد. کاربرد داده کاوی در حوزه هوش مصنوعی پیش بینی وضع آینده بازار، گرایش مشتریان و شناخت سلیقه های عمومی آنها را برای شرکت ها ممکن می سازد [۲۱].

در حوزه دیگر کتابخانه ها و موسسات آموزشی با مشکل مدیریت کارآمد بار سنگین داده ها که دائما نیز در حال افزایش است روبرو می باشند. نرم افزارهای کامپیوتری بکار گرفته شده برای این منظور، غالبا فقط برای پرس و جوهای معمولی و پشتیبانی از مسائل مدیریتی و برنامه ریزی کوتاه مدت اداری جوابگو هستند. در حالیکه در عمق درون این حجم داده ها، الگوها و روابط بسیار جالبی میان پارامترهای مختلف بصورت پنهان باقی میماند. داده کاوی یکی از پیشرفتهای اخیر در حوزه کامپیوتر برای اکتشاف عمیق داده هاست. داده کاوی از اطلاعات پنهانی که برای برنامه ریزیهای استراتژیک و طولانی مدت میتواند حیاتی باشد پرده برداری میکند. تبیین مشخصه های اساسی فرایند داده کاوی و کشف کاربردهای ممکن آن در کتابداری و موسسات دانشگاهی اهداف اصلی این مقاله را شکل میدهند. پیشرفتهای حاصله در علم اطلاع رسانی و تکنولوژی اطلاعات، فنون و ابزارهای جدیدی برای غلبه بر رشد مستمر و تنوع بانکهای اطلاعاتی تامین می کنند. این پیشرفتها هم در بعد سخت افزاری و هم نرم افزاری حاصل شده اند. ریزپردازنده های سریع، ابزارهای ذخیره داده های انبوه پیوسته و غیر پیوسته، اسکنرها، چاپگرها و دیگر ابزارهای جانبی نمایانگر پیشرفتهای حوزه سخت افزار هستند. پیشرفتهای حاصل در نظامهای مدیریت بانک اطلاعات در طی چهار دهه گذشته نمایانگر تلاشهای بخش نرم افزاری است. این تلاشها در بخش نرم افزار را میتوان بعنوان یک حرکت پیشرونده از ایجاد یک بانک اطلاعات ساده تا شبکه ها و بانکهای اطلاعاتی رابطه ای و سلسله مراتبی برای پاسخگویی به نیاز روزافزون سازماندهی و بازیابی اطلاعات ملاحظه نمود. بدین منظور در هر دوره، نظامهای مدیریت بانک اطلاعاتی مناسب سازگار با نرم افزار سیستم عامل و سخت افزار رایج گسترش یافته اند. مثلا داده کاوی در حوزه بازاریابی، بدلیل پیوستگی غیرقابل انتظاری که بین پروفایل یک مشتری و الگوی خرید او ایجاد میکند اهمیتی خاص دارد. تحلیل رکوردهای حجیم نگهداری سخت افزارهای صنعتی، داده های هواشناسی و دیدن کانالهای تلوزیونی از دیگر کاربردهای آن است. در حوزه مدیریت کتابخانه کاربرد داده کاوی بعنوان فرایند ماخذ کاوی نامگذاری شده است. برخی دیگر از کاربردهای مهم داده کاوی عبارتند از [۲۴-۲۲]:

- ابزارهای پرس و جو: ابزارهای متداول زبان پرس و جوی ساختاربندی شده در ابتدا برای انجام تحلیل های اولیه بکار گرفته شدند که در حوزه داده کاوی می تواند مسیرهایی برای تفحص بیشتر نشان دهد.
- فنون آماری: مشخصات اصلی داده ها لازم است با کاربرد انواع مختلفی از تحلیلهای آماری شامل جدول بندی ساده و متقاطع داده ها و محاسبه پارامترهای آماری مهم بدست آید تا کارائی نتایج داده کاوی افزایش یابد.
- مصور سازی: با نمایش داده ها در قالب نمودارها و عکسها مانند نمودار پراکندگی؛ گروه بندی داده ها در خوشه های متناسب تسهیل میشود. استنباط عمیق تر ممکن است با بکارگیری تکنیکهای گرافیکی پیشرفته حاصل شود.
- پردازش تحلیلی پیوسته: از آنجا که مجموعه داده ها ممکن است روابط چندین بعدی داشته باشند، روشهای متعددی برای ترکیب کردن آنها وجود دارد. ابزارهای پردازش تحلیلی پیوسته به ذخیره چنین ترکیباتی کمک میکند و ابزارهای

ابتدا-انتهای پیوسته برای انجام پرس و جو ایجاد میکند. اما این ابزارها هیچ دانش جدیدی ایجاد نمی کنند.

- یادگیری مبتنی بر مورد: این تکنیک مشخصات گروههای داده ها را تحلیل میکند و به پیش بینی هر نهاد واقع شده در همسایگی شان کمک میکند. الگوریتمهایی که استراتژی یادگیری تعاملی را برای کاوش در یک فضای چندین بعدی بکار میگیرند برای این منظور مفیدند.
- درختان تصمیم گیری: این تکنیک بخشهای مختلف فهرست پاسخهای موفق داده شده مربوط به یک پرس و جو را بازیابی می کند و به این ترتیب به ارزیابی صحیح گزینه های مختلف در نتایج داده کاوی کمک میکند.
- قوانین وابستگی: اغلب مشاهده میشود که یک وابستگی نزدیک (مثبت یا منفی) بین مجموعه ای از داده های معین وجود دارد. بنابراین قوانین رسمی وابستگی برای تولید الگوهای جدید ساخته و بکار گرفته میشوند.
- شبکه های عصبی : شبکه های عصبی شامل الگوریتم های یادگیری ماشینی است که عملکرد خود را بر اساس کاربرد و ارزیابی نتایج بهبود می بخشند.
- الگوریتم های ژنتیکی: این هم تکنیک مفید دیگری برای پیش بینی هدف است. به این ترتیب که با یک گروه یا خوشه شروع میشود و رشدش در آینده را با حضور در برخی مراحل فرایند محاسبه احتمال جهش تصادفی؛ همانطور که در تکامل طبیعی فرض میشود طرح ریزی می نماید. این تکنیک به چند روش میتواند عملی شود. و ترکیب غیرقابل انتظار یا نادری را از عواملی که در حال وقوع بوده و مسیر منحنی طراحی داده ها را تغییر میدهند؛ منعکس میکند.

در نهایت می توان گفت با توجه به رشد محبوبیت کارکردها در حوزه داده کاوی، کاربردهای داده کاوی بصورت فزاینده درحال توسعه می باشد و سازمان ها و دولت ها باید در این مسیر اهتمام ویژه ورزیده و بسترهای توسعه کاربردهای داده کاوی را بیش از پیش فراهم سازند. در قسمت آخر مقاله به برخی ابزارهای کاربردی و مهم این حوزه نیز اشاره خواهیم داشت.

ابزارها

ابزارهای داده کاوی می‌توانند در شاخه‌های علوم، مهندسی و تجارت مورد استفاده قرار بگیرند و مشکلات دنیای واقعی را حل کنند. اینترنت مملو از ابزارهای داده کاوی است که عملکرد های مختلفی را برای کمک به کسب و کار ها در جهت درک داده های خود ارائه می دهند. هر ابزار دارای ویژگی ها و عملکرد های منحصر به فردی است که نیاز های مختلف را برآورده می کند. برخی از رایج ترین ابزارهای رایج بکار گرفته شده تحت عنوان داده کاوی عبارتند از [۲۷-۲۵و۱۹و۱۴]:

RapidMiner

RapidMiner یک ابزار داده کاوی است که توسط شرکتی به همین نام ساخته شده است . محیطی یکپارچه را برای آماده سازی داده ها، یادگیری ماشین، یادگیری عمیق، متن کاوی و تجزیه و تحلیل پیش بینی فراهم می کند. این نرم افزار برای کاربردهای تجاری و همچنین برای تحقیق، آموزش، آموزش، نمونه سازی سریع و توسعه برنامه کاربرد دارد. از تمام مراحل فرآیند یادگیری ماشین از جمله آماده سازی داده ها، تجسم نتایج، اعتبارسنجی و بهینه سازی مدل پشتیبانی می کند.

SAS

این ابزار یکی دیگر از ابزار های داده کاوی است که به طور واضح برای فعالیت های آماری در نظر گرفته شده است. SAS یک برنامه نویسی منحصر به فرد است که توسط شرکت های عظیمی برای تجزیه و تحلیل داده ها استفاده می شود. SAS کتابخانه ها و ابزارهای آماری مختلفی را ارائه می دهد که شما به عنوان دانشمند داده می توانید از آنها برای مدل سازی و ترتیب اطلاعات آنها استفاده کنید. در حالی که SAS کاملاً قابل اعتماد است ولی استثنایی که وجود دارد که آن هزینه استفاده از آن است و به دلیل هزینه بالا فقط توسط مشاغل بزرگتر مورد استفاده قرار می گیرد.

R Tool

R یک ابزار پیشگام در حوزه داده کاوی است زیرا شما را قادر می سازد سه وظیفه مشخص را فقط در یک سیستم عامل انجام دهید. توسعه دهندگان می توانند از R برای دستکاری داده ها استفاده کنند. به همین ترتیب ، توسعه دهندگان می توانند مجموعه داده های عظیم چند متغیره را به سرعت و با در نظر گرفتن قالبی که تحلیل آن دشوار است ، کاهش دهند. بعلاوه ، تجسم داده ها نیز آسان می شود. این تجسم علاوه بر طیف گسترده ای از نمودارهای متحرک و بصری را شامل می شود.

Apache Spark

Apache Spark یا اساساً Spark یک موتور تجزیه و تحلیل قدرتمند است و پر کاربردترین ابزار Data Science است. Spark به صراحت برای مقابله با پردازش دسته ای و پردازش خطی است. این برنامه با API های متعددی همراه است که دانشمندان داده را به دسترسی مکرر به داده ها برای یادگیری ماشین ، ذخیره سازی در SQL و غیره ترغیب می کند. این یک پیشرفت نسبت به Hadoop است و می تواند چندین برابر سریعتر از MapReduce عمل کند Spark دارای API های یادگیری ماشینی زیادی است. Spark می تواند به دانشمندان داده کمک کند تا با اطلاعات داده شده پیش بینی های شگفت انگیزی انجام دهند. Spark کاملاً در مدیریت خوشه تبحر دارد و این باعث می شود که از Hadoop برای ذخیره سازی استفاده شود. این چارچوب مدیریت خوشه است که Sparkرا قادر می سازد تا با سرعت بالا برنامه را پردازش کند.

Python

پایتون یک زبان برنامه نویسی تفسیر شده ، شی گرا و سطح بالا با معنایی پویا است. سطح بالای ساخته شده در ساختار داده ها ، همراه با تایپ پویا و اتصال پویا ، آن را برای توسعه سریع برنامه و همچنین استفاده به عنوان یک اسکریپت یا زبان چسب برای اتصال اجزای موجود به یکدیگر بسیار جذاب می کند. نحوه ساده و آسان پایتون بر خوانایی کد ها تأکید دارد و بنابراین هزینه نگهداری برنامه را کاهش می دهد. پایتون از ماژول ها و بسته ها بسیر زیادی پشتیبانی می کند که باعث انعطاف بالای این نرم افزار می شود. مفسر پایتون و کتابخانه استاندارد گسترده به صورت منبع یا باینری بدون هیچ گونه هزینه ای برای همه سیستم عامل های اصلی در دسترس هستند.

BigML

BigML یکی دیگر از ابزار های داده کاوی است که به طور گسترده مورد استفاده قرار می گیرد. این ابزار یک محیط GUI کاملاً تعاملی و مبتنی بر ابر است که می توانید برای پردازش الگوریتم های یادگیری ماشین استفاده کنید. BigML با استفاده از رایانش ابری برای نیازهای صنعت ، یک برنامه نویسی نهادینه ارائه می دهد. از طریق آن ، سازمان ها می توانند از الگوریتم های یادگیری ماشین در بخشهای مختلف سازمان خود استفاده کنند. به عنوان مثال ، می تواند از این نرم افزار برای پیش بینی فروش ، تجزیه و تحلیل ریسک و نوآوری در محصول استفاده کند. BigMLدر مدل سازی پیش بینی تخصص دارد. این ابزار از طیف گسترده ای از الگوریتم های یادگیری ماشین مانند خوشه بندی ، طبقه بندی ، پیش بینی سری زمانی و غیره استفاده می کند.

IBM SPSS Modeler

اگر علاوه بر این در حال طراحی مقیاس گسترده ای از پروژه ها مانند تجزیه و تحلیل متنی هستید . در آن مرحله میز کار IBM SPSS و رابط بصری آن را کشف خواهید کرد. این حتی شما را قادر می سازد تا طیف گسترده ای از الگوریتم های داده کاوی را بدون داشتن هیچ اطلاعاتی در مورد برنامه نویسی تولید کنید. به همین ترتیب می توان از این ابزار برای تشخیص ناهنجاری ، CARMA، شبکه های بی طرف پایه ، رگرسیون و شبکه های بیزی استفاده کنید که هفت مورد از تمایزهای چند لایه با یادگیری انتشار مجدد استفاده می کنند.

Tableau

پلتفرم تجزیه و تحلیل Tableau به عنوان گزینه پیشرو در بازار برای کسب و کار مدرن ، کاوش و مدیریت داده ها را برای مردم آسان تر بسیار آسان می کند. نرم افزار تبلو کشف و به اشتراک گذاشتن بینش هایی بدست آمده ای که می تواند مشاغل و جهان را تغییر دهد. مأموریت Tableau برای کمک به مردم برای دیدن و درک بهتر داده ها می باشد. به همین دلیل محصولات تبلو به گونه ای طراحی شده اند تا کاربر بهتر بتواند از نرم افزار استفاده کند. خواه یک تحلیلگر ، دانشمند داده ، دانشجو ، معلم ، مدیر اجرایی یا کاربر تجاری. نرم افزار تبلو قدرتمندترین ، ایمن ترین و انعطاف پذیرترین پلتفرم تجزیه و که بصورت end-to-end analytics است.

Orange

یک مورد خارق العاده از آنچه زبان برنامه نویسی پایتون می تواند ایجاد کند ، Orange مجموعه ای از نرم افزارها با کمک قطعات یادگیری ماشین و فرایندهای دستکاری داده است. این نرم افزار کاملاً رایگان است ، همراه با کمک به تمرینات مختلف آموزشی که از قبل بارگیری شده با گردش کار داده کاوی است. در این ابزار برخی از متداول ترین تجسم های مورد نیاز برای یک حرفه متخصص تنها با چند کلیک فاصله است که شامل نقشه های حرارتی ، نمودارهای پراکندگی ، استخراج متن است . حتی با توجه به تصاویر بصری بیش از حد ساده بصری که می تواند توسط هر کسی ، پیشرفته یا حتی در سطح تازه کار ساخته شود ، بهترین ابزار رایگان برای داده کاوی را ارائه می دهد.

NLTK

این ابزار پردازش زبان طبیعی توسعه مدل های آماری را مدیریت می کند که به کامپیوترها در درک زبان انسان کمک می کند. این مدل های آماری بخشی از یادگیری ماشین است و از طریق چند الگوریتم آن می تواند به کامپیوترها در درک زبان طبیعی کمک کند. NLTKبه طور گسترده ای برای روش های مختلف پردازش زبان مانند توکن سازی ، ساقه سازی ، برچسب گذاری ، تجزیه و ML استفاده می شود.

در نهایت لازم به ذکر است، عواملی که هنگام انتخاب یک ابزار داده کاوی در نظر باید گرفت و پیشنهاد می کنیم تا شما نیز در نظر بگیرید، شامل مقیاس پذیری داده ها، سهولت استفاده، دقت و قیمت است. مقیاس پذیری به توانایی ابزار برای مدیریت مجموعه داده های بزرگ اشاره دارد، زیرا مشاغل مختلف در طول روز با حجم فزاینده ای از داده ها سروکار دارند. سهولت استفاده برای قادر ساختن کاربران به دستکاری و تجزیه و تحلیل داده ها با تخصص فنی کم یا بدون تخصص فنی ضروری است. دقت یک عامل بسیار مهم می باشد، زیرا که ارزش اطلاعات داده ها در دقت آنها نهفته است. در نهایت، قیمت ابزار داده کاوی، به این دلیل که تعادل ویژگی ها و قیمت، به ویژه برای مشاغل کوچک بسیار مهم می باشد.

بحث و نتیجه‌گیری

امروزه با پیشرفت فناوری و حضور گسترده‌ی آن در زندگی روزمره‌مان شاهد کاربرد پررنگ داده و اطلاعات هستیم. در نتیجه تلاش برای استخراج اطلاعات از داده ها توجه بسیاری از افراد دخیل در صنعت اطلاعات و حوزه های وابسته را به خود جلب نموده است. حجم بالای داده های دائما در حال رشد در همه حوزه ها می باشد و تفاوت وسیع در فرآیندهای تولید داده پیچیدگی مدیریت و استخراج اطلاعات را افزایش داده است. اخیرا استراتژیها و فنون متعددی برای گردآوری، ذخیره، سازماندهی و مدیریت کارآمد داده های موجود و رسیدن به نتایج معنی دار بکار گرفته شده انداده کاوی یکی از پیشرفتهای اخیر

در راستای فن آوریهای مدیریت داده هاست. اصطلاح داده کاوی به فرایند نیم خودکار تجزیه و تحلیل پایگاه داده های بزرگ به منظور یافتن الگوهای مفید اطلاق می شود. در این تحقیق قصد داریم فرایند عملیاتی سازی داده کاوی را بررسی نموده و به تحلیل دقیق و کاربردی این موضوع بپردازیم. بعلاوه درخصوص کاربردها و ابزارهای مهم این حوزه تحقیق نمائیم. لازم به ذکر است، عواملی که هنگام انتخاب یک ابزار داده کاوی در نظر باید گرفت، شامل مقیاس پذیری داده ها، سهولت استفاده، دقت و قیمت است. در نهایت می توان گفت با توجه به رشد محبوبیت کارکردها در حوزه داده کاوی، کاربردهای داده کاوی بصورت فزاینده درحال توسعه می باشد و سازمان ها و دولت ها باید در این مسیر اهتمام ویژه ورزیده و بسترهای توسعه کاربردهای داده کاوی را بیش از پیش فراهم سازند.

تعارض منافع

«هیچ‌گونه تعارض منافع توسط نویسندگان بیان نشده است»

منابع و مآخذ

[1] Feng Z, Zhu Y. A survey on trajectory data mining: Techniques and applications. IEEE Access. 2016 Apr 13;4:2056-67.

[2] Obstfeld AE, Patel K, Boyd JC, Drees J, Holmes DT, Ioannidis JP, Manrai AK. Data mining approaches to reference interval studies. Clinical Chemistry. 2021 Sep 1;67(9):1175-81.

[3] Shekhar S, Jiang Z, Ali RY, Eftelioglu E, Tang X, Gunturi VM, Zhou X. Spatiotemporal data mining: A computational perspective. ISPRS International Journal of Geo-Information. 2015 Oct 28;4(4):2306-38.

[4] Romero C, Ventura S. Data mining in education. Wiley Interdisciplinary Reviews: Data mining and knowledge discovery. 2013 Jan;3(1):12-27.

[5] Mobasher B. Data mining for web personalization. InThe adaptive web: Methods and strategies of web personalization 2007 Jan 1 (pp. 90-135). Berlin, Heidelberg: Springer Berlin Heidelberg.

[6] Romero C, Ventura S. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews). 2010 Jul 26;40(6):601-18.

[7] Grossi V, Romei A, Turini F. Survey on using constraints in data mining. Data mining and knowledge discovery. 2017 Mar;31:424-64.

[8] Getoor L, Diehl CP. Link mining: a survey. Acm Sigkdd Explorations Newsletter. 2005 Dec 1;7(2):3-12.

[9] Miller HJ, Han J. Geographic data mining and knowledge discovery. CRC press; 2009 May 27.

[10] Liñán LC, Pérez ÁA. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. RUSC. Universities and Knowledge Society Journal. 2015 Jul 13;12(3):98-112.

[11] Chakrabarti S, Neapolitan RE, Pyle D, Refaat M, Schneider M, Teorey TJ, Witten IH, Cox E, Frank E, Güting RH, Han J. Data mining: know it all. Morgan Kaufmann; 2008 Oct 31.

[12] Alsrehin NO, Klaib AF, Magableh A. Intelligent transportation and control systems using data mining and machine learning techniques: A comprehensive study. IEEE Access. 2019 Apr 3;7:49830-57.

[13] Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook. New York: Springer; 2005 Sep 1.

[14] Ruan D, Chen G, Kerre EE, Wets G, editors. Intelligent data mining: techniques and applications. Springer Science & Business Media; 2005 Aug 24.

[15] Harding JA, Shahbaz M, Srinivas, Kusiak A. Data mining in manufacturing: a review.

[16] Westphal C, Blaxton T. Data mining solutions: methods and tools for solving real-world problems. John Wiley & Sons, Inc.; 1998 Jul 16.

[17] Barai SK. Data mining applications in transportation engineering. Transport. 2003 Sep 1;18(5):216-23.

[18] Gera M, Goel S. Data mining-techniques, methods and algorithms: A review on tools and their validity. International Journal of Computer Applications. 2015 Jan 1;113(18).

[19] Collier K, Carey B, Grusy E, Marjaniemi C, Sautter D. A perspective on data mining. Centre for Data Insight, Northern Arizona University, USA. 1998 Jul:2-4.

[20] Kleissner C. Data mining for the enterprise. InProceedings of the Thirty-First Hawaii International Conference on System Sciences 1998 Jan 9 (Vol. 7, pp. 295-304). IEEE.

[21] Gheware SD, Kejkar AS, Tondare SM. Data mining: Task, tools, techniques and applications. International Journal of Advanced Research in Computer and Communication Engineering. 2014 Oct;3(10).

[22] Peacock PR. Data mining in marketing: Part 1. Marketing Management. 1998;6(4):8.

[23] Kusiak A. Data mining: manufacturing and service applications. International Journal of Production Research. 2006 Sep 15;44(18-19):4175-91.

[24] Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P. Distributed data mining on grids: services, tools, and applications. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2004 Nov 15;34(6):2451-65.

[25] Solanki H. Comparative study of data mining tools and analysis with unified data mining theory. International Journal of Computer Applications. 2013 Aug;75(16):23-8.

[26] Sumathi S, Sivanandam SN. Introduction to data mining and its applications. Springer; 2006 Oct 12.

[27] Köksal G, Batmaz I, Testik MC. A review of data mining applications for quality improvement in manufacturing industry. Expert systems with Applications. 2011 Sep 15;38(10):13448-67.