**داده کاوی (Data Mining)**



**فهرست مطالب**

**عنوان صفحه**

فصل اول/تعریف داده کاوی

1-1نعاریف داده کاوی.....................................................................................................................2

1-2 مراحل فرایند کشف دانش از پایگاه .........................................................................4

1-3 تاریخچه داده کاوی............................................................................................8

1-4 کارکرد داده کاوی.............................................................................................9

1-5 مفاهیم اساسی در داده کاوی.................................................................................11

1-6 داده کاوی چگونه کار می کند..............................................................................12

1-7 قابلیتهای داده کاوی..........................................................................................13

فصل دوم/کاربرد داده کاوی

2-1 مثالی کلاسیک از کاربرد داده کاوی....................................................................17

2-2داده کاوی و آمار...........................................................................................18

2-2-1 تفاوت داده کاوی و آنالیز آمار.......................................................................18

2-2-2 آنالیز آماری...........................................................................................19

2-2-3 روش آنالیز آماری. ..................................................................................19

2-3 فواید و نقش داده کاوی در فعالیت شرکت ها............................................................20

2-4 کاربرد داده کاوی در کتابخانه ها و موسسات دانشگاهی...............................................21

2-4-1 پیشرفت در تکنولوژی داده پردازی..................................................................23

2-4-2 عناصر داده کاوی.....................................................................................24

2-4-3 فنون داده کاوی........................................................................................25

2-5 پروژه داده کاوی پیوند......................................................................................27

2-6 داده کاوی و کاربرد آن در کسب و کار هوشمند بانک...................................................29

2-6-1 بازاریابی ................................................................................................30

2-6-2 مدیریت ریسک..........................................................................................31

2-6-3 تشخیص تقلب............................................................................................31

2-6-4 به دست آوردن و حفظ مشتری.........................................................................31

2-7 فرصت و چالش های داده کاوی در شهر الکترونیکی......................................................32

2-7-1 داده کاوی چیست..........................................................................................36

2-7-2 زمینه و تکامل تدریجی داده کاوی در شهر الکترونیک...............................................37

2-8 کاربرد داده کاوی در شهر الکترونیکی......................................................................39

2-8-1 کشف علايق و انگيزه­هاي شهروندان و توليد سرويس­هاي شخصي­سازي.........................39

2-8-2 تجديد ساختار سايت وب شهر و افزايش کارايي سيستم...............................................40

2-8-3 تقويت برنامه­ريزي­هاي دولت و ترويج نو­آوري.......................................................41

2-8-4 بهبود تحليل­ها و تصميمات دولت........................................................................42

2-9 کاربردهای داده کاوی در کتابخانه ها و محیط های دانشگاهی............................................43

2-9-1 مدیریت و خدمات کتابخانه...............................................................................43

2-9-2 تذکرات نهایی.............................................................................................47

2-10 الگوريتم هاي داده كاوي براي طبقه بندي دانشجويان....................................................49

2-10-1 نتیجه های آزمایشی....................................................................................51

2-11 مشكلات تجاري براي داده كاوي..........................................................................52

2-11-1 روش كار داده كاوي...................................................................................53

فصل سوم/الگوریتم های داده کاوی

3-1 استخراج الگوريتم ها(تجزيه و تحليل سرويس ها-داده كاوي) ...........................................56

3-2 انواعي از الگوريتم ها داده كاوي............................................................................56

3-2-1 بکار بردن این الگوریتم ها..............................................................................57

3-3 جرئیات الگوریتم.............................................................................................................58

3-4 طراحي SPAM............................................................................................59

3-4-1 ابزاري براي کاوش در شبکة گستردة جهان.........................................................59

3-4-3 داده‌كاوي...............................................................................................................61

3-5 معماري يك سيستم داده‌كاوي..................................................................................62

3-6 وب‌كاوي................................................................................................................63

3-6-1 انواع عملياتي که در وب‌کاوي انجام مي‌شوند...........................................................63

3-6-2 كاوش محتوايي وب.......... .............................................................................64

3-6-3 کاوش ساختاري وب........................................................................................64

3-6-4 کاوش کاربردی وب .......................................................................................67

3-6-5 روشها و ابزارهاي وب كاوي............................................................................67

3-6-6 مقياسهاي اصلي براي مقايسه روشهاي وب كاوي......................................................69

3-7 انواع موتورهاي جستجو........................................................................................70

3-7-1 کاوشگر وب.................................................................................................70

3-7-2 دروازه وب .................................................................................................71

3-7-3 مقايسة موتور جستجوهاي دروازه وب و کاوشگر وب...............................................73

فصل چهارم/نرم افزار های داده کاوی

4-1 نرم افزار داده کاوی SPSS Clementine ..........................................................77

4-2 داده كاوي و OLAP..........................................................................................77

4-2-1 داده كاوي موفق................................................................................................................78

4-2-1-1 تحليل ارتباطات.........................................................................................78

4-2-1-2 سلسله مراتبي از انتخاب ها............................................................................82

4-2-2 طبقه بندي....................................................................................................84

4-2-3 حدس بازگشتي .............................................................................................84

4-2-4 سري هاي زماني...........................................................................................85

4-2-5 شبكه هاي عصبي..........................................................................................85

4-2-6 درخت هاي انتخاب........................................................................................86

4-2-7 استنتاج قانون.................................................................................................87

4-3 ساختن يك پايگاه داده داده كاوي...............................................................................87

4-3-1 جستجوي داده................................................................................................89

4-3-2 آماده سازي داده براي مدل سازي..........................................................................89

4-4 ساختن مدل داده كاوي...........................................................................................89

4-4-1 ارزيابي و تفسير..............................................................................................90

4-4-2 ايجاد معماري مدل و نتايج.................................................................................90

4-5 نرم افزارWeka ..............................................................................................91

4-5-1 تعریف.................................................................................................................91

4-5-2 قابليتهاي95

4-5-3 دريافت Weka ...................................................................................... 96

4-5-3-1 مروري بر Explorer .........................................................................96

4-5-3-2 به کارگیری فیلترها .............................................................................................................102

4-5-4 الگوريتم­‏هاي يادگيري....................................................................................107

4-5-5 رده بندهاي Lazy .......................................................................................109

4-5-6 خوشه................................................................................................................110

4-5-7 منابع وابسته...............................................................................................116

منابع.......................................................................................................131

**چکیده**

در این تحقیق به صورت کلی از داده کاوی تعریف شده و به بررسی آن به صورت دقیق و کاربردی پرداخته شده است. در بخشی به نرم افزار هایی پرداخته شده است که بزرگترین کاربردها را در نرم افزار های داده کاوی دارد.و در نهایت به نرم افزار وکا پراخته شده است که با توجه به داده های موجود تمامی منوهای نرم افزار توضیح داده شده است.

و در نهایت با توجه به تحقیقات انجام گرفته در زمینه ی دانشگاهی و به کارگیری نتایج به دست آمده مدیران آموزشی می توانند مشاوره های لازم را برای پیشگیری از رسیدن دانشجویان به وضعیت بحرانی بکار گیرند.همچنین این مدل ها می توانند به عنوان یک ابزار پشتیبان تصمیم گیری در سیستم های آموزشی مورد بهره برداری قرار گرفته و نقش مهمی را در ارتقا سطح علمی دانشگاه ها داشته باشند.

**فصل اول**

**داده کاوی چیست؟**

**1-1 تعاریف داده کاوی**

داده کاوی(Data Mining)پل ارتباطی میان علم آمار،علم کامپیوتر،هوش مصنوعی،الگوشناسی،فراگیری ماشین و بازنمایی بصری داده می باشد. داده کاوی فرآیندی پیچیده جهت شناساییالگوها و مدل های صحیح، جدید و به صورت بالقوه مفید، در حجم وسیعی از داده می باشد،به طریقی که این الگو ها و مدلها برای انسانها قابل درک باشند.

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.

داده کاوی استخراج اطلاعات مفهومی، ناشناخته و به صورت بالقوه مفید از پایگاه داده میباشد.[[1]](#footnote-1)

داده کاوی علم استخراج اطلاعات مفید از پایگاه های داده یا مجموعه داده ای می باشد.[[2]](#footnote-2)

داده کاوی استخراج نیمه اتوماتیک الگوها، تغییرات،وابستگی ها، نابهنجاری ها و دیگر ساختارهای معنی دار آماری از پایگاه های بزرگ داده می باشد.[[3]](#footnote-3)

داده کاوی 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" [1]

در داده کاوي معمولا به کشف الگوهاي مفيد از ميان داده ها اشاره مي شود . منظور از الگوي مفيد ، مدلي در داده ها است که ارتباط ميان يک زير مجموعه از داده ها را توصيف مي کند و معتبر ، ساده ، قابل فهم و جديد است .

در متون آکادميک تعاريف گوناگوني براي داده کاوي ارائه شده اند . در برخي از اين تعاريف داده کاوي در حد ابزاري که کاربران را قادر به ارتباط مستقيم با حجم عظيم داده ها مي سازد معرفي گرديده است و در برخي ديگر ، تعاريف دقيقتر که درآنها به کاوش در داده ها توجه مي شود موجود است . برخي از اين تعاريف عبارتند از :

* داده کاوي عبارت است از فرايند استخراج اطلاعات معتبر ، از پيش ناشناخته ، قابل فهم و قابل اعتماد از پايگاه داده هاي بزرگ و استفاده از آن در تصميم گيري در فعاليت هاي تجاري مهم.
* اصطلاح داده کاوي به فرايند نيم خودکار تجزيه و تحليل پايگاه داده هاي بزرگ به منظور يافتن الگوهاي مفيد اطلاق مي شود .
* داده کاوي يعني جستجو در يک پايگاه داده ها براي يافتن الگوهايي ميان داده ها .
* داده کاوي يعني استخراج دانش کلان ، قابل استناد و جديد از پايگاه داده ها ي بزرگ .
* داده کاوي يعني تجزيه و تحليل مجموعه داده هاي قابل مشاهده براي يافتن روابط مطمئن بين داده ها .

همانگونه که در تعاريف گوناگون داده کاوي مشاهده مي شود ، تقريبا در تمامي تعاريف به مفاهيمي چون استخراج دانش ، تحليل و يافتن الگوي بين داده ها اشاره شده است .

داده كاوي معمولا به عنوان جستجو ،آناليز كردن ، باقيمانده از مقدار زيادي از داده براي يافتن ارتباط ،الگو ها يا هر همبستگي آماري معني داري ، تعريف مي شود.با پيدايش كامپيوتر ، پايگاه داده بزرگ و اينترنت ، آسانتر مي توان ميليون ، بيليون و حتي تريليون قسمت هاي از داده را جمع كرد كه مي تواند از روي قاعده آناليز انجام داد و به جستجو ارتباط ها وپيدا كردن راه حل در مششكلات مختلف كمك كرد.به علاوه دولتمندان در بيشتر خريد و فروش از داده كاوي در پيدا كردن الگو ها و ارتباط هاي توان مشتري استفاده مي كنند.سازمان هاي بزرگ و سازمان آموزشي نيز از داده كاوي براي فهميدن همبستگي پر معنا كه مي تواند در جامعه ما توسعه يابد ، استفاده مي شود.

**نکته**: همانگونه که در تعاريف گوناگون داده کاوي مشاهده مي شود، تقريبا در تمامي تعاريف به مفاهيمي چون استخراج دانش ، تحليل و يافتن الگوي بين داده ها اشاره شده است.

" داده کاوي فرآيندي است که طي آن با استفاده از ابزار های تحليل داده به دنبال کشف الگوها و ارتباطات ميان داده هاي موجود که ممکن است منجر به استخراج اطلاعات جديدي از پايگاه داده گردند، مي باشد."

در داده کاوي از بخشي از به نام تحليل اکتشافي داده ها استفاده مي شود که در آن بر کشف اطلاعات نهفته و ناشناخته از درون حجم انبوه داده ها تاکيد مي شودبنابراين مي توان گفت در داده کاوي تئوريهاي پايگاه داده ها، هوش مصنوعي، يادگيري ماشين وعلم آمار را در هم مي آميزند تا زمينه کاربردي فراهم شود.

بايد توجه داشت که اصطلاح داده کاوي زماني به کار برده مي شود که با حجم بزرگي از داده ها در حد گیگابایت يا ترابايت، مواجه باشيم 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،مخابرات، بیمه و... می باشد.

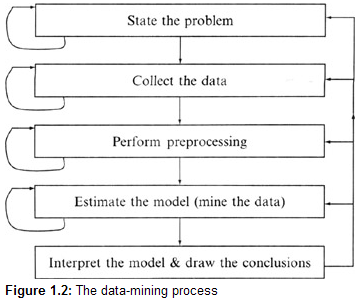
نکته: در تعاریفی که از داده کاوی ارائه شد به اصطلاح "فرایند" اشاره شد. حتی در بعضی محیط های حرفه ای این نظر وجود دادرد که داده کاوی شامل انتخاب و بکارگیری ابزارهای مبتنی بر کامپیوتر برای حل مسائل فعلی و بدست آوردن یک راه حل بطور اتوماتیک و خودکار میباشد.

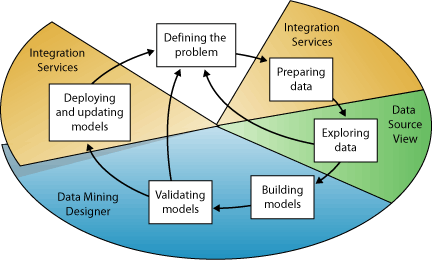
برای آموزش داده کاوی، باید بر مفاهیم و روش های اعمال شده برخلاف همه جاذبه های ابزارهای مبتنی بر کامپیوتر که امور رابا جزئیات ودستورات با فرمت های خاصی باید به خیلی از سوالات از جمله چگونگی طراحی واستفاده از فرایندها را پاسخ داد به جای بیان جزئیات عملی ابزار مختلف داده کاوی تکیه نمود.

**1-2 مراحل فرايند کشف دانش از پايگاه داده ها**

فرايند کشف دانش از پايگاه داده ها شامل پنج مرحله است که عبارتند از:

1. درک قلمرو یا بیان مسئله و فرموله کردن فرضیه
2. انتخاب و جمع آوری داده ها
3. تبديل داده ها
4. کاوش در داده ها
5. تفسير نتيجه یا تفسیر مدل و رسیدن به نتایج





شکل 1-1

بیان مسئله و فرموله کردن فرضیه:

در ابتداي امر پيش زمينه كشف دانش، فهم درست داده و مساله مي باشد. بدون اين فهم درست هيچ الگوريتمي صرف نظر از خبره بودن آن نمي تواند نتيجه مطمئني براي شما حاصل نمايد و داده را جهت كاوش آماده نموده يا نتايج را به طور صحيح تفسير نمود. براي استفاده بهتر از داده كاوي بايد يك بيان 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.

1. 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 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.)

نکته: براي فرايند داده کاوي داده ها ي مورد نياز موجود در انبار داده ها بايد 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 گوناگوني از داده ها است به عنوان مثال در يک پايگاه داده هاي مربوط به سيستم فروشگاهي، اطلاعاتي در مورد خريد مشتريان، خصوصيات آماري آنها،dispatcher ها (توزیع کنندگان)، مشتریان، حسابداري و ... وجود دارند که همه آنها در داده کاوي مورد نياز نيستند.

1. پیش پردازش ها یا تبديل داده ها

زمانی که که داده هاي مورد نياز از پایگاه داده های موجود در انبار داده ها "جمع اوری" شدند و داده هاي مورد کاوش مشخص گرديدند، معمولا به تبديلات خاصي روي داده ها نياز است که شامل حداقل دو مرحله متداول می باشد:

1. 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 1. داده های غیرعادی را تشخیص داد و خذف کرد ویا 2. باید روش های قوی مدل سازی رابگونه ای توسعه داد که نسبت به این نوع داده ها غیر حساس باشند.

2. 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.

1. برآورد مدل یا کاوش در داده ها

در این مرحله داده هاي تبديل شده با استفاده از تکنيکها و عملياتهاي داده کاوي مورد کاوش قرار مي گيرند تا الگوهاي 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 فرآيند تكراريباشد. ( 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)

1. تفسير نتيجه یا تفسیر مدل و رسیدن به نتایج

اطلاعات استخراج شده با توجه به هدف کاربر تجزيه و تحليل شده و بهترين نتايج باید در تصمیم گیری کاربر موثر مي باشند. هدف از اين مرحله تنها ارائه نتيجه (بصورت منطقي و يا نموداري) نيست، بلکه پالايش اطلاعات ارايه شده به کاربر نيز از اهداف مهم اين مرحله است .

هشدار: اگرچه تاکید بر مراحل 3و4 فرایند داده کاوی بیشتر است اما باید به این نکته توجه داشت که اینها فقط دو مرحله از یک فرایند پیچیده هستند. همه فرایند داده کاوی و تک تک مراحل بطور مجزا بسیار تکرار پذیر هستند.

هشدار : باید توجه داشت که بدون توجه به صحت و درستی مراحل 5گانه داده کاوی، ممکن است که مدل و داده حاصل انچنان معتبر نباشد.

**1-3 تاريخچه داده کاوي**

اخيرا داده کاوي موضوع بسياري از مقالات ، کنفرانس ها و رساله ها ي عملي شده است ، اما اين واژه تا اوايل دهه نود مفهومي نداشت وبه کار برده نمي شد .

در دهه شصت و پيش از آن زمينه هايي براي ايجاد سيستم ها ي جمع آوري و مديريت داده ها ايجاد شد و تحقيقاتي در اين زمينه انجام پذيرفت که منجر به معرفي و ايجاد سيستم هاي مديريت پايگاه داده ها گرديد .

ايجاد و توسعه مدلهاي داده اي براي پايگاه سلسله مراتبي ، شبکه اي و بخصوص رابطه اي در دهه هفتاد ، منجر به معرفي مفاهيمي همچون شاخص گذاري و سازماندهي داده ها و در نهايت ايجاد زبان پرسش SQL در اوايل دهه هشتاد گرديد تا کاربران بتوانند گزارشات و فرمهاي اطلاعاتي مورد نظر خود را ، از اين طريق ايجاد نمايند .

توسعه سيستم هاي پايگاهي پيشرفته در دهه هشتاد و ايجاد پايگاه هاي شي گرا ، کاربرد گرا(Application Oriented )و فعالباعث توسعه همه جانبه و کاربردي شدن اين سيستم ها در سراسر جهان گرديد . بدين ترتيب DBMS هايي همچون DB2 ، Oracle ، Sybase ، ... ايجاد شدند و حجم زيادي از اطلاعات با استفاده از اين سيستم ها مورد پردازش قرار گرفتند . شايد بتوان مهمترين جنبه در معرفي داده کاوي را مبحث کشف دانش از پايگاه داده ها ([[4]](#footnote-4)KDD) دانست بطوري که در بسياري موارد DM و KDD بصورت مترادف مورد استفاده قرار مي گيرند .

همانطور که در تعريف داده کاوي ذکر شد ، هدف از جستجو و کشف الگوهايي در پايگاه داده ها و استفاده از آنها در اخذ تصميمات حياتي است ، بنابراين مي توان گفت که DM بخشي از فرايند KDD است که در نهايت به ايجاد سيستم هاي DSS[[5]](#footnote-5) نقش داده کاوي در فرايند کشف دانش از پايگاه داده ها را نشان مي دهد .

براي اولين بار مفهوم داده کاوي در کارگاهIJCAI در زمينه KDD توسط Shapir مطرح گرديد . به دنبال آن در سالهاي 1991 تا 1994 ، کارگاههاي KDD مفاهيم جديدي را در اين شاخه از علم ارائه کردند بطوري که بسياري از علوم و مفاهيم با آن مرتبط مي باشد.

**1-4 كاركرد داده كاوي**

داده كاوی فرایندی تحلیلی است كه برای كاوش داده ها ( معمولا حجم عظیمی از داده ها - در زمینه های كسب وكار و بازار) صورت می‌گیرد و یافته‌ها‌با‌به‌كارگیری الگوهایی‌،‌احراز اعتبار می‌شوند . هدف اصلی داده كاوی پیش بینی است. ‌فرایند داده ‌كاوی شامل سه مرحله می باشد : 1. كاوش اولیه 2. ساخت مدل یا شناسایی الگو با كمك احراز اعتبار/ تایید و 3. بهره برداری.

**مرحله 1 : كاوش.** معمولا‌این‌مرحله با آماده سازی داده ها صورت می گیرد كه ممكن است شامل پاك سازی داده ها ،‌تبدیل داده ها‌و‌انتخاب زیرمجموعه‌هايي‌‌ از ركوردها‌با‌حجم‌عظیمی‌از‌متغييرها( فیلدها ) باشد . سپس با توجه‌به‌ماهیت‌مساله تحلیلی‌، این‌مرحله‌به‌مدل‌هاي‌‌ ‌پیش بیني ساده یا مدل‌های‌آماری‌و‌گرافیكی برای شناسایی متغیرهاي مورد نظر و تعیین پیچیدگی مدل‌ها برای استفاده در مرحله بعدی نیاز دارد .

**مرحله 2:ساخت و احراز اعتبار مدل.** این‌مرحله‌ به بررسی‌مدل‌هاي مختلف و گزینش بهترین مدل با توجه به كارآیی پيش‌بيني آن می پردازد. شاید این مرحله ساده به نظر برسد، اما اينطورنیست. تكنیك‌های‌متعددی‌برای‌رسیدن‌به‌این‌هدف توسعه یافتند.و " ارزیابی رقابتی مدل ها"‌نام گرفتند. بدین منظور مدل‌های مختلف برای مجموعه داده‌های یكسان‌‌به‌كار‌می‌روند ‌تا‌ كارآیی‌شان ‌با‌ هم مقایسه‌شود ،‌سپس مدلی كه‌بهترین كارآیی راداشته باشد‌، انتخاب می‌شود.‌این‌تكنیك‌ها عبارتند از : Bagging,Boosting ,Stacking  و Meta-learning.

**مرحله 3 : بهره برداری.** آخرین‌مرحله‌مدلی‌راكه‌درمرحله قبل‌انتخاب‌شده است، در داده‌های‌جدیدبه كار‌می‌گیردتا پیش‌بینی‌هاي‌خروجی‌های مورد انتظار را تولید نماید.

داده كاوی‌به‌عنوان‌ابزار‌مدیریت‌اطلاعات‌برای‌تصمیم گیری‌،‌عمومیت‌یافته‌است. اخیرا‌،‌توسعه تكنیك های تحلیلی جدید در این زمینه مورد توجه قرار گرفته است (مثلا *Classification Trees*)،اما هنوز داده كاوی مبتنی بر اصول آماری نظیر(*Exploratory Data Analysis (EDA*)می باشد.

با این وجود تفاوت عمده ای بین داده كاوی و EDA وجود ‌دارد‌. داده‌كاوی‌بیشتر‌به‌برنامه‌های‌كاربردی گرایش دارد تا ماهیت اصلی پدیده .به عبارتی‌داده كاوی كمتر با شناسایی روابط بین متغیرها سروكار دارد .

و به عبارت ديگر

عمل داده کاوی از یک پایگاه داده به چند مرحله مشخص تقسیم می شود:

**1.** مرحله اول : تشکیل انبار داده .

با توجه به عنوان ، این مرحله برای تشکیل محیطی پیوسته و یک پارچه جهت انجام مراحل بعدی و  داده کاوی در آن، انجام می گیرد.در حالت کلی انبار داده مجموعه پیوسته و طبقه بندی شده است که دائما در حال تغییر بوده و دینامیک است که برای کاوش  آماده می شود.

**2.** مر حله دوم : انتخاب داده ها

در این مرحله برای کم کردن هزینه های عملیات داده کاوی، داده هایی از پایگاه داده انتخاب می شوند که مورد مطالعه هستند و هدف داده کاوی دادن نتایجی در مورد آنهاست.

**3.** مرحله سوم : تبدیل داده ها .

مشخص است برای انجام عملیات داده کاوی لزوما باید تبدیلات خاصی روی داده ها انجام گیرد ممکن است این تبدیلات خیلی راحت و مختصر مثل تبدیل byte به  integer باشد یا خیلی پیچیده و زمان بر و با هزینه های بالا مثل تعریف صفات جدید و  یا تبدیل و استخراج داده ها از مقادیر رشته ای و ... باشد.

**4.** مرحله چهارم : کاوش در داده ها .

در این مرحله است که داده کاوی انجام می شود.در این مرحله با استفاده از تکنیک های داده کاوی داده ها مورد کاوش قرار گرفته ، دانش نهفته در آنها استخراج شده و الگو سازی صورت می گیرد.

**5.** مر حله پنجم : تفسیر نتیجه .

در این مرحله نتایج و الگو های ارائه شده توسط ابزار داده کاو مورد بررسی قرار گرفته و نتایج مفید معیین می شود.

**1-5 مفاهیم اساسی در داده كاوی**

**Bagging:** این مفهوم برای تركیب رده بندی های پیش بینی شده از چند مدل  به كار می رود.فرض كنیدكه قصدداریدمدلی برای رده بندی پیش بيني بسازیدو مجموعه داده های مورد نظرتان كوچك است.شمامی توانید نمونه هایی( با جایگزینی) را از مجموعه داده ها انتخاب و برای نمونه های حاصل ازدرخت رده بندی (مثلا C&RT وCHAID)استفاده نمایید.به طوركلی برای نمونه های مختلف به درخت های متفاوتی خواهید رسید.سپس برای پیش بینی با كمك درخت های متفاوت به دست آمده از نمونه ها ،‌یك رای گیری ساده انجام دهید.رده بندی نهایی ، رده بندی ای‌خواهد بود كه درخت های مختلف آنرا پیش بینی كرده اند .

**Boosting:** این مفهوم برای تولید مدل‌های چندگانه (برای پیش بینی یا رده بندی)به كار می‌رود. Boosting نیزاز روش C&RT یا CHAID استفاده وترتیبی از classifierها را تولید خواهد كرد .

**Meta-Learning:** این مفهوم برای تركیب پیش بینی‌های حاصل از چند مدل به كار می‌رود.و هنگامی كه انواع مدل‌های موجود در پروژه خیلی متفاوت هستند، كاربرد دارد. فرض كنید كه پروژه داده كاوی شما شامل Tree classifierها نظیر C&RTو CHAID، تحلیل خطی و شبكه های عصبی است.هر یك از كامپیوترها،رده بندی هایی رابرای نمونه ها‌پیش بینی كرده اند.تجربه نشان می‌دهدكه تركیب پیش بینی های چند روش دقیق تراز پیش بینی های هریك از روشهاست.پیش بینی های حاصل از چند classifier را می توان به عنوان ورودی meta-linear مورد استفاده قرار داد. Meta-linear پیش بینی هارا تركیب می كند تا بهترین رده بندی پیش بینی شده حاصل شود.

**1-6 داده كاوي چگونه كار مي كند؟**

از زماني كه مقدار بزرگي از تكنولوژي اطلاعات به طور جداگانه سيستم هاي تحليلي و ارتباطي توسعه يافتند ، داده كاوي ارتباطي بين اين دو فراهم كرد.نرم افزار داده كاوي ارتباط ها و الگو ها رادر تبديل ذخيره شده داده تجزيه و تحليل مي كند كه بر اساس جستجو كاربراست.برخي انواع از نرم افزار تحليلي ، آماري ، يادگيري ماشين و شبكه هاي عصبي را فراهم مي كنند.معمولا هر چهار نوع از ارتباط ها موارد زير را جستجو مي كنند :

* **كلاسها**: داده را به منظور استهاده در موقعيت گروهاي تصميم گيري ذخيره مي كند.براي مثال ، يك رستوران زنجيرهاي مي تواند اطلاعات خريد هاي مشتريان را زماني كه آنها بازديد مي كنند و چيزي كه معمولا سفارش مي دهند را تعيين كند.اين اطلاعات مي تواند به منظور كاهش ترافيك روزانه استفاده شود.
* **خوشه بندي**: اصلاح داده از ميان ارتباط هاي منطقي يا حق تقدم مشتري گروه بندي شود. براي مثال ، داده مي تواند قطعات فروش يا رابطه مشتري را شناسايي كند.
* **پيوستگي**: داده مي تواند رابطه را شناسايي كند.
* **الگو هاي مرتب**: داده رفتار الگو ها و گرايش ها را پيش بيني مي كند.براي مثال ، در خرده فروشي تجهيزات بيروني مي تواند احتمال سبد خريد را بر اساس خريد هاي مشتري از كيسه خواب تا كفش هاي راه پيمايي را پيش بيني كند .[[6]](#footnote-6)

داده كاوي شامل 5 عنصر مهم است:

* استخراج،انتقال و بار گيري داده تبديل به سيستم انبار داده.
* ذخيره و مديريت داده در سيستم پايگاه داده چند بعدي.
* تامين كردن دسترسي داده در آناليز هاي تجاري و تكنولوژي اطلاعات حرفه اي.
* آناليز داده از طريق نرم افزار هاي كاربردي.
* معرفي كردن داده در فرمتي مفيد از قبيل گراف يا جدول.

**1-7 قابليتهاي داده کاوی:**

* بايد توجه داشته باشيد كه داده كاوي يك ابزار جادويي نيست كه بتواند در پايگاه داده شما به دنبال الگوهاي جالب بگردد و اگر به الگويي جديدي برخورد كرد آن را به شما اعلام كند بلكه صرفا الگوها و روابط بين داده ها را به شما اعلام مي كند بدون توجه به ارزش آنها. بنابراين الگوهايي كه به اين وسيله كشف مي شوند بايد با جهان واقع تطابق داشته باشند.
* به اين ترتيب شما مي توانيد از هدف خريد مردم بدون اينكه فاكتورهايي براي خريد كالاهاي خود در نظر بگيريد مطلع شويد؟
* براي تضمين بدست آمدن نتايج با معني لازم است كه شما بتوانيد داده هاي خود را تحليل كنيد كيفيت خروجي شما به اطلاعات خارج از پايگاه داده(به عنوان مثال داده اي باارزشي كه متفاوت از داده هاي نوعي در پايگاه داده شماست) ستونهاي ظاهرا بي ارتباط يا با ارتباط نزديك به بقيه پايگاه داده(مانند تاريخ توليد يا انقضاي كالا) بستگي نزديكي دارند .الگوريتم بر اساس حساسيتشان به داده ها روشهاي متفاوتي دارند. اما غير عاقلانه است كه به محصول داده كاوي صرفا به براي تمام تصميم گيري هايمان تكيه كنيم.
* داده كاوي بطور اتوماتيك و بدون رهنمايي قادر به كشف راه حل ها نيست. شما ترجيحا به جاي بيان يك هدف مبهم مانند "كمك به ارتقاي پاسخ دهي به در خواست ها mail من " شما بايد از داده كاوي براي يافتن خصيصه هاي افرادي كه
* (1): به درخواست هاي شما پاسخ مي دهند
* (2): به درخواست هاي شما پاسخ داده و خريد زيادي مي كنند
* استفاده كنيد. الگو هايي كه داده كاوي براي يافتن به اين دو هدف استفاده مي كنند متفاوت است.
* اگر چه يك ابزار خوب براي داده كاوي شما را از پيچيدگي هاي تكنيكهاي آماري راحت مي سازد اما به شما براي فهميدن كار هاي ابزاري كه انتخاب كرده ايد و همچنين الگوريتمهايي كه بر پايه آن كار مي كند نيازمند است. انتخابي كه شما براي ابزار مورد نياز انجام مي دهيد و بهينه سازي هايي را كه شما انجام مي دهيد در دقت و سرعت كار بسيار تاثير دارد.
* داده كاوي و انبار داده ها :
* اغلب داده اي كه مورد كاوش قرار مي گيرد ابتدا از يك انبار داده آماده شده به داخل يك پايگاه داده كاوي سرازير مي شود. اين كار مزاياي زيادي دارد. پايگاه داده كاوي مي تواند به جاي يك انبار فيزيكي داده يك انبار منطقي از داده ها باشد. به شرط آنكه انبار دادهDBMS بتواند دامنه هاي منابع اضافي از داده كاوي را نيز پوشش دهد.

فصل دوم

کاربرد داده کاوی

**2-1 مثالی کلاسیک از کاربرد داده کاوی**

اغلب تجارت ها به تصميم گيريهاي استراتژيك و يا اتخاذ خط مشي هاي جديد براي خدمت رساني بهتر به مشتريان نياز دارند. به عنوان مثال فروشگاهها آرايش مغازه خود را براي ايجاد ميل بيشتر به خريد مجدداً طراحي مي كنند. اين  مثال به داده هايي در مورد رفتار مصرفي گذشته مشتريان براي تعيين الگوهايي به وسيله داده كاوي، نياز دارند.

برای روشن تر شدن مسئله می توان مثال را اینگونه بیان کرد که در یک فروشگاه زنجیره ای پس از داده کاوی مشخص میشود که درصدی از مشتریان خرید تلویزیون ، میز تلوزیون و گلدان کریستالی را هم در همان روز و بعد از خرید تلوزیون میخرند.مدیر فروشگاه می تواند بلافاصله دستوراتی صادر کند که براساس مدلهای تلوزیون موجود میزهایی و براساس مدل میزها گلدانهای کریستالی برای فروش سفارش داده شود و غرفه های جنبی غرفه تلویزیون را به میز و گلدان کریستالی اختصاص دهد. مطمئنا حتی پس از مدت کوتاهی سود حاصل از این بخش از فروشگاه به طور قابل ملاحضه ای ترقی خواهد کرد.

 در واقع ابزار داده كــــاوي، داده را مي گيرد و يك تصوير از واقعيت به شكل مدل مي سازد، اين مدل روابط موجود در داده ها را شرح مي 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 يا خدماتي، به چه مشترياني، در چه زماني و از طريق چه كانالي عرضه شود.

بيشتر شركتها، بانكهاي داده اي عظيمي شامل داده هاي بازاريابي، منابع انساني و مالي را دارا هستند. بنابراين، سرمايه گذاري در زمينه انبار داده، يكي از اجزاي حياتي در استراتژي مديريت ارتباط با مشتري است.

رابطه مشتري با زمان تغيير مي كند و چنانچه تجارت و مشتري درباره يكديگر بيشتر بدانند اين رابطه تكامل و رشد مي يابد. چرخه زندگي مشتري چارچوب خوبي براي به كارگيري داده كاوي در مديريت ارتباط با مشتري فراهم مي كند. در بخش ورودي داده كاوي، چرخه زندگي مشتري مي گويد چه اطلاعاتي در دسترس است و در بخش خروجي آن، چرخه زندگي مي گويد چه چيزي احتمالاً جالب توجه است و چه تصميماتي بايد گرفته شود. داده كاوي مي تواند سودآوري مشتري هاي بالقوه را كه مي توانند به مشتريان بالفعل تبديل شوند، پيش بيني كند و اينكه تا چه مدت به صورت مشتريان وفادار خواهند ماند و چگونه احتمالاً ما را ترك خواهند كرد.

بعضي از مشتريان مرتباً مراجعاتشان را به شركتها براي كسب مزيتهايي كه طي رقابت ميان آنها به وجود مي آيد، تغيير مي دهند. در اين صورت شركتها مي توانند هدفشان را روي مشترياني متمركز كنند كه سودآوري بيشتري دارند.

بنابراين مي توان از طريق داده كاوي ارزش مشتريان را تعيين، رفتار آينده آنها را پيش بيني و تصميمات آگاهانه اي را در اين رابطه اتخاذ كرد. [1]

**2-2 داده كاوي و آمار**

**2-2-1 تفاوت داده کاوی و آنالیز های آماری**

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 کاوي با آنالیزهای متداول آماری نیز متفاوت است.در زیرمی توان برخی از اصلی ترین تفاوت های داده کاوي و آنالیز آماری را مشاهده نمود:

**2-2-2 آنالیز آماری:**

**•** آمار شناسان همیشه با یک فرضیه شروع به کار می کنند.

**•** آنها از داده های عددی استفاده می کنند.

**•** آمارشناسان باید رابطه هایی را ایجاد کنند که به فرضیه آنها مربوط است.

**•** آنها می توانند داده های نابجا و نادرست را در طول آنالیز مشخص کنند.

**•** آنها می توانند نتایج کار خود را تفسیر و برای مدیران بیان کنند.

**داده کاوی:**

**•** به فرضیه احتیاجی ندارد.

**•** ابزارهای داده کاوی از انواع مختلف داده ، نه تنها عددی می توانند استفاده کنند.

**•** الگوریتمهای داده کاوی به طور اتوماتیک روابط را ایجاد می کنند.

**•** داده کاوی به داده های صحیح و درست نیاز دارد.

**•** نتایج داده کاوی نسبتا پیچیده می باشد و نیاز به متخصصانی جهت بیان آنها به مدیران دارد.

جهت درک بهتر تفاوت داده کاوی و آنالیزهای آماری به مثال زیر که در مورد شناخت کلاهبرداری های شرکت بیمه می باشد، توجه کنید.

**2-2-3 روش آنالیز آماری:**

یک مفسر ممکن است متوجه الگوی رفتاری شود که سبب کلاهبرداری بیمه گردد. بر اساس این فرضیه، مفسر به طرح یک سری سوال می پردازد تا این موضوع را بررسی کند. اگر نتایج حاصله مناسب نبود، مفسر فرضیه را اصلاح می کند و یا با انتخاب فرضیه دیگری مجددا شروع می کند. این روش نه تنها وقت گیر است بلکه به قدرت تجزیه و تحلیل مفسر نیز بستگی دارد.

مهمتر از همه اینکه این روش هیچ وقت الگوهای کلاهبرداری دیگری را که مفسر به آنها مظنون نشده و در فرضیه جا نداده ، پیدا نمی کند.

**2-3 فواید و نقش داده کاوی در فعالیت شرکتها**

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 یکیاز اصلی ترین  عنــاصـر 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.

داده کاوی پیش بینی وضع آینده بازار، گرایش مشتریان و شناخت سلیقه های عمومی آنها را برای شرکت ها ممکن می سازد.

**مراحل اصلی داده کاوی**

داده کاوی را " کشف دانش در داده ها " نیز می نامند. کشف دانش داده ها دارای مراحل مختلفی می باشد که در اینجا به صورت خلاصه آنها را بیان می کنیم :

**•** استخراج اطلاعات از چندین منبع داده ( پایگاه داده).

**•** یکپارچه سازی اطلاعات و حذف داده های زاید.

**•** قرار دادن اطلاعات اصلاح شده در انبار داده ها.

**•** انجام عملیات داده کاوی توسط نرم افزار های مخصوص.

**•** نمایش نتایج به صورت قابل فهم مانند گزارش و گراف.

**2-4 کاربردهای داده کاوی در کتابخانه ها و موسسات دانشگاهی**

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 فراینده داده کاوی و کشف کاربردهای ممکن آن در کتابداری و موسسات دانشگاهی اهداف اصلی این مقاله را شکل میدهند.

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[[7]](#footnote-7) که داده ای درباره داده است در عمل عالی بنظر میرسد.

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. در این رابطه میتوان از محصولاتی مانند، Dbase-IV, Unify, Sybase, Oracle  و غیره نام برد.

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. انگیزه برای گسترش داده کاوی بطور عمده از دنیای تجارت در دهه 1990 پدید آمد. مثلا داده کاوی در حوزه بازاریابی، بدلیل پیوستگی غیرقابل انتظاری که بین پروفایل یک مشتری و الگوی خرید او ایجاد میکند اهمیتی خاص دارد. (Barry and Linoff, 1997)تحلیل رکوردهای حجیم نگهداری سختافزارهای صنعتی، داده های هواشناسی و دیدن کانالهای تلوزیونی از دیگر کاربردهای آن است. در حوزه مدیریت کتابخانه کاربرد داده کاوی بعنوان فرایند ماخذ کاوی (bibliomining )نامگذاری شده است. این مقاله به کاربردهای داده کاوی در مدیریت کتابخانه ها و موسسات آموزشی می پردازد. در ابتدا به چند سیستم سازماندهی داده ها که ارتباط نزدیکی به داده کاوی دارند می پردازد؛ سپس عناصر داده ای توصیف میشوند و درپایان چگونگی بکارگیری داده کاوی در کتابخانه ها و موسسات آموزشی مورد بحث قرار گرفته و مسائل عملی مرتبط در نظر گرفته می شوند.

**2-4-1 پیشرفت در تکنولوژیهای داده پردازی**

سازمانهای بزرگ و چند- مکانه مثل بانکها، دفاتر هواپیمایی و فروشگاههای زنجیره ای با حجم زیادی از داده ها که ناشی از عملکرد روزانه آنهاست روبرو هستند. بطور سنتی چنین داده هایی به دو دسته تقسیم شده اند:

 1. رکوردهای اصلی (Master records)

 2.  رکوردهای عملیاتی (Transaction records)

  فرض بر این است که رکوردهای اصلی حاوی اطلاعات پایه هستند که معمولا چندان تغییر نمی کنند در حالیکه رکوردهای عملیاتی با توجه به طبیعت عملیات تجاری حتی بطور ساعتی تغییر خواهند کرد. سیستمهای مدیریت پایگاه داده مناسب برای پیوند دادن این دو مجموعه اطلاعاتی  و تهیه گزارشهای استاندارد جهت کنترل فعالیتها گسترش یافتند.  سیستم اطلاعات مدیریت رایج برای پشتیبانی عملیات و سرویس دهی به چند کاربر در سطوح مختلف سازمان مبتنی بر این نظریه است.

بمنظور کمک به تصمیم گیری راهبردی، نظریه تاسیس بانک اطلاعات رکوردهای اصلی به نظریه سازماندهی دیتا مارت (Data Mart)و انبار داده ها تغییر یافت.  استخراج اطلاعات از رکوردهای عملیاتی یا پایگاههای اطلاعات عملیاتی و سازماندهی آن برای تحلیل استاندارد یا زمانی فلسفه اولیه و اصولی چنین پیشرفتهایی است. گرچه، دیتا مارت و انبار داده ها از نظر هدف و ساختار با هم متفاوتند. (Inmon, 1998)

**2-4-2 عناصر داده کاوی**

 توصیف و کمک به پیش بینی دو کارکرد اصلی داده کاوی هستند. تحلیل داده مربوط به مشخصه های انتخابی متغیرها؛ از گذاشته و حال، و درک الگو مثالی از تحلیل توصیفی است. برآورد ارزش آینده یک متغیر و طرح ریزی کردن روند مثالی از توانایی پیشگویانه داده کاوی است.

برای عملی شدن هریک از دو کارکرد فوق الذکر داده کاوی، چند گام ابتدایی اما مهم باید اجرا شوند که از این قرارند:

1. انتخاب داده ها
2. پاک سازی داد ها
3. غنی سازی داده ها
4. کد گذاری داده ها

 با دارا بودن هدف کلی در مطالعه، انتخاب مجموعه داده های اصلی برای تحلیل، اولین ضرورت است. رکوردهای لازم میتواند از انبار داده ها و یا بانک اطلاعاتی عملیاتی استخراج شود. این رکوردهای داده جمع آوری شده؛ اغلب از آنچه آلودگی داده ها نامگذاری شده است رنج می برند و بنابراین لازم است پاکسازی شوند تا از یکدستی فرمت (شکلی) آنها اطمینان حاصل شود، موارد تکراری حذف شده و کنترل سازگاری دامنه بعمل آید. ممکن است داده های گردآوری شده از جنبه های خاصی ناقص یا ناکافی باشند. در این صورت داده های مشخصی باید گردآوری شوند تا بانک اطلاعات اصلی را تکمیل کنند. منابع مناسب برای این منظور باید شناسایی شوند. این فرایند مرحله غنی سازی داده ها را تکمیل میکند. یک سیستم کدگذاری مناسب معمولا" جهت انتقال داده ها به فرم ساختار-بندی شده جدید؛ متناسب برای عملیات داده کاوی تعبیه میشود .

**2-4-3 فنون داده کاوی**

 ممکن است متوجه شده باشید که فنون داده کاوی یک گروه نامتجانس را شکل میدهند چرا که هر تکنیکی که بتواند بینش جدیدی از داده ها را استخراج کند میتواند داده کاوی به حساب آید. برخی از ابزارهای رایج بکار گرفته شده تحت عنوان داده کاوی عبارتند از:  (Adriaans and Zantinge, 2003)

**ابزارهای پرس و جو:** ابزارهای متداول زبان پرس و جوی ساختاربندی شدهدر ابتدا برای انجام تحلیلهای اولیه بکار گرفته شدند که می تواند مسیرهایی برای تفحص بیشتر نشان دهد.

فنون آماری: مشخصات اصلی داده ها لازمست با کاربرد انواع مختلفی از تحلیلهای آماری شامل جدول بندی ساده و متقاطع داده ها و محاسبه پارامترهای آماری مهم بدست آید.

**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(**Online analytical processing) )OALP) **:** 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 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 :** این یک الگوریتم یادگیری ماشینی است که عملکرد خودش را بر اساس کاربرد و ارزیابی نتایج بهبود می بخشد.

**الگوریتم 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.

گام نهایی فرایند داده کاوی، گزارش دادن است. گزارش شامل تحلیل نتایج و کاربردهای پروژه، درصورت بکارگیری آنها، است . و متن مناسب، جداول و گرافیکها را در خود جای می دهد. بیشتر اوقات گزارش دهی یک فرایند تعاملی است که تصمیم گیرنده با داده ها در پایانه کامپیوتری بازی میکند و فرم چاپی برخی نتایج واسطه محتمل را برای عملیات فوری بدست می آورد.داده کاوی در تولید چهار نوع دانش ذیل مفید است: ((Fayyad et al., 1996))

- دانش سطحی (کاربردهای SQL)

- دانش چند وجهی (کاربردهای OALP)

- دانش نهان (تشخیص الگو و کاربردهای الگوریتم یادگیری ماشینی)

- دانش عمیق (کاربردهای الگوریتم بهینه سازی داخلی)

**2-5 پروژه داده کاوی پیوند**

[پروژه پیوند](http://www.peivand.org/)، یک طرح علمی- پژوهشی می باشد که با هدف گسترش فرهنگ ازدواج بر اساس علم و منطق، در قالب انجمن علمی مهندسی فناوری اطلاعات دانشگاه آزاد واحد تهران جنوب شروع به کار کرده است. فکر و ایده اصلی این طرح توسط استاد مجید خلیلیان مطرح شده و توسط یک تیم پروژه 4 نفره دانشجویی که همگی از دانشجویان رشته ی مهندسی فناوری اطلاعات می باشند، در حال پیگیری است. این طرح با بهره گیری از الگوریتم های داده کاوی و بر پایه علایق و ابعاد شخصیتی هر کاربر، به کاوش و جستجو در میان بانک اطلاعاتی خود می پردازد و قصد دارد تا با استفاده از یک روش هوشمندانه و مبتنی بر اصول روانشناسی، فرد را در نحوه انتخاب کمک کرده و مشاوره های لازم را به وی عرضه دارد. همان طور که در بالا اشاره شد، [**پروژه پیوند**](http://www.peivand.org/) با استفاده از علم داده كاوي سعی در رسیدن به اهداف خود را دارد. این علم فرآيندي است كه ابزارهای مختلف جستجو و بررسي داده ها را به كار مي گيرد تا الگوها و روابطي را ميان داده ها كشف كند كه بتوانند پيش بيني هايي معتبر انجام دهند. الگوریتم های داده کاوی پیوند نیز با انجام آنالیزهای مختلف بر روی ویژگی های شخصیتی هر فرد، مدل و نتیجه ای در قالب یک کمیت عددی که نشاندهنده میزان هماهنگی آن فرد با هر یک از کاربران ثبت شده است را به وی ارائه می دهند.

پس از انجام گرفتن مراحل اولیه تحقیقات و مشاوره با متخصصین روانشناسی و همچنین تکمیل الگوریتم داده کاوی مورد استفاده، سؤالات مورد نظر طرح گردید. در حال حاضر این پروژه، از طریق وب سایت [پروژه پیوند](http://www.peivand.org/) قابل دسترسی است. ممکن است در ابتدا به دلیل کمبود اطلاعات ثبت شده، هدف اصلی این طرح برای کاربران به طور کامل محقق نگردد، ولی استفاده از پیوند مزایای دیگری را به دنبال دارد که در زیر به تعدادی از آنها اشاره شده است:

1-در پایان پاسخگویی به سؤالات، عددی به کاربر ارائه می شود که نشان دهنده میزان تفاوت وی با دیگر افراد و معیار های ازدواج آنها می باشد.(میزان غیرعادی بودن فرد)

2-در طراحی اولیه سؤالات سعی شده است تا تمامی ابعاد اصلی افراد برای ازدواج مورد بررسی قرار گیرد که این خود می تواند فکر کاربر را به نکاتی جلب نماید که شاید پیش از این به آنها توجه نمی کرد.

3- با پاسخگویی به سؤالات، حداقل یک بار معیارهای خود را برای ازدواج مرور کرده اید و نظرتان بیشتر به ابعاد شخصیتی خودتان و فرد ایده آلتان جلب می شود.

لازم به ذکر است که در این مرحله از اجرای طرح، افراد می توانند بدون وارد کردن نام و نام خوانوادگی خود، ثبت نام نمایند. در ضمن اطلاعات تکمیلی در مورد این طرح از طریق سایت اختصاصی پروژه پیوند به اطلاع شما خواهد رسید.سیستم پیوند همواره آماده ی دریافت پیشنهادات و انتقادات شما عزیزان می باشد. [[8]](#footnote-8)

2-6 [**داده کاوی و کاربرد آن در کسب و کار هوشمند بانک**](http://ece.ut.ac.ir/dbrg/dmining/sdma.htm)

با رشد فزاینده حجم داده‌ها در سیستمهای متنوع کسب و کار، و همچنین نیاز روز افزون جهت دستیابی به اطلاعات ارزشمند و معرفت از این داده‌های خام، داده کاوی به عنوان روشی مهم و پرکاربرد برای استخراج اطلاعات و ارضاء این نیاز مطرح شده است. در واقع داده کاوی بخشی از فرایند استخراج معرفت (Knowledge Discovery)است که در آن الگوهای مفید و ضمنی در پایگاه داده ها جستجو می‌شوند. در این میان با افزایش کاربرد سیستمهای اطلاعات جغرافیایی، پایگاه‌های بزرگی از داده‌های متنوع جغرافیایی در دسترس قرارگرفته‌اند که کمک شایانی به انجام تحلیل‌های کامل‌تر و دقیق‌تر می‌نمایند.داده کاوی روی داده‌هایی که دارای یک یا چند ویژگی مکانی، فضایی و یا جغرافیایی باشند، داده کاوی فضایی (Spatial Data Mining)نامیده می‌شود و خروجی آن اطلاعات و معرفتی است که دارای خصوصیات فضایی و جغرافیایی، مانند مکان، جهت، فاصله، شکل هندسی و مانند آن می باشد. هدف از این پروژه بررسی و اجرای یک روش داده کاوی پیشرفته روی داده‌های فضايي موجود در بانک ملت ایران می‌باشد که با داده‌های مختلف بانکی از قبیل مکان شعب، شاخصهای بانکی مانند درآمد، سود، هزینه، تعداد کارکنان، میزان مراجعه و مانند آن تلفیق خواهند شد. بدین معنی که بعد از انجام مراحل لازم جهت آماده سازی داده‌ها -با ملاحضات لازم به دلیل فضایی بودن آنها- برای عملیات داده کاوی، شامل پردازش و پاکسازی داده‌ها(Data Processing and Cleaning) و ساخت انبار داده‌ها (Data warehousing)،و همچنین در نظر گرفتن روشهای دسترسی به داده‌های فضایی (Spatial Data Access) ،الگوریتمی برای استخراج قوانین وابستگی (Association Rule Mining) توسعه و پیاده سازی خواهد شد و از آن برای کشف روابط موجود ما بین مقادیر مختلف فضايي و جغرافیایی مانند ترکیب جمعیتی، کاربری‌های منطقه، وضعیت سنی، درآمد، تحصیلات، موقعیت رقبا، شبکه معابر و مانند آن از یک طرف و شاخصهای بانکی شعب مانند سود، هزینه، درآمد، کارایی و مانند آن از طرفی دیگر استفاده خواهد شد. دانش استخراج شده از این فرایند، در تصمیم گیریی های مختلف مدیران در حوزه مدیریت شعب، مانند مکانیابی، توسعه، تلفیق و تنظیم شعب، کاربرد و اهمیت بالایی خواهد داشت[2].

**2-6-1 بازاريابي**

بازاريابي يكي از مهمترين كاربردهاي داده‌كاوي در صنعت بانكداري در حوزه‌ي بازاريابي است. دپارتمان بازاريابي بانك مي‌تواند از داده‌كاوي براي تحليل پايگاه داده مشتريان استفاده كند و سرويس‌ها و محصولات ترجيحي مشتريان را تشخيص دهد. با ارائه سرويس‌ها و محصولاتي كه مشتريان واقعاً متقاضي آنها هستند از هدر رفتن هزينه براي تامين سرويس‌هاي بدون متقاضي جلوگيري مي‌شود. بنابراين واحد بازاريابي بانك مي‌تواند با كسب دانش بيشتر در مورد مشتريان،  بيشتر بر روي  نيازهاي آنها تمركز كند.

**2-6-2 مديريت ريسك**

مديران تصميم گيري بانك براي بسياري از تصميم گيري‌ها نياز دارند بدانند كه آيا مشتريان بانك ها قابل اطمينان هستند يا خير. اگر آنها اطلاعات كاملي در مورد مشتريان خود نداشته باشند، ارائه‌ي كارت‌هاي اعتباري به مشتريان جديد، گسترش اعتبار مشتريان كنوني و تصويب وام‌ها، تصميمات داراي ريسك براي بانك هستند.

داده‌كاوي مي تواند با استخراج داده در مورد مشتريان ريسك بانك‌ها را در اين موارد تصميم‌گيري كاهش دهد. با تحليل رفتارهاي تراكنشي مشتريان با حساب‌هاي سپرده‌هايشان مي‌توان رفتار آنها را در ارتباط با وام و بازپرداخت وام‌ها پبش‌بيني كرد. Credit scoring يكي از اولين ابزارهاي مديريت ريسك توليد شده، مي‌باشد و در صنعت بانكداري براي گرفتن تصميم در مورد اعطاي وام به مشتريان بسيار با ارزش است . وام دهنده‌ها بهتر است براي تصميم گيري، ابزاري ارزيابي ريسك داشته باشند. سوابق متقاضيان وام خوب و بد مي‌تواند براي ايجاد يك الگو براي تشخيص خوب يا بد بودن متقاضيان جديد به كار گرفته شود.

**2-6-3 تشخيص تقلب**

يكي ديگر از حوزه‌هاي مهم در صنعت بانكداري كه داده‌كاوي مي‌تواند در آن موثر باشد تشخيص تقلب است. "ّFlacon’s fraud assessment system"  يكي از سيستم‌هايي كه در تشخيص تقلب موثر بوده، اين سيستم توسط بسياري از بانك‌هاي صادر كننده كارت اعتبار ي به كار گرفته شده است. مثلاً ‌اين سيستم در آمريكا و با استفاده از داده كاوي 80% تراكنش‌هاي كارت‌هاي اعتباري صادر شده را بررسي كرده است.

**2-6-4 به دست آوردن و حفظ مشتري**

داده كاوي علاوه بر كمك به بانك‌ها براي بدست آوردن مشتريان جديد، به حفظ مشتريان جديد نيز كمك مي‌كند. كسب و حقظ مشتري براي هر صنعتي يك عامل اساسي است. امروزه مشتريان براي انتخاب بانكي به منظور انجام فعاليت‌هاي اقتصادي خود، گزينه‌هاي متعددي دارند در صورتي كه مسئولين اجرايي نتوانند توجه كامل مشتريان را جلب كنند، مشتري به راحتي مي‌تواند بانكي را پيدا كند كه توقعاتش را برآورده كند.

**2-7 فرصت­ها و چالش­هاي داده­کاوي در شهر الکترونيکي**

فناوري اطلاعات و ارتباطات خصوصا اينترنت، به­صورت فعال حكومت­ها را در قرن بيست و يكم به­شدت تغيير­داده­است. داده­کاوي يکي از دستاورد­هاي فناوري اطلاعات و ارتباطات براي بهبود تحليل­ها و تصميمات سازماني و استراتژي­هاي مديريتي مي­باشد. ادبيات بسيار وسيعي روي داده­کاوي و تکنولوژي­هاي بکارگيري موثر آن وجود­دارد. در سال­هاي اخير، شهر الکترونيک توجه بسياري از محققان را به­خود جلب­کرده­است. در واقع شهر الکترونيک از استفاده موثر دولت از تکنولوژي­هاي پيشرفته اطلاعات و ارتباطات و از طريق کانل­هاي اطلاعاتي مختلف همچون تلفن، اينترنت و ايستگاه­هاي عمومي کامپيوتري، ناشي مي­شود. هر روزه حجم عظيمي از داده­هاي دسترسي کاربران در سيستم­هاي شهر الکترونيک توليد مي­شوند. کاوش اين داده­ها در تجزيه و تحليل­هاي دولت، پيشگويي­ها، استراتژي­ها، کنترل و برنامه­ريزي­هاي عملياتي و تاکتيکي تاثير­گذار هستند. داده­کاوي در شهر الکترونيک نقش مهمي در تنظيم اهداف دولتي، کشف رفتار شهروندان و تخمين زمينه­هاي فعاليت آينده آنها دارد. البته در نقش اين داده­ها در موفقيت حکومت نبايد گزافه­گويي کرد. زيرا اتوماتيزه­کردن عوامل داده­کاوي در خدمات شهر الکترونيک، هنوز نيازمند حل­کردن بعضي مشکلات است. در اينجا براي فهم بهتر از اهميت و جايگاه داده­کاوي در شهر الکترونيک، مطالعه­اي شامل کاربرد­هاي داده­کاوي و چالش­هاي پياده­سازي موفق آن در شهر الکترونيک، توسعه­داده­شده­است. اين مطالعه مي­تواند منافعي را در اختيار سهام­داران مختلف و صاحبان اختيار که نياز به دستگيري دانش مخفي و ضمني از شهروندان، سازمان­ها و يا کسب و کار­ها دارند، قرار دهد.

با گسترش فناوري اطلاعات و ارتباطات[[9]](#footnote-9) درجهان و ورود سريع آن به زندگي روزمره مردم مسائل و ضرورت­هاي تازه­اي به­وجود­آمده­است .امروزه انسان توسعه يافته كسي است كه به اطلاعات دسترسي­داشته­باشد و دسترسي به اطلاعات نه يك ضرورت،كه يك قدرت محسوب­مي­شود. دراين­ميان شهر­ها به عنوان مراكز قدرت انساني و تمدن­هاي بشري بيش از پيش اهميت­يافته­اند. به اعتقاد الوين تافلر، مردم كره زمين تا به امروز سه موج اساسي تحول راپشت سرگذاشته اند :موج اول، موج انقلاب كشاوزي است كه زمان آغاز آن بركسي مشخص نيست. موج دوم، انقلاب صنعتي است كه به دنبال اختراع ماشين بخاردرسال 1764آغاز­شد و بالاخره موج سوم يا انقلاب انفورماتيك است كه ازسال 1946كه بشر به ساخت كامپيوتر نائل آمده آغاز­گشته­است. اگر در موج دوم سخت­افزارها به كمك انسان­ها مي­آمدند، درموج سوم اين، نرم­افزار­ها هستند که به خدمت بشر مي­شتابند و تفكرات و تصورات آدمي را به شكل كدهاي صفر و يك و با كمك امواج ماهواره­اي مبادله­مي­كنند. در موج سوم، انسان هر روز که بيشتر ياد­مي­گيرد بيشترمي فهمدكه با حقيقت فاصله دارد .موج سوم راموج خرد ورزي نيزلقب داده اند زيرا در اين عرصه­ها، انسان­ها ديگر فرصت ندارند زياد با هم صحبت­كنند، همه چيز تعريف­شده و براي هر تعريف، يك كد درنظرگرفته­شده­است. شهرالكترونيكي ((ElectronicCity)E-city)يكي از دستاوردهاي موج سوم است. با گسترش صنعت ارتباطات و سپس فراگير­شدن اينترنت، فضاي جديدي براي شهرها به­وجودآمده است كه از آن به شهر الکترونيکي يا شهر مجازي، تعبيرمي­شود. تحولات حوزه فناوري بويژه گسترش شبكه اينترنت در دهه اخير و انقلاب اطلاعاتي، موجب تحولات عميقي شده و جوامع انساني را به­سوي تبديل­شدن هر چه بيشتر و سريعتر، به جوامع الكترونيك پيش­مي­برد. جوامعي كه درآنها، مرزهاي ميان واقعيت و خيال تا اندازه زيادي مخدوش شده و يا به عبارت ديگر، وارد رابطه­اي كاملا تازه شده است. اينترنت، به­نوعي آغاز ساختمان شهري تازه است. شهربالقوه، شهري متعلق به زمان واقعي، شهري با يک شبکه ارتباطي دروني براساس سرعت مطلق امواج الکترومغناطيسي نوعي فرامرکز شهر جهان است که تمام شهرهاي واقعي اقمار و حاشيه اين شهر بالقوه­اند و اين شهر بالقوه درهيچ­جا نيست. تنها در آنجاست که آنتن­هاي بشقابي و فيبرهاي نوري در آنجا هستند .شهر مجازي، شهري نيست که در مکان واقعي قرارگرفته­باشد. اين شهر به مکان-زمان جغرافيايي وابسته نيست بلکه با زمان واقعي پخش و دريافت فوري يک پيام، رابطه دارد . در چنين شرايطي، هويت شهري و حتي شهروندي نيز معاني متفاوتي مي­يابند.

مطابق تعريف کارشناسان، شهر الکترونيکي شهري است که در آن از ابزار  ICT نظير برنامه­هاي کاربردي و کامپيوتر براي افزايش کارايي و اثر بخشي خدمات به مردم، بنگاه­هاي اقتصادي وکارمندان ديگر بخش­هاي دولت استفاده­مي­شود . به­طور­کلي مي­­توان مزاياي بالقوه شهر الکترونيک را براي شهروندان، کسب­و­کار­ها و موسسات اداري عمومي، به صورت زير خلاصه­کرد:

- **افزايش دسترسي:** شهروندان و کسب­و­کار­ها مي­توانند از طريق کانال­هاي مبتني بر تکنولوژي­هاي اطلاعاتي و ارتباطاتي از قبيل کيوسک­هاي اطلاعاتي، اينترنت و دستگاه­هاي موبايل، با دولت تعامل داشته­باشند.

- **انعطاف­پذيري:** توانايي تعامل در زمان­هاي مناسب­تر و توليد انواع سرويس­ها از ويژگي­هاي شهر الکترونيک محسوب مي­شود.

- **کارايي:** يک دولت کارامد­تر در شهر الکترونيک، منجر به خدمات بهتر و استفاده بهتري از منابع موجود مي­شود.

- **پوشش بيشتر:** توانايي دسترسي به درصد بيشتري از افراد جامعه مثل مناطق روستايي و محروم، افراد معلول و بيکار و غيره.

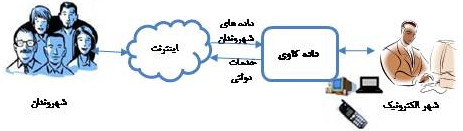
شهر الکترونيک، از شايسته­ترين کاربرد­هاي داده­کاوي به حساب­مي­آيد. در متون آکادميک تعاريف گوناگوني براي داده­کاوي ارائه شده­است. در واقع داده­كاوي روشي براي پشتيباني تصميم گيري مبتني بر كامپيوتر است و با گرفتن الگوريتم­هاي زيادي از آمار، هوش مصنوعي و ساير زمينه­ها كاري جديد را انجام مي­دهد. نقطه تحول الگوريتم­ها، داده كاوي نيست بلكه، ايده استخراج دانش به صورت خودكار از پايگاه داده­هاي بزرگ است.

با گسترش سيستم­هاي پايگاهي و حجم بالاي داده­هاي ذخيره شده در سيستم­هاي شهر الکترونيک، نياز به ابزاري است تا بتوان داده­هاي ذخيره شده را پردازش­کرد و اطلاعات حاصل از اين پردازش را در اختيار دولت و کاربران قرار­داد. يكي از اين ابزار­ها، داده­كاوي است که همزمان با ايجاد و استفاده از پايگاه داده­ها در اوايل دهه 80 براي جستجوي دانش در داده­ها شکل گرفت. اين داده­ها براي ثبت رفتار کاربران (شهروندان و سازمان­ها)، اهداف و انگيزه­هاي کاربران بکار­مي­روند (شکل پايين). بهره­برداري از چنين داده­هايي دولت را در شناسايي بيشتر شهروندان و سازمان­ها، نيازمندي­ها و رفتار­هاي آنان، کشف تخلفات و اتخاذ سياست­هاي مناسب، ياري­مي­دهد.

در سال­هاي اخير تحقيقات رو­به­رشدي در حوزه داده­کاوي توسط محققين انجام شده­است. از نقطه­نظر شهر الکترونيک، داده­کاوي اين پتانسيل را دارد که موجب کاهش هزينه­ها و کسب مزاياي رقابتي براي همه سهام­داران يعني شهروندان و کسب­و­کا­رها شود.

در حالي که ازدياد کاربرد­هاي داده­کاوي مي­تواند ابزار­هاي مديريتي داده بسياري عرضه کند، سازمان­هايي که اين ابتکارات را استعمال مي­کنند به محدوديت­هاي سو­تدبير پياده­سازي آن و اشتباهات و غفلت­هاي آن آسيب­پذير هستند. به عبارت ديگر در­حالي­که ابتکارات داده­کاوي رو به تکامل است، مسائل و چالش­هايي در رابطه با پياده­سازي و بررسي آن وجود­دارد که جوامع سياسي و دولتي براي جلوگيري از خطرات زيان­بار بايستي از آنها آگاه­باشند.

بقيه اين تحقيق به اين صورت سازمان­دهي شده­است: در ابتدا مقدمه­اي از داده­کاوي و بعضي کاربرد­هاي مهم آن شرح داده­مي­شود، سپس در بخش سوم زمينه و تکامل تدريجي داده­کاوي در شهر الکترونيک ارائه شده­است و سرانجام در بخش چهارم به توصيف کاربرد­ها و منافع داده­کاوي در شهر الکترونيک پرداخته­شده­است. چالش­هاي داده­کاوي در شهر الکترونيک بخش بعدي را تشکيل­مي­دهد و در آخر نتيجه­گيري اين مطالعه را به پايان مي­رساند.



شکل 2-1 جايگاه داده­کاوي در شهر الکترونيک

**2-7-1 داده­كاوي چيست؟**

فناوري­هاي نوين اطلاعاتي و ارتباطي، و همچنين تكنولوژي­هاي پشتيبان تصميم، با جمع­آوري، ذخيره، ارزيابي، تفسير و تحليل، بازيابي و اشاعه اطلاعات و دانش به كاربران خاص، مي­توانند در اطلاع­يابي به­موقع، صحيح و مورد­نياز به افراد تاثير زيادي داشته­باشند. يكي از ابزارهاي مورد استفاده در اين فناوري­ها، داده كاوي مي باشد. داده­کاوي شامل استفاده از ابزار­هاي پيشرفته تحليل داده به منظور کشف الگو­هاي معتبر، از قبل نا­شناخته و روابط در مجموعه داده­هاي بزرگ است. اين ابزار­ها، مدل­هاي آماري، الگوريتم­هاي رياضي و متد­هاي يادگيري ماشين (Machine Learning) (الگوريتم­هايي که عملکرد خود را از طريق تجربه به­صورت اتوماتيک بهبود مي­دهند) مي­باشد. داده­کاوي فراتر از جمع­آوري و مديريت داده است، و شامل تجزيه و تحليل و پيش­گويي مي­شود. نام ديگر آن کشف دانش در پايگاه داده يا به اختصار KDD (Knowledge Discover in Database) است.

داده­کاوي مي­تواند روي داده­هاي کمي، متني، يا چند­رسانه­اي انجام­گيرد. کاربرد­هاي آن شامل موارد زير مي­باشد:

- **قوانين وابستگي(**Acossiation Rule**) :** الگو­هايي که در آن وجود يك آيتم دلالت بر وجود آيتم ديگر دارد،

- **کلاس­بندي:** انتساب الگو­ها به يک مجموعه کوچک از کلاس­هاي از قبل تعريف شده به وسيله­ کشف بعضي روابط بين ويژگي­ها،

- **خوشه­بندي (**Clustering**):** گروه­بندي مشتريان يا مجموعه الگو­هايي که ويژگي­هاي مشابهي دارند،

- **پيش­گويي(**prediction**) :** کشف الگو­ها براي پيش­گويي منطقي درباره آينده،

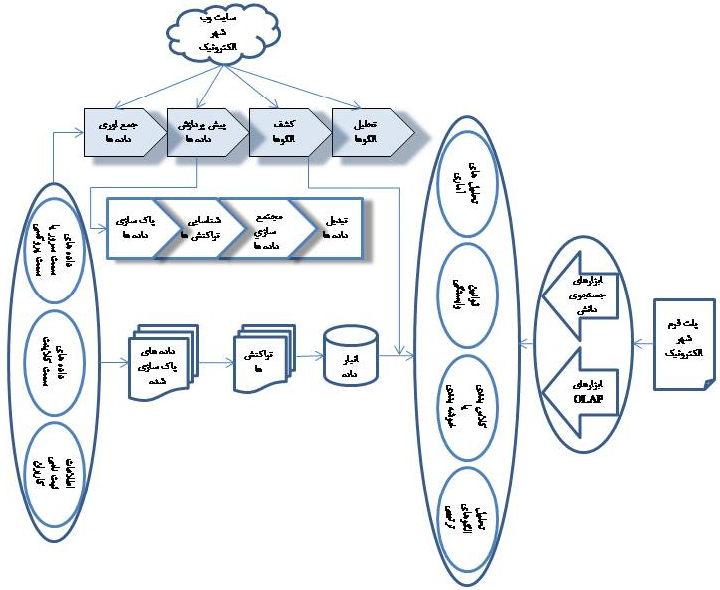
- **تحليل مسير (**Pth Analysis**)يا الگو­هاي ترتيبي:** الگو­هايي که در آن يک رخداد منجر به وقوع رخداد ديگر مي­شود.

داده­کاوي يک تکنولوژي جديد نيست ولي کاربرد آن به­طور معناداري در بخش­هاي مختلف خصوصي و عمومي رو­به­رشد بوده و عموما صنايعي چون بانک، بيمه، پزشکي و خرده­فروشي از داده­کاوي به هدف کاهش هزينه­ها، افزايش تحقيقات و افزايش فروش استفاده مي­کنند.

**2-7-2 زمينه و تکامل تدريجي داده­کاوي در شهر الکترونيک**

در­حالي­که پردازنده­هاي مرکزي بزرگ و گران­قيمت در دهه 60 و 70 ميلادي به منظور پردازش کنترل موجودي­ها و مديريت داده­ها برنامه­نويسي شدند، اکثر هزينه­ها به­سادگي بر روي دستگيري داده به فرمت قابل خواندن براي ماشين، و تحويل آن به سيستم­ها متمرکز شد، به­طوري که متجاوز از 80 درصد هزينه پروژه­هاي نرم­افزاري در دوره دهه 1960 تا اوايل دهه 90 صرف­شده­است. تاکيد روي مسايل تماميت داده­ها و کنترل صحت داده­ها، در اواخر دهه 90 به نقطه اوج خود رسيد و در حالي که سال 2000 سپري مي­شد اين مسئله منجر به سرمايه­گذاري بسياري از سازمان­ها بر روي ايجاد پايگاه­هاي داده به­منظور جلوگيري از خطاهاي فاجعه­انگيز شد. يک واکنش ثانويه حرکت به­سمت اجتناب از مشکلات Y2K، اين بود که شرکت­ها اقدام به پياده­سازي سيستم­هاي مديريت پايگاه داده گران­قيمتي کردند که توانايي ضبط و نگهداري، پردازش و اشتراک­ بهتر داده­ها را داشته­باشد.

با وجود اين موهبت تازه و ارزان، داده­هاي صحيح، بروز، قابل دسترس و قابل پردازش از سيستم­هاي اطلاعاتي، پس از سال 2000 هزينه­هاي گسترش کاربرد­هاي داده­کاوي کاهش يافت و متد­هاي جديدي از دستگيري صريحانه دانش يا هوش از انبار­هاي داده­اي بزرگ و توسعه­پذير شکل گرفت. نتيجه اينکه متد­هاي داده­کاوي نه فقط براي انبار داده شخصي يک سازمان، بلکه بر روي داده­هاي کپي­شده، خريداري­شده و يا حتي از منابع خارجي يا ديگر ادارات دولتي به­سرقت­رفته نيز قابل استفاده بود.



شکل 2-2چهارچوب داده­کاوي در شهر الکترونيک

**2-8 کاربرد­هاي داده­کاوي در شهر الکترونيک**

در اين بخش، کاربرد­هاي مهم داده­کاوي در شهر الکترونيک ارائه شده­است .اين کاربرد­ها، پروسيجر­هايي هستند که داده­هاي به­دست­آمده از تعامل شهروندان يا سازمان­ها با سيستم­هاي الكترونيكي شهری را به منظور پشتيباني تصميم­هاي شهر در امور مختلف، به دانش با ارزش تبديل مي­کند. اين پشتيباني­هاي تصميم­، شامل يافتن علايق، انگيزه­ها و خواسته­هاي شهروندان و بهبود رضايت شهروندان و سازمان­ها، سازمان­دهي مجدد سايت وب شهر و افزايش عملکرد و کارايي سيستم، تقويت برنامه­ريزي­هاي خدمات شهری و ترويج نو­آوري و بهبود تحليل­ها و تصميم­گيري­ها مي­شود. در ادامه به تفصيل هر يک از موارد مذکور مي­پردازيم.

2-8-1 **کشف علايق و انگيزه­هاي شهروندان و توليد سرويس­هاي شخصي­سازي**

فعاليت­هاي کاربران مي­تواند نمايانگر رفتار­هاي آنان و علايق آنها باشد. مثلا کجاها را کليلک مي­کنند، چه­مدت روي يک صفحه خاص توقف مي­کنند، چه کلماتي را جستجو مي­کنند، از چه سايت وبي مراجعه مي­کنند، روابط متقابل انها با سايت وب چيست و از اين قبيل. سرويس­هاي شخصي­سازي (Personalization Service)، يک کاربرد مبتني بر وب و يک هدف نهايي است که بر مبناي داده­هاي ثبت­نام کاربران و داده­هاي به­دست­آمده از تعامل آنها با سايت وب، به يافتن خوشه­هايي از کاربران با الگو­هاي دستيابي مشابه و سلايق يکسان مي­پردازد. آن­گاه هر کاربر را با توجه به الگوي پيمايشي فعلي، به يک خوشه واحد منسوب مي­کند و سپس پيشنهاداتي براي بقيه صفحات درون سايت وب، به او عرضه مي­کنند. اين پيشنهاد­ها بر اساس اينکه چه صفحاتي توسط کاربران همان خوشه رويت شده­است، به صورت پويا توليد­مي­شوند . کل اين فرايند در شکل زير به نمايش در­آمده­است.



شکل 2-3 فرايند توليد سرويس­هاي شخصي

بنابراين در شهر الکترونيک، ثبت کردن داده­هاي کاربراني که در سايت شهر حرکت­مي­کنند و شناسايي لينک­هايي که به­طور بالقوه مورد علاقه آنها است، به­منظور يافتن رفتار شهروندان، پي­بردن به نياز­هاي آنان و توليد سرويس­هاي شخصي شهروندان و مديريت نياز­هاي آنان، امري ضروري مي­باشد.

**2-8-2 تجديد ساختار سايت وب شهر و افزايش کارايي سيستم**

محتوي و ساختار سايت وب شهر هميشه ثابت نيست. طراحان وب­سايت نمي­توانند فقط به متخصصين در زمينه طراحي وب­سايت متکي باشند. آنها بايستي قادر باشند براي سهولت دسترسي شهروندان و سازمان­ها، ساختار و محتوي سايت وب شهر را بر اساس نتايج به­دست­آمده از تحليل­هاي داده­کاوي در شهر الکترونيک، به صورت پويا تغيير­دهند. به­عنوان مثال سايت وب را مي­توان با توجه به مسير­هاي دستيابي متوالي بازديد­کنندگان، دوباره­ساختار­دهي­کرد. اين کار موجب صرفه­جويي در زمان دسترسي کاربران و هزينه­ها خواهد­شد.

به­طور­کل به­منظور بهينه­سازي ساختار سايت وب شهر، پيشنهادات زير مورد استفاده قرار مي­گيرد :

- کاوش فايل­هاي لاگ کاربران به منظور استخراج صفحات دسترسي مرتبط با هم. بدين ترتيب لينک­هاي جديد بين صفحات اضافه مي­شود تا حرکت شهروندان را در سايت وب تسهيل سازد.

- استفاده از تکنيک­هاي تحليل مسير براي يافتن پر­تکرار­ترين مسير­هاي دستيابي و ايجاد پيام­هاي مهم در بين آنها به­منظور افزايش تمايل به شهروندان و بهبود کيفيت خدمات.

- کاوش فايل­هاي لاگ، مي­تواند موقعيت مورد­انتظار کاربران را براي اطلاعات آشکار­سازد. اگر تعداد تکرار موقعيت مورد­انتظار دسترسي از مکان واقعي دسترسي بيشتر بود، آن­گاه براي هدايت بهتر شهروندان، يک لينک جديد در بين صفحات افزده­مي­گردد.

داده­کاوي، با مشاهده و تحليل رفتار­ها و فعاليت­هاي شهروندان در سايت وب شهر الكترونيك، تاثير چشمگيري در در شهر الکترونيک دارد. و با ثبت دانش در حوزه اجتماعي و انساني مرتبط با فعاليت­ شهروندان، منجر به بهبود سيستم مي­شود. به عنوان مثال از آنجا که امنيت يکي از موضوعات مهم براي سايت وب شهر الکترونيک محسوب مي­شود با ثبت الگو­هاي دسترسي کاربران و تحليل مسير­هاي دستيابي آنها توسط تکنيک­هاي داده­کاوي، مي­توان به­سرعت و سهولت به مزاحمت­ها و تهاجم­ها پي برد و با آنها مقابله کرد، مانند تکنيک­هاي داده­کاوي براي تشخيص کلاهبرداري در کارت­هاي اعتباري. محققان نشان­داده­اند که بسياري از اين تکنيک­ها از سيستم­هاي مهندسي­شده و نيز افراد خبره، موفق­تر عمل­مي­کنند .

**2-8-3 تقويت برنامه­ريزي­هاي دولت و ترويج نو­آوري**

با بکارگيري تکنولوژي داده­کاوي، دولت مي­تواند منابع انساني، منابع اطلاعاتي و نيازمندي­ها را براي تعديل روابط بين منابع دروني و بيروني شهر مديريت­کند مثلا مديريت کل فرايند از طرح­ريزي برنامه­ها گرفته تا پياده­سازي آنها با استفاده از داده­کاوي. ابزار­هاي (Online Analysis and Process)OLAP مي­توانند جريان پروژه را متناسب با منابع اجتماعي بهينه کنند، اين، به مقدار زيادي هزينه منابع اجتماعي و جابجايي اطلاعات را کاهش­مي­دهد و سبب ترفيع علمي، آگاهي و هوشمندي برنامه­ريزي­ها و خدمات شهری مي­شود. به عبارتي ديگر، يک برنامه­ريزي دولتي موثر و علمي مي­تواند در سايه کنترل و مديريت بلادرنگ و با ابزار­هاي تکنولوژي هوشمند و تجسمي تقويت شود.

کاوش داده­هاي موجود در سيستم­هاي شهر الکترونيک، به صورت معنا­داري حساسيت دولت را به همه موضوعات از درخواست­هاي روزمره شهروندان گرفته تا رخداد­هاي تشنجي بهبود مي­بخشد. گذشته بر اين کاوش چنين داده­هايي، موجب حصول سريع و بموقع اطلاعات راجع به خدمات شهری و رويه­هاي اجتماعي توسط دولت مي­شود. بدين ترتيب مديريت و نقل و انتقال منابع اجتماعي را سيتماتيک­تر و پويا­تر ساخته و توانايي­هاي نو­آوري دولت را بهبود مي­بخشد.

نو­آوري فقط محدود به متد اجرايي و حکومتي و جريان فرايند امورات و خدمات شهری نمي­شود، و شامل وضع استراتژي­ها و سياست­هاي عمومي مي­شود و اين يک الزام اساسي است که دولت را از سازماني عملياتي به سازماني خدماتي تبديل مي­کند .

**2-8-4 بهبود تحليل­ها و تصميمات دولت**

90­درصد يک تصميم بر اساس اطلاعات و 10­درصد آن بر اساس مهارت­هاي ادراکي فرد صورت مي­پذيرد. تصميمات دولت از طريق تحليل اطلاعات مربوط به شهروندان از قبيل درخواست­ها، پيشنهاد­ها و تمايلات آنها اتخاذ­مي­شود. بنابراين داده­هاي حاصل از تعامل کاربران، منبعي از بينش خاص، اطلاعات، دانش و تجربه را توليد مي­کند که با صحت راهکار­هاي دولت در مواجه با مشکلات دخيل مي­باشد.

مشارکت شهروندان در مسائل عمومي و اجتماعي، به مهار­کردن فعاليت­هاي سياسي و خط مشي کشور کمک­مي­کند. تعامل شهروندان با سيستم­هاي الکترونيکي شهري و کاوش داده­هاي حاصل از آن مي­تواند يک برنامه، فعاليت، طرح و رهبري را قانوني و مشروع نمايد. از طرف ديگر کاوش اين اطلاعات منجر به شناسايي اطلاعات مهم و پنهاني مي­شود که تمام سطوح بخش­هاي دولتي را پشتيباني­کرده و منجر به کاهش هزينه­هاي پرسنلي و هزينه­هاي ناشي از انجام فعاليت­هاي سياسي مي­شود .[[10]](#endnote-1)

**2-9 کاربردهای داده کاوی در کتابخانه ها و محیط های دانشگاهی**

داده کاوی در ابتدا از حوزه تجارت برخاست اما کاربردهای آن در سایر حوزه هائی که به گردآوری حجم وسیعی از داده هائی می پردازند که دستخوش تغییرات پویا نیز می گردند؛ مفید شناخته شد. بخشهایی مثل بانکداری، تجارت الکترونیک، تجارت سهام، بیمارستان و هتل از این نمونه اند. انتظار میرود که استفاده از داده کاوی در بخش آموزش بطور عام امکانهای جدید بسیاری ارائه دهد. برخی کاربردهای داده کاوی در کتابخانه ها و قسمت اداری آموزش در ذیل مورد بحث قرار گرفته اند.

2-9-1 **مدیریت و خدمات کتابخانه**

عملیات کتابداری بطور کلی شامل مدیریت مدارک، ارائه خدمات و امور اداره و نگهداری است. هر کدام از این کارکردها با انواع مختلفی از داده ها سروکار دارد و بطور جداگانه پردازش میشود. اگرچه، انجام تحلیل ترکیبی براین مجموعه های داده  نیز میتواند افق تازه ای را بگشاید که به طرح خدمات جدید و تحول رویه ها و عملیات جاری کمک نماید. جدول یک برخی از کاربردهای ممکن داده کاوی را که میتواند در کتابداری مفید باشد ارائه میکند.

جدول 2-1 کاربردهای داده کاوی در کتابخانه ها

|  |  |
| --- | --- |
| **کاربرد متصور** | **بانک اطلاعاتی** |
| برای تعیین نقاط قوت و ضعف مجموعه | گردآوری منابع |
| برای ایجاد رابطه بین خواننده، منابع کتابخانه و زمان مشخصی از سال | استفاده از مجموعه |
| برای تحلیل سفارشهای پاسخ داده شده و سفارشهای دریافت شده | امانت بین کتابخانه ای |
| برای پیش بینی روند بازگشت منابع | داده های بخش امانت |
| برای نشان دادن منابع مالی بکار گرفته شده | داده های هزینه |

 داده کاوی میتواند برای پاسخ دادن به یک سوال خاص مربوط به کتابخانه و نیز برای کشف روندهای عمومی که به تصمیم گیری کمک میکنند، استفاده شود . برای مثال سوال میتواند چنین باشد: امکان اینکه امانت گیرندگان منابع را یک هفته بعد از تاریخ عودت برگردانند تا نامه های یادآوری کمتری فرستاده شود چقدر است؟ یا میزان اشتراک مورد انتظار برای نشریات بین المللی انتخاب شده برای سال آینده چقدر است؟ درک الگوی استفاده کلی مجلات الکترونیکی یا تحلیل درخواستهای  اعضا  برای میکروفیلمها طی 5 سال گذشته نیز همگی مثالهایی از کشف روندهای عمومی اند. دامنه تحلیل استنادی هم میتواند با استفاده از داده کاوی گسترش داده شود.

در ارتباط با کتابخانه ها، وب کاوی حوزه دیگری از علاقمندی است. وب کاوی شامل محتوا کاوی وب، ساختار کاوی وب و استفاده کاوی وب با توجه به یک موضوع خاص است که در طراحی خدمات جدید مبتنی بر وب کمک خواهد کرد.

مدیریت موسسات دانشگاهی اداره موسسات دانشگاهی کار پیچیده ای است. در این موسسات دائما" نیاز به درآمدزایی و خود- کارآمدی و کاهش وابستگی به بودجه دولتی احساس میشود. این مساله کنترل دائمی جنبه های مختلف هر فعالیت و پروژه را می طلبد. بانکهای اطلاعاتی برای چنین موسساتی  مربوط به دانشجویان، دانشکده، اساتید و کارمندان، تعداد رشته ها و چند مورد دیگر است . ارزیابی تقاضا و وضعیت عرضه نقش مهمی بازی میکند. مرور بانکهای اطلاعاتی نمونه نمایانگر کاربردهای بالقوه داده کاویست.

جدول 2-2 کاربردهای داده کاوی در موسسات دانشگاهی

|  |  |
| --- | --- |
| **کاربرد متصور** | **بانک اطلاعاتی** |
| برای درک رابطه های جمعیت شناختی، اقتصادی و اجتماعی | ثبت نام دانشگاهی |
| برای ایجاد رابطه بین عوامل اقتصادی-اجتماعی و نمرات اخذ شده | کارایی دانشگاهی |
| برای تعیین میزان مفید بودن سیستم با استناد به نمرات امتحان | بانک سوالات |
| برای ارزیابی همکاری دانشکده با توجه به میزان استفاده از کتابخانه | همکاری فکری |
| برای پیدا کردن تأثیر انتشارات در تقاضا برای رشته ها | انتشارات |
| برای تحلیل سوالات دریافت شده در وب سایت دانشگاه و کمک به ایجاد رشته های جدید دانشگاهی | بازدید از وب سایت |

کاربرد داده کاوی در دانشگاه ملی سنگاپور قابل ملاحظه است. در این دانشگاه از ابزارهای داده کاوی برای شناسایی و دسته بندی دانشجویانی که به کلاسهای پیش نیاز برای واحد درسی ارائه شده نیاز داشتند استفاده شد.  (Kurian and John, 2005)علاوه بر آن، مسائلی مانند اختصاص بهتر منابع و نیروی انسانی، مدیریت روابط دانشجو و به تصویر کشیدن رفتار گروههای مختلف میتواند بوسیله ابزارهای داده کاوی انجام شود.

 محدودیت هاکاربرد داده کاوی با چند عامل محدود شده است. اولین مورد به سخت افزار و نرم افزار لازم و موقعیت بانک اطلاعاتی مربوط میشود . برای مثال در هند، داده های غیر مجتمع که برای کاربردهای داده کاوی لازم است ممکن است به فرم دیجیتالی در دسترس نباشد. در دسترس بودن نیروی انسانی ماهر در داده کاوی نیز مسأله مهم دیگری است. محرمانه بودن رکوردهای مراجعان ممکن است در نتیجه پردازش داده های مبتنی بر داده کاوی آسیب پذیر شود. کتابداران و مؤسسات آموزشی باید این مسأله را در نظر داشته باشند؛ چرا که در غیر اینصورت ممکن است گرفتار شکایات قانونی گردند.محدودیت دیگراز ضعف ذاتی نهفته  در ابزارهای نظری ناشی میگردد. ابزارهایی مانند یادگیری ماشینی و الگوریتمهای ژنتیکی بکار گرفته شده در  فعالیتهای داده کاوی به مفاهیم وفنون منطق و آمار بستگی دارد. در این حد نتایج به روش مکانیکی تولید شده و بنابراین به یک بررسی دقیق نیاز دارند. اعتبار الگوهای بدست آمده به این طریق؛ باید آزمایش شود. چرا که که در بسیاری موارد روابط علل و معلول مشتق شده؛ از برخی استدلالات غلط ذیل رنج میبرند.(Cannavo, 2003)

●  علت دور مثلا" امکانات ضعیف خوابگاه باعث می شود دانشجویان نمرات پایینی کسب نمایند.

●  علت مجردمثلا" بودجه محدود بر بازدهی پژوهشی دانشکده تأثیر می گذارد.  
● علائم در نظر گرفته شده برای این عواملمثلا" مجموعه کتابخانه ممکن است افزایش نیابد چون تعداد خوانندگان مرتبا" کاهش می یابد.

● سفسطه دسته بندی مثلا" مدرسان حقوق بسیار بالا دریافت می کنند و کل حقوقشان بالغ بر میلیونها میشود.

●سفسطه ترکیب مثلا" اگر هر مدرس در دانشکده شایسته و واجد صلاحیت باشد کل دانشکده عملکرد بهتری خواهد داشت.

●  سوگیری در انتخاب نمونه:مثلا"استناد به یافته های یک پیمایش نمونه گیری شده از دانشجویان یک دانشکده که از خانواده های ثروتمند هستند و مخارج روزانه در خوابگاه برای هر دانشجو 100 دلار است. از آنجایی که مطالعه الگوها و استخراج روابط میان رکوردها مستلزم کاربرد منطق قیاسی و استقرایی است فرد باید مراقب اشتباهاتی که عموما" رخ میدهد باشد. برای مثال بحثهای قیاسی یا استقرایی، تا زمانیکه وضعیت درست بودن فرضیه آزمایش نشود چیزی درباره درست یا غلط بودن نتایجشان نمی گویند. طبیعتا، نتایج تولید شده ماشینی ممکن است از چنین نقایصی رنج ببرند.

**2-9-2 تذکرات نهایی**

 بکارگیری تکنولوژی اطلاعات توسط هر سازمان در عمل یک فرایند هموار نیست. کتابخانه یا مؤسسه دانشگاهی از این قضیه استثنا نیست. اما، تجربه نشان میدهد که یک برنامه نظام مند میتواند ظهور و نگهداری تکنولوژی اطلاعات در محیط کتابخانه را تسهیل کند. (Patkar and Iyer, 1990; Patkar, 2000, 2004) حتی کاربرد تکنولوژی های پیشرفته پردازش اطلاعات مثل سیستمهای خبره و سیستم اطلاعات جغرافیایی (جی.آی.اس) در کتابخانه گزارش شده  است. (Myers, 1992; Patkar, 1999)  با این پیش زمینه ، کاربرد داده کاوی بوسیله کتابخانه ها و موسسات دانشگاهی ، به شرط آماده سازی مناسب، بطور قابل توجهی عملی است.

برای دانشگاهها، کالجها، مدارس و موسسات آموزش از راه دور که بانکهای اطلاعاتی عظیمی دارند، ابزارهای داده کاوی میتواند الگوها و روابطی را که خیلی عیان نیستند آشکار کند. این نتایج ممکن است به طراحی دوباره  فرایندها و رویه های مرتبط منجر شود. تحلیلهای پشتیبانی شده توسط داده کاوی در کل موسسات و محیط ها میتواند مسائل متنوع مدیریت آموزشی؛ از جمله درک بهتر مشخصه های اقتصادی اجتماعی دانشجویان، مندرجات رشته ها و آموزش و پرورش و ساختار هزینه را  مخاطب قرار دهد.

آنچه لازم است اینست که فراتر از عملکرد داده پردازی استاندارد قدم برداریم مخصوصا" کتابخانه ها و موسسات دانشگاهی که با انواع مختلفی از بانکهای اطلاعاتی سروکار دارند و به سطوح معقولی از کامپیوتری کردن و دیجیتالی کردن داده ها دست یافته اند. در یک نظر، ابزارهای داده کاوی نمایانگر پیشرفت در زنجیره تکنولوژی اطلاعات هستند. داده کاوی همچنین میتواند بعنوان بخشی از فرایند بزرگتر کشف دانش در بانکهای اطلاعاتی در محیط های مختلف در نظر گرفته شود. البته نباید چنین پنداشت که ابزارهایی مثل داده کاوی نیاز به مداخله انسانی را کاهش خواهد داد.

همچنانکه در بالا نشان داده شد، ارزیابی و تعدیل نتایج بدست آمده بوسیله چنین ابزارهای خودکاری؛ به آزمایش نیاز دارد تا در برابر کاربردهای غلط محافظت شود.

انتظار میرود داده کاوی در گسترش سازمان خودیادگیرنده مشارکت کند.  کشف انتخابهای نوین با بهره گیری از داده کاوی اطمینان بخش بهترین کاربرد ممکن منابع موجود است. داده کاوی ماهیت چرخه مانند دارد. برای اینکه در پی کشف الگوها، سوالات بیشتری پدید خواهند آمد که دور بعدی فرایند را شکل میدهند. بهره برداری از تکنولوژیهای پیشرفته مثل داده کاوی مطمئنا" برای متخصصان کتابداری و مدیران موسسات آموزشی یک چالش دائمی خواهد بودچرا که آنها خلاقیت طلبند و برای نوآوری تلاش می کنند. [[11]](#endnote-2)

**2-10 الگوريتم هاي داده كاوي براي طبقه بندي دانشجويان**

در اين مقاله متد هاي مختلف و مهارت هاي داده كاوي را براي طبقه بندي دانشجويان بر اساس كاربرد مدل داده و ماركها شامل دوره مخصوصشان است را مقايسه مي كنيم.ما ابزار كاوي مخصوصي را براي ساختن پيكربندي و اجراي داده كاوي آسانتر براي مدرس توسعه مي دهيم.از داده هاي واقعي براي دوره Moodle با دانشجويان دانشگاه كردوبا استفاده مي كنيم.ما همچنين مهارت هاي پيش پردازش بر روي داده هاي عددي در دستورات را انجام مي دهيم تا تاييد كند كه اگر مدلهاي طبقه بندي كننده بهتر است ، بدست آورد.سرانجام ما ادعا مي كنيم كه مدهاي طبقه بندي كننده براي استفاده در آموزش مناسب است ودقيق و قابل درك براي مدرسان به منظور براي ساختن تصميم استفاده شود.

توانايي پيش بيني يا طبقه بندي عملكرد دانشجويان در محيط هاي آموزشي بر پايه وب خيلي مهم است.در عرصه وعده زياد ، اين هدف در داده كاوي (DM)بدست مي آيد.در حقيقت يكي از مفيدترين وظايف DM در يادگيري ،طبقه بندي است.اهداف اموزشي متفاوت در استفاده از طبقه بندي از قبيل : كشف استعداد گروهي دانشجويي با شخصيت هاي شبيه وعكس العمل به استراتژي وابسته به علم مخصوص،تا بدرفتاري دانشجويان يا بازي كردن را شناسايي كند،گروهاي دانشجويي كساني هستند كه با اشاره رانده شدند يا شكست خوردند و تصور غلط در فرايند دانشجويي پيدا كردند، شناسايي ياد دهنده با انگيزه كم و پيدا كردن عمل هاي مفيد در كم كردن نرخ حذف تصادفي ، براي پيش بيني يا طبقه بندي دانشجويان،زماني كه از سيستم مدرس هوشمند استفاده مي كنند.، غيره انواع مختلفي در متد هاي طبقه بندي و الگوريتم هاي هوش مصنوعي است كه در درآمد،ارزش يا نمره دانشجويان را پيشبيني مي كند.برخي مثالها : پيش بيني نمره دانشجويان (طبقه بندي در 5 كلاس: A,B,C,D,E,F) براي تست نمره از شبكه هاي عصبي استفاده مي شود.پيش بيني موفقيت آكادميك (كلاس هاي كه موفق هستند يا نيستند) با استفاده از تحليل توابع مميز.طبقه بندي دانشجويان با استفاده از الگوريتم ژنتيك يا پيش بيني نمره نهايي.پيش بيني موفقيت هاي آكادميك دانشجويان (طبقه بندي مانند كم،متوسط و زياد در خطرات كلاس ها)استفده از روش هاي مختلف داده كاوي .پيش بيني ارزش دانشحويان (كلاس هاي شكست خورده و گذشته) استفاده از مهارت رگرسيون در داده دانشگاه باز يونان يا استفاده از مدل هاي شبكه عصبي براي گزارش Moodle.

**2-10-1 پيش زمينه**

طبقه بندي يكي از مسايلي است كه به طور مكرر در DM و تحقيق هاي يادگيري ماشين مطالعه مي شود.شامل پيش بيني ارزشي از (طبقه بندي) خصوصيت ها(كلاس) بر اساس ديگي خصوصيت (پيش بيني خصوصيت ه)است. روشهاي متفاوت در طبقه بندي است از جمله:

* طبقه بندي آماري توليد كننده اي است كه آيتم هاي شخصي را جايگزين گروها بر اساس اطلاعات مقداري از آيتم هاي اصلي مي كند(مانند متغيير ،كاراكتر و غيره) و مجموعه آموزشي را با اصطلاحات قبلي برچسب گذاري مي كند.برخي مثالهادر الگوريتم هاي آماري ، تحليل هاي مشخص كننده خطي،ميانگين مربعات درجه دوم،هسته و نزديكترين همسايه K هستند.
* درخت تصميم ، مجموعه اي از وضعيت سازماندهي شده در ساختار سلسله مرتبي است .آن يك مدل پيش گويانه است كه مثالي از بقه بندي را از طريق مسير هاي مشخص شده با وضعيت ها از ريشه درخت تا وقتي كه به برگ برسد كه مشابه برچسب كلاس خواهد بود.يك درخت تصميم مي تواند به آساني مجموعه اي نقشه هاي طبقه بندي را تبديل كند.برخي از شناخته شده ترين الگوريتم هاي درخت تصميم ها ،C4.5 و CART است.
* مقايسه قانون فازي كه منطق فازي را با داده هاي غير خطي تفسير مي كند.
* شبكه هاي عصبي نيز مي تواند بر اساس مقايسه قانون باشد.يك شبكه عصبي مانند توصيف موازي از فرايند شبكه است.

**2-10-2 نتيجه هاي آزمايشي**

ما برخي آزمايش ها را به منظور اجراي الگوريتم هاي طبقه بندي غير مفيد براي پيش بيني ارزش نهايي دانشجويان بر اساس داده هاي معمول دانشجويان در سيستم يادگيري انجام داديم.هدف ما طبقه بندي دانشجويان با ارزش هاي نهايي يكسان در گروه هاي مختلف كه وابسته به نمره است.ما داده را از 438 دانشجويان دانشگاه كردوبا در نمره هاي Moodle (امنيت و بهداشت در كار،پروژه ها ، شركت هاي مهندسي ، برنامه نويسي براي مهندسي ،علم كامپيوتر و برنامه هاي مشخص) انتخاب كرديم.Moodle (<http://moodle.org>) يكي از تكراري ترين استفاده رايگان ازسيستم مديريت مفهوم يادگيري (LCMS) است.Discretization نيازمند به برخي الگوريتم هاست.ما روش هاي دستي (كه مجبور است نكات بريده شده را نشان دهد)را در ارزش خصوصيت ،برمي گزينيم.از 4 فاصله و برچسب استفاده مي كنيم.(FAIL:if value is <5 ; PASS: if value is >=5 and <7;GOOD: if value is >=7 and <9 ; and EXCELLENT : if value is >=9) به علاوه ما از متدهاي يكسان هم در همه خصوصيت ها با فاصله و برچسب ها استفاده مي كنيم (كم،متوسط،زياد) سپس هر ورژن از جدول خلاصه ها را كشف مي كنيم(با مقادير قطعي و عددي ) در فايل هاي متني و فرمت KEEl. بعد قسمتهاي فايل هاي كامل (فايل عددي و قطعي)در جفت اموزشي وتست فايلها را مي سازيم.هر الگوريتم استفاده لايه اي از 10 حالت اعتبار صليبي را ارزيابي مي كند.مجموعه داده ، تصادفي 10 زير مجموعه نامربوط از اندازه هاي مساوي را به روش هاي لايه اي تقسيم مي كند(با نگهداشتن وضعيت كلاس اصلي )در هر تكرار ، از هر 10 زير مجموعه به عنوان تست مجموعه و ديگر 9 زير مجموعه كه مجموعه آموزش را تركيب مي كند،استفاده مي شود.در اين كار ما نيز مسائل يادگيري از داده نامتعادل در نظر مي گيريم.ما مي گوييم داده نامتعادل است زماني كه برخي كلاس ها از لحاظ معني با ديگر رابطه در متغير هايي عددي فرق دارند.داده ما نامتعادل واضح است كه توصيف مي كند: EXCELLENT 3.89% , GOOD 14.15% , PASS 22.15% , FALL 59.81% يكي از متدها مكرر در يادگيري از داده نامتعادل استفاده مي شود و شامل تكرار ساده از داده يا از طريق ساده سازي كلاس هاي اقليت يا زير ساده سازي از اكثريت است.تا هر كلاس مساوي نمايش داده مي شود.زماني به طور واقعي با داده تعادل برقرار مي كنيم كيفيت شامل طبقه بندي است و معمولااصطلاحات دقيق را اندازه گيري مي كند.

**2-10-2 در نهایت**

در اين مقاله عملكرد و نااطميناني را از مهارتهاي مختلف داده كاوي براي طبقه بندي دانشجويان با استفاده از ابزار كاوي Moodle مقايسه كرديم زماني كه ما بعضي وظايف پيش پردازش همچون دادهاي متعادل را درخواست مي كنيم نشان داديم كه برخي الگوريتم ها اجراي طبقه بندي شان را بالا بردند اما ديگران اين طور نيست.ما نيز نشان داديم كه مدل طبقه بندي خوب مجبور است به هر دو دقت و قابل درك براي صنعت توجه كند.در آزمايشات آينده ما مي خواهيم مقايسه از هر مدل طبقه بندي را اندازه گيري كنيم و از داده با اطلاعات بيشتر درباره دانشجويان استفاده كنيم(مانند پروفايل و برنامه آموزشي) و با كيفيت بيشتر (داده كامل درباره دانشجوياني كه همه نمرات آنها فعال است).در اين روش مي توانيم از چگونگي كيفيت و كميت داده اندازه گيري كنيم كه مي تواند بر عملكرد الگوريتم ها تاثير بگذارد.سرانجام مامي خواهيم تست انجام دهيم و از ابزار از طريق مدرسان استفاده كنيم تا از موقعيت وابسته به آموزش وپرورش به منظور پيشرفت در شايستگي بهره بريم.

**2-10=3 مشكلات تجاري براي داده كاوي**

تكنولوژي داده كاوي مي تواند برنامه هاي كاربردي زيادي را انجام دهد.و به انواع مختلفي از سوالات تجاري جواب دهد.اين ليست مشكلات عمومي كمي را نشان مي دهد كه مي تواند استفاده از داده كاوي را حل كند:

تحليل هاي كره سازي: چه مشترياني بشتر دوست دارند رقابت را تغيي دهند؟ ارتباطات، بانك ،موسسات بيمه و رقابت هاي شديدي كه در طول روز دارند. به طور متوسط ، هر تلفن همراه جديد تعهد كننده هزينه شركت هاي تلفن بيشتر از 200 دلار در سرمايه گذاري بازاريابي است.هر موسسه بازرگاني دوست دارد همه مشتريان را تا حد امكان نگهدارد.تحليل كره سازي مي تواند به مديان بازارياب كمك كند تا دليل ي براي مشتريان كره سازي را بفهمد، مشتريان مربوط را نگهدارد و سرتنجام مشتريان وفادار را افزايش دهد.

**فروش متقاطع:** اهداف مشترياني كه دوست دار خريدن هستند چيست؟ فروش متقاطع ،تجارت مهمي براي تغيير در خرده فروشي است.بيشتر خرده فروشان ، مخصوصا خرده فروشان آنلاين از اين ويژگي براي فروش خودشان استفاده مي كنند.براي مثال ، اگر شما از مغازه كتاب فروشي آنلاين از قبيل Amazon.comيا Barnes andNoble.com براي كتاب خريداري كنيد شايد به اين نكته توجه كنيد كه اين وب سايت مجموعه اي از ارتباط ها درباره كتاب هاي مربوطه مي دهد.اين ارتباط ها مي توانداز تحليل داده كاوي بدست آمده باشد.

**تشخيص تقلب:** ادعاي متقلب نامطئن است؟كمپاني هاي نامطمئن هزاران ادعا را در روز پردازش مي كنند.اين امكان براي آنها دارد كه به هر موردي رسيدگي كنند.داده كاوي مي تواند به شناسايي ادعايشان كه بيشتر نادرست است كمك كند.

**مديريت خطر:** بايد براي اين مشتري وام تصويب كنيد؟

اين يك سوال عادي در سناريوي بانك است. تكنولوژي داده كاوي مي تواند از نمره براي سطح خطر مشتري استفاده كند.به مدير براي ساختن تصميم درست براي هر برنامه كاربردي كمك كنيد.

**تقسيم بندي مشتري:** مشتري من چه كسي است؟ تقسيم بندي مشتري اين كمك را به مدير شركت مي كند كه تا پروفايل مختلف مشتري ها بفهمد و برا اساس تقسيم بندي عمل مناسب انجام دهد.

**فصل سوم**

**الگوریتم های داده کاوی**

**3-1 استخراج الگوريتم ها(تجزيه و تحليل سرويس ها-داده كاوي)**

الگوريتم داده كاوي مكانيزمي است كه يك مدل داده كاوي ايجاد مي كند. به منظور ايحاد يك مدل از الگوريتم ، ابتدا مجموعهاي از داده ها را اناليز مي كند و الگو ها و گرايش ها را جستجومي كند.اين الگويتم از نتيجه آناليز ها ي تعريف الگو هايي از مدل mining استفاده مي كند. اينها پارامتر هايي هستند كه از يك سو كاربردي هستند و از سوي ديگر مجموعه داده كامل را از الگو ها و جزئيات آماري استخراج مي كنند.

مدل mining كه الگوريتم را ايجاد مي كنند مي تواند شكل هاي متفاوتي داشته باشند از جمله:

* مجموعه قاعده كه چگونگي توليدات را توصيف مي كند كه با هم در يك تبديل گروه بندي شده اند.
* توصيف درخت كه پيش بيني مي كند آيا مشتري يك محصول را خواهد خريد.
* مدل رياضي كه خريد ها را پيشبيني مي كند.
* مجموعه اي از خوشه بندي كه توصيف ميكند چگونه نمونه هاي پايگاه داده با هم وابسته هستند.

سرويس آناليز Microsoft SQL Server  برخي الگوريتم ها براي استفاده راه حل داده كاوي شما را فراهم مي كند.اين الگوريتم ها زير مجموعه اي از الگوريتم هايي هستند كه مي تواند براي دادهكاوي استفاده شود. شما نيز مي توانيد از الگوريتم هاي سه قسمتي استفاده كنيد كه از OLE DB براي تعريف داده كاوي استفاده مي كند.

**3-2 انواعي از الگوريتم ها داده كاوي**

Analysis Services شامل دنبال كردن انواع الگوريتم است:

* **الگوريتم هاي طبقه بندي** يكي يا بيشتر متغير هاي مجزا را پيش بيني مي كند كه بر اساس ساير خواص در پايگاه داده هستند.يك مثال از الگوريتم هاي طبقه بندي [Microsoft Decision Trees Algorithm](http://technet.microsoft.com/en-us/library/ms175312.aspx). است.
* **الگوريتم هاي بازگشتي** يكي يا بيشتر متغير هاي پيوسته را پيش بيني مي كند.از قبيل سود يا ضرر ، بر اساس خواص ديگر از پايگاه داده.يك مثال از الگوريتم هاي بازگشتي [Microsoft Time Series Algorithm](http://technet.microsoft.com/en-us/library/ms174923.aspx). است.
* **الگوريتم هاي تقسيم** داده را به گروه ها يا خوشه بندي ،اصطلاحي كه ويژگي هاي شبيه دارد ،تقسيم مي كند.مثالي از الگوريتم تقسيم [Microsoft Clustering Algorithm](http://technet.microsoft.com/en-us/library/ms174879.aspx). است.
* **الگوريتم پيوند** پيدا كردن كاراكتر هايي بين خاصيت هاي مختلف از يايگاه داده است.معمول تر است كه برنامه كاربردي از این نوع الگوریتم ها برای ایجاد نقشه های همبستگی استفاده کنند که می تواند در تحلیل بسته های فروش استفاده شود.مثالی از این الگوریتم در [Microsoft Association Algorithm](http://technet.microsoft.com/en-us/library/ms174916.aspx). است.
* **الگوریتم های تحلیل متوالی** خلاصه ای از تکرار متوالی یا داستان های فرعی در داده است از قبیل جریان مسیر وب.مثالی از الگوریتم تحلیل متوالی در [Microsoft Sequence Clustering Algorithm](http://technet.microsoft.com/en-us/library/ms175462.aspx) است.

**3-2-1 بکار بردن این الگوریتم ها**

انتخاب بهترین الگوریتم در استفاده از وظایف تجاری مخصوص می تواند تغییر کند.زمانی می توانید از الگوریتم های مختلف در اجرای برخی از وظایف تجاری استفاده می کنید که هر الگوریتم نتایج مختلف تولید کند و الگوریتم هایی که می توند بیشتر از یک نوع نتیجه تولید کند.برای مثال می توانید از الگوریتم درخت تصمیم میکروسافت که فقط برای پیش بینی نیست استفاده کنید .اما همچون یک روش شماری از ردیف های مجموعه داده را کاهش می دهد چون درخت تصمیم می تواند ردیف هایی که بر مدل کاوی نهایی تاثیر می گذارد را نشان دهد.شما نیز مجبور نیستید از الگوریتم های مستقل استفاده کنیددر ابتدا راه حل داده کاوی می توانید از برخی الگوریتم ها در کشف داده استفاده کنید و بعداز دیگر الگوریتم ها در پیش بینی درآمد خاص بر اساس داده استفاده کنید.برای مثال می توانی از الگوریتم خوشه ای استفاده کنید که الگوها را در گروه هایی که کم یا بیشتر مشابه هستند.و سپس از نتایج ای که تصمیم بهتر در مدل درخت ایجاد می کند استفاده می شود.م توانید از الگوریتم های چندگانه با راه حلی در ایجاد وظایف جدا استفاده کنید.برای مثال با استفاده از الگوریتم درخت بازگشتی اطلاعات پیش بینی مالی را نگهدارید و یک الگوریتم نشان راه را برای تحلیل بسته های فروش اجرا کنید.

**3-3 جرئیات الگوریتم**

دنبال کردن جدول پیوندهایی را فراهم می کند تا انواعی از اطلاعات قابل دسترسی در هر الگوریتم است:

* **توصیف الگوریتم پایه** شرح پایه ای را فراهم می کند که الگوریتم چیست و چگونه کار می کند؟با سناریوی تجاری جایکه الگوریتم مفید به نظر می رسد.
* **مراجع تکنولوژی** لیستی از پارامتر ها که می توانید مجموعه کنترل رفتار الگوریتم و نتایج را در مدل تغییردهید.جزئیات تکنولوژی اضافی درباره تاثیر الگوریتم ها ،عملکرد باکها و نیازهای داده فراهم شده است.
* **جستجوی یک مدل** مثالهایی از جستجو که می تواند در هر نوع مدل استفاده می شود.می توانید از جستجو در یک مدل با یادگیری بیشتر درباره الگوهایی از مدل یا ساختن پیش گویی بر اساس این الگوها است.
* **مفهوم مدل کاوی** چگونگی اطلاعات ذخیره شده در ساختار معمول برای همه انواع مدل هارا توصیف می کند. و چگونگی تفسیر اطلاعات را توضیح می دهد.بعد از اینکه یک مدل را ساختین می توانید مدل ها را از طریق بیننده با BIDevelopment Studio فراهم کنید یا می توانید از جستجو ها در برگشت اطلاعات به طور مستقیم در محتوای مدل از طریق استفاده با DMX بنویسید.

**3-4 طراحي SPAM:**

**3-4-1 ابزاري براي کاوش در شبکة گستردة جهان**

کشف و استخراج الگوهاي مفيد از شبکة گستردة جهاني که با عنوان وب‌کاوي مطرح مي‌شود، از کاربردهاي عمدة داده‌کاوي به شمار مي‌رود. موتورهاي جستجوي امروزي، چه آنها که از نوع کاوشگر وب هستند و چه آنها که از نوع دروازة وب هستند، ديگر از عهدة پاسخگويي به جستجوي کاربران در اين کتابخانة عظيم نامنظم برنمي‌آيند. در اين مقاله بازيابي هر چه سريعتر و بهتر اطلاعات از شبکة گستردة جهاني مورد توجه قرار گرفته و به همين منظور يك معماري كلي براي سيستمهاي داده‌كاوي ارائه مي‌شود و با معرفي جايگاه وب‌كاوي در داده‌كاوي و بررسي مشكلات موجود در كشف مؤثر اطلاعات از وب راه‌حلهاي موجود براي برخورد با هر يك از اين مشكلات در قالب جداولي به نامهاي SPAM (Solution\_Problem Analysis Matrix) و T&A (Tools and Approaches) ارائه مي شوند. SPAM ابزاري است که چارچوب مورد نياز براي انتخاب روش مناسب براي پياده‌سازي يک سيستم وب‌کاوي را در اختيار قرارمي‌دهد و T&A ابزاري است که براي انتخاب ابزارها و روشهاي مناسبي بکار مي‌رود که براي ساخت يک سيستم وب کاوي در دسترسند. اين ابزارها در سيستمي به نام AUT\_UniversitiesPortal که دروازه‌اي براي دستيابي به مجموعة طبقه‌بندي ‌شده‌اي از سايتهاي دانشگاهاي مختلف دنيا است، مورد تست و ارزيابي قرار گرفته‌اند.

**3-4-2 تعریف**

ما در «عصر اطلاعات» زندگي مي‌کنيم، عصري که در آن انسانها بيش از هر زمان ديگري در گذشته داده و اطلاعات توليد و منتشر مي‌کنند. در واقع اطلاعاتي که در اختيار ما قرار دارد خيلي بيشتر از آن است که بتوانيم آن را تجزيه و تحليل کنيم. به اين ترتيب انتخاب منابع مورد نظر كاربر از اين حجم رياد اطلاعات روز به روز مشكلتر مي شود، که اين مسئله به دلايل مختلفي است که از آن جمله مي توان به موارد زير اشاره کرد:

1- وب خيلي بزرگ و متنوع است. مطالعه‌ای نشان داده است که بيش از 10 بيليون صفحه با دسترسي عمومي در وب وجود دارد. علاوه بر اين، تقريباً هر ماه 6 ترا بايت اطلاعات جديد به وب اضافه مي‌شود.

2- داده هاي وب به سرعت در حال تغييرند. در حيني که وب به سرعت از لحاظ اندازه رشد مي کند، اطلاعات آن هم مدام تغيير مي‌کنند و به روز مي‌شوند.

3- وب خيلي بد سازماندهي شده است. گر چه ممکن است بخشهاي کوچکي از وب خوب سازماندهي شده باشند و به خوبي نگهداري شوند اما کل وب کاملاً غير ساخت يافته است.

علاوه بر اينها وب کاربران خيلي متفاوتي داردو هرکاربری ممکن است تنها به بخش کوچکي از وب علاقمند باشد. در نتيجه، کاربران بر خط مشکلات زيادي در يافتن اطلاعات موردنظرشان دارند.

موتورهاي جستجو به کاربران در يافتن منابع اطلاعاتي مرتبط کمک مي‌کنند. اما موتورهاي جستجو داراي مشکلات زيادي هستند. يک عنوان ممکن است صدها يا هزاران سند را شامل شود که باعث می‌شود موتور جستجو تعداد زيادي سند شامل اسناد غيرمرتبط برگرداند. تعداد زيادي از اسنادي که کاملاً به عنوان مورد نظر وابسته هستند هم ممکن است داراي کلمات کليدي که آن موضوع را تعريف مي‌کند، نباشند.

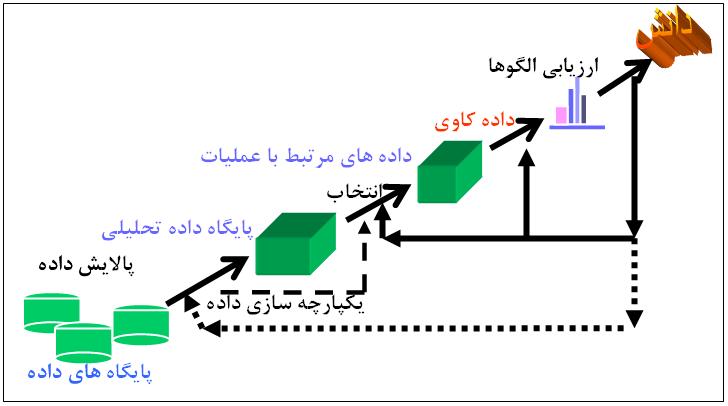
مجموعه اطلاعات موجود در وب منابع غني براي وب‌کاوي بوجود مي‌آورند که مي‌تواند به ما در پيدا کردن و استخراج منابع اطلاعاتي با بهبود عملکرد موتورهای جستجو کمک کند.

در اين مقاله ابتدا يك بررسي اجمالي بر داده‌كاوي و مراحل انجام آن خواهيم داشت و بر اساس مطالعات انجام شده يك معماري كلي براي داده‌كاوي ارائه مي شود. در ادامه با معرفی جايگاه وب‌كاوي در داده‌كاوي و بررسي مشكلات موجود در كشف مؤثر اطلاعات از وب ابزارهای SPAM و T&A برای حل این مشکلات ارائه می‌شوند. در نهايت نتايج بررسي‌هاي انجام شده در راستاي بکارگيري تکنيکهاي داده‌کاوي و بخصوص روشهاي وب‌كاوي در موتورهاي جستجو به منظور بهبود بخشيدن به عمليات جستجو و دريافت نتايج به اندازة کافي معتبر از اين جستجو ارائه خواهد شد.

3-4-3 داده‌كاوي

داده‌كاوي كه با عنوان كشف دانش از پايگاه‌هاي داده هم شناخته مي‌شود، فرايندي براي استخراج الگو هاي مفيد از پايگاه‌هاي داده مي‌باشد[2]. داده‌کاوي مي تواند الگوهاي مفيد مورد نظر کاربران خود را از انواع مختلفي از پايگاه‌هاي داده استخراج کند. بيشتر محققان داده‌كاوي را مترادف با كشف دانش در پايگاه‌هاي داده مي‌دانند. كشف دانش شامل مراحل زير است كه همانطوركه در شكل 1 نشان داده شده است به صورت پي در پي انجام مي شوند:

* + پالايش داده:‌ آشفتگي‌ها و دادة ناسازگار را حذف می‌کند.
* يکپارچه‌سازي داده: درصورت لزوم منابع داده‌اي را ترکيب می‌کند.
* تبديل داده: داده را به فرم مناسب براي داده كاوي تبديل مي‌كند.
* داده كاوي:‌ فرايندی ضروري است كه در آن متدهاي هوشمند الگوهاي داده‌اي مناسب را استخراج مي‌كنند.
* ارزيابي الگو: الگوهاي استخراج شده را ارزيابی می‌کند.
* نمايش دانش: در اين مرحله تكنيكهاي مختلف نمايش دانش براي نشان دادن دانش كشف و کاوش شده به كاربر استفاده مي‌شوند.

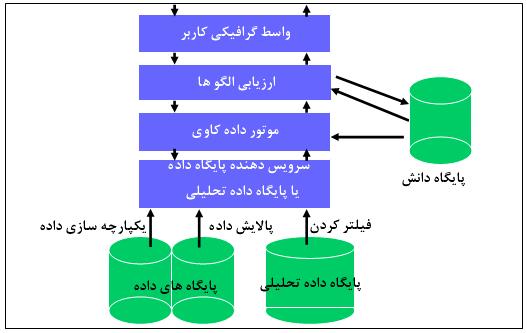


شکل 3-1 مراحل عمليات کشف دانش از پايگاه‌هاي داده

3-5 معماري يك سيستم داده‌كاوي

معماريهاي مختلفي براي سيستمهاي داده‌كاوي ارائه‌ شده‌اند[2، 3، 19، 20]. بطور كلي معماري يك سيستم داده كاوي كه در شكل 2 هم ارائه شده است، داراي مؤلفه هاياصلي زير است:

* پايگاه داده، پايگاه دادة تحليلي، يا مخزن اطلاعاتي ديگري: اين جزء شامل يك يا چند پايگاه داده، پايگاه دادة تحليلي يا مخازن اطلاعاتي ديگري است كه تكنيكهاي پاك كردن و يکپارچه سازي داده روي آنها اعمال مي‌شود.
* سرويس دهندة پايگاه‌داده: اين سرويس دهنده مسئوليت واكشيداده را براساس درخواست داده‌كاوي كاربر بر عهده دارد.
* پايگاه دانش: اين مخزن حوزه يا دامنة دانش است، كه براي هدايت عمل جستجو يا ارزيابي الگوهاي بدست آمده استفاده مي‌شود.
* موتور داده كاوي: شامل مجموعه‌اي از مؤلفه‌های عملياتي جهت تعيين ويژگيها، تعيين قوانين انجمني، توصيف ساختار داده‌ها از طريق خوشه بندي و تحليل ارزيابي است.
* مؤلفه ارزيابي الگوها: اين مؤلفه مرتبط بودن الگوها را بررسي مي‌كند و با ديگر مؤلفه‌های داده‌كاوي در ارتباط است، به طوريكه جستجوي الگوهاي مورد نظر به درستي پيش برود.
* واسط كاربري گرافيكي: اين مؤلفه امکان مرور پايگاه‌هاي داده‌ای، جستجوی الگوها و ارزيابی الگوهاي كشف و استخراج شده را فراهم می‌کند‌.



شکل 3-2 معماري يک سيستم داده‌کاوي

3-6 وب‌كاوي

وب‌كاوي از مهمترين کاربردهاي داده‌کاوي در کاوش شبکة گستردة جهاني براي کشف و استخراج الگوهاي مفيد است. وب علاوه بر اينکه داراي مجموعه عظيمی از اطلاعات است، حاوی مجموعه‌ای پويا از پيوندها براي دسترسي به صفحات وب و استفاده از اطلاعات نيز می‌باشد که يک مجموعة غني براي داده‌کاوي ايجاد مي‌کند.

3-6-1 انواع عملياتي که در وب‌کاوي انجام مي‌شوند

به طور کلي عمليات مورد نظر در وب‌کاوي شامل کاوش محتوايي وب، کاوش ساختاری وب و کاوش کاربردی وب است. بايد توجه داشت كه اين شاخه هاي مختلف كاملاً از هم مجزا نيستند. از اين عمليات براي حل مشكلات مطرح شده در كاوش شبكة گستردة جهاني استفاده مي‌شود. نحوة تشخيص عمليات مناسب براي حل مشكلات موتورهای جستجو در جدول 1 تحت عنوان جدول SPAM ارائه شده است. *در اين جدول مسائل موجود و کاربردهای وب‌کاوي به ‌همراه روشهايي که براي حل آنها وجود دارد از مقالات مختلف جمع آوري شده و در قالب جدول SPAM ارائه‌شده است. با استفاده از اين جدول و با توجه به دسته بندي انجام شده، به راحتي مي توان عمليات مورد نياز مسائل موجود را استخراج کرد. لازم به ذکر است که خانه هاي خالي نشان دهندة اين است که در رابطه با مسئلة عنوان شده هيچ روشي از آن دسته عمليات وجود ندارد.*

در ادامة اين بخش كاربردها، نحوة اجرا و نتيجة اجراي عمليات مذكور به اجمال مورد بررسی قرار می‌گيرند.

3-6-2 كاوش محتوايي وب

كاوش محتوايي وب كشف اطلاعات مفيد از اسناد داده و محتواي وب مي باشد. آنچه با عنوان محتواي وب شناخته مي‌شود، مي‌تواند شامل انواع مختلفي از داده باشد.

**جدول 3-1 جدول** SPAM **براي انتخاب عمليات وب كاوي مناسب در رابطه با حل مشكلات مختلف موجود در كشف و استخراج مؤثر اطلاعات از شبكة گستردة جهاني**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **كاوش محتوايي وب** | **كاوش ساختاري وب** | **كاوش در كاربردهاي وب** |
| **دسترسي به اطلاعات مورد نظر كاربر از طريق موتورهاي جستجو با وجود بزرگي و پويايي وب** | **روشهاي متن كاوي**  **بخش بندي اسناد وب**  **روشهاي بازيابي اطلاعات** |  |  |
| **افزايش كيفيت سرويسهاي اطلاعاتي اينترنت** |  | **كشف اعتبار صفحات از طريق ساختار پيوندهاي هر صفحه و Hub** | **كاوش گزارشهاي دسترسي به سرويس دهنده هاي وب و متا اطلاعات** |
| **كاربر فقط بخشي از وب را كه به مربوط اوست ببيند.** | **بخش بندي اسناد وب** |  | **كاوش گزارشهاي دسترسي به سرويس دهنده هاي وب و متا اطلاعات** |
| **مدل كردن كاربر بر اساس علايق و اهداف او** |  |  | **كاوش گزارشهاي دسترسي به سرويس دهنده هاي وب و متا اطلاعات** |
| **اعتبار دهي صفحات وب** |  | **كشف اعتبار صفحات از طريق ساختار پيوندهاي هر صفحه و Hub** |  |
| **سازماندهي مناسب وب** | **طبقه بندي صفحات وب** | **بخش بندي وب از طريق بررسي ساختار پيوندها بين صفحات مختلف** | **كاوش گزارشهاي دسترسي به سرويس دهنده هاي وب و متا داده** |

قبلا اينترنت انواع مختلفي سرويس و منابع داده داشت مثلFTP، Use Net و Gopher. امروزه بيشتر اين داده‌ها از طريق وبدر دسترس هستند. در چند سال اخير اطلاعات دولتي به صورت عظيمي رشد داشته است،محتواي وب شامل داده غيرساخت‌يافته مثلمتون آزاد و داده نيمه‌ساخت‌يافته مثل اسناد HTML و يك نوع داده ساخت‌يافته‌تر مثل داده موجود در جداول يا پايگاه‌داده‌هاي توليد کننده صفحات HTML است. در هر حال بيشتر محتواي وب داده غيرساخت‌يافته است. تحقيقات دربارة بكاربردن تكنيكهاي داده كاوي ازمتون غير ساخت يافته با عنوان متن‌كاوي ناميده مي شود. بنابراين مي‌توانيم متن‌كاوي را به عنوان نمونه اي از كاوش محتواي متن بدانيم. ما مي‌توانيم تحقيقات انجام شده در زمينه كاوش محتوايي وب را از دو نقطه نظر متفاوت بررسي كنيم: ديدگاه‌هاي بازيابي اطلاعات و پايگاه‌هاي داده. هدف كاوش محتوايي وب از ديدگاه بازيابی اطلاعات اساساً كمك به بهبود جستجوي اطلاعات يا فيلتر كردن اطلاعات براي كاربران و معمولاً بر اساس پروفايلهايي استنباط شده يا تقاضا شده كاربران است. در حاليكه هدف كاوش محتوايي وب از نقطه نظر پايگاه‌داده اساساً تلاش براي مدل كردن داده روي وب و استفاده از آن به نحوي است كه بتوان به جاي جستجو هاي ساده كه تنها براساس كلمات كليدي هستند پرس و جوهاي پيشرفته‌تري داشت.

3-6-3 کاوش ساختاري وب

كاوش ساختاري وب به دنبال كشف مدلي است كه در زيرساختار پيوندهاي وب وجود دارد. اين مدل بر اساس توپولوژي ابرپيوندها است. اين مدل مي تواند براي بخش‌بندي صفحات وب بكار رود و مي‌تواند در توليد اطلاعاتي مثل تشابه و ارتباط بين وب‌سايتهاي مختلف مفيد باشد. كاوش ساختاري وب مي تواند براي كشف اعتبار صفحات وب استفاده شود. تعيين اعتبار صفحات وب در پيوندهايي که به آن صفحه وجود دارد، پنهان است. وب تنها شامل يک سري صفحات جدا از هم نيست بلکه مهمترين ويژگي آن پيوندهايي است که صفحات را به هم متصل مي‌کنند. اين پيوندها داراي مقدار زيادي اطلاعات نهفته هستند که در ايجاد اين تصور که اعتبارسنجي صفحات به صورت خودکار انجام مي‌شود، کمک زيادي مي‌کنند. در واقع وقتي نويسندة يک صفحة وب پيوندي به صفحة ديگري در صفحه‌اش مي گذارد اين پيوند به منزلة مهر تأييدي است که اين نويسنده روي آن صفحه زده است. هر چه يک صفحة وب داراي پيوندهاي بيشتري از جاهاي مختلف باشد، اين نشان‌دهندة اهميت بيشتر آن صفحه است و طبعاً اعتبار بيشتر آن صفحه را نشان مي‌دهد. بنابراين پيوندهاي وب اطلاعات غني و کاملي در رابطه با مرتبط بودن، کيفيت و ساختار محتوايي وب ايجاد مي‌کند و يک منبع غني براي وب کاوي محسوب مي‌شود. مطرح شدن اين ايده موجب شده است که مطالعات زيادي در رابطه با کاوش صفحات معتبر در وب انجام شود. در سالهاي 1970 محققان به منظور ارزيابي مقاله‌هاي تحقيقاتي روشهايي را مطرح کردند که اساس آنها استفاده از اظهار نظرهاي موجود در مجلات معتبر در رابطه با آن مقاله بود. اما بر خلاف اظهار نظرهايي که در ژورنالها ارائه مي‌شود، پيوندهاي وب ويژگي‌هاي خاصي دارند که استفاده از آنها را براي تعيين اعتبار صفحات با مشکل مواجه مي‌کند. اول اينکه هر پيوند نشاندهندة تأييدي که ما به دنبال آن هستيم نيست. در واقع بعضي از اين پيوندها به هدف ديگري در نظر گرفته شده‌اند. دوم اينکه ممکن است به دلايل تجاري يا رقابتي يک صفحه داراي پيوندي از طرف نويسندگان رقيب نباشد که اين مسئله سبب پايين آمدن اعتبار آن صفحه مي‌شود. براي مثال ممکن است شرکت کوکاکولا ترجيح دهد که شرکت رقيبش يعني پپسي را با پيوند زدن به صفحات وب آن تأييد نکند. سوم اينکه صفحات معتبر معمولاً توصيفي از خودشان دارند. براي مثال صفحة اصلي Yahoo ممکن است داراي توصيفي با عنوان «موتور جستجوي وب» نباشد.

3-6-4 کاوش کاربردی وب

در کاوش کاربردی وب رکوردهاي گزارشهاي وب براي کشف الگوهاي دسترسي کاربر به صفحات وب بررسي مي‌شوند. در حاليكه کاوش محتوايي وب و كاوش ساختاري وب از داده اوليه و واقعي وب استفاده مي‌كنند، كاوش كاربردي وب داده ثانويه مشتق شده از فعل و انفعالات كاربران با وبرا مي كاود. اطلاعات كاربردي وب شامل اطلاعات گزارشهاي دسترسي به سرويس دهنده هاي وب، گزارشهاي سرويس‌دهندگانproxy، گزارشهاي كاوشگرها، پروفايلهاي كاربران،اطلاعات تراكنشها يا نشستهاي كاربران، كوكي ها، پرس و جوهاي کاربران، ليستهاي آدرسهاي مورد نظر كاربران، كليك هاي موس و هر نوع داده ديگري را كه بتوان به عنوان نتيجه يك کار كاربر در نظر گرفت، مي باشد. تحليل و پيدا کردن نظم موجود در رکوردهاي گزارشهاي وب مي‌تواند مشتريهاي احتمالي موجود در تجارت الکترونيکي را مشخص کند، کيفيت سرويسهاي اطلاعاتي اينترنت را براي کاربران افزايش دهد و کارايي سيستم سرويس دهندة وب را بهتر کند.

3-6-5 روشها و ابزارهاي وب كاوي

براي اجراي عمليات مختلف وب كاوي روشها و ابزارهاي مختلفي بكار برده شده اند. تا اينجا با توجه به جدول SPAM مشخص شد كه هر يك از عمليات وب‌كاوي چه كاربردي دارند و براي حل كداميك از مشكلات موجود در استخراج اطلاعات از وب مناسبند. در اين بخش ابزارها و روشهاي مناسب براي انجام هر يك از اين عمليات در قالب جدولي با نام جدول T&A ارائه مي شوند كه در جدول 2-2 نشان داد.

جدول 3-2 جدول T&A ابزارها و روشهاي وب كاوي

|  |  |  |  |
| --- | --- | --- | --- |
| روشها و ابزار وب كاوي | | تکنيکها | كاربرد |
| **روشهاي وب کاوي** | روشهاي متن كاوي | قوانين اپيزودي، روشهاي پردازش زبانهاي طبيعي، ابزارهاي شبكه هاي عصبي، ابزارهاي خوشه بندي | كاوش محتواي متون غير ساختيافته وب براي بهبود كارايي موتورهاي جستجو |
| روشهاي بازيابي اطلاعات | مدل ماركوف نهان(HMM)، نقشه هاي خود سازمان ده (SOM)، انتروژي بيشينه، روشهاي استقرايي | كاوش محتواي اسناد وب براي بهبود كارايي موتورهاي جستجو، پيدا كردن الگوهايي در متون اسناد وب |
| روشهاي بخش بندي متن | Rocchio، نزديكترين K همسايه ها (KNN)، بردارهاي پشتيباني ماشين (SVM)، روشهاي استقرايي، سيستمهاي خبره، شبکه هاي عصبي، يادگيري آماري | كاوش محتواي اسناد وب براي بهبود كارايي موتورهاي جستجو |
| كاوش قوانين انجمني | قوانين انجمني | مدل كردن كاربر، مشخص كردن شماي وب سايت براي سازماندهي مناسب وب |
| روشهاي طبقه بندي | درختهاي تصميم گيري، قانون بيز، KNN، الگوريتمهاي ژنتيك، شبکه هاي عصبي | طبقه بندي اسناد وب |
| روشهاي خوشه بندي | روشهاي جزء بندي، روشهاي سلسله مراتبي، روشهاي برمبناي تراكم، روشهاي شبكه اي | خوشه بندي اسناد وب و تعيين بخشهاي اوليه به منظور بخش بندي |
| **ابزار** | روشهاي يادگيري آماري | رگرسيون، كوچكترين مربع خطي مناسب (LLSF) | بخش بندي اسناد وب، مدل كردن كاربر |
| شبكه هاي عصبي | شبكه انتشار رو به عقب | بخش بندي اسناد وب |
| سيستمهاي خبره | CONSTRUE، MedIndex | بخش بندي اسناد وب، مدل كردن كاربر |

براي هر روش تکنيکهاي مختلف آن ارائه شده است. با مراجعه به اين جدول مي توان با توجه به زمينة عملياتي و کاربرد مورد نظر، روش مناسب را انتخاب کرد. براي انتخاب تکنيک مناسب از بين تکنيکهاي معرفي شده بايد از مقياسهاي ارزيابي استفاده کرد. در ادامة اين بخش مقياسهاي اصلي براي ارزيابي روشهاي مختلف ارائه شده است.

3-6-6 مقياسهاي اصلي براي مقايسه روشهاي وب كاوي

فرض كنيد يک سيستم وب كاوي تعدادي سند را بر اساس ورودي كه به صورت يك پرس و جو بوده است،بازيابي كرده است. چطور مي توانيم تشخيص دهيم كه سيستم چقدر دقيق يا درست بوده است؟

مجموعه اسناد مرتبط با يك پرس و جو را با {Relevant} نشان مي دهيم، و مجموعه اسناد بازيابي شده را با {Retrieved} نشان مي دهيم. مجموعه اسنادي كه هم مرتبط هستند و هم بازيابي شده اند با {Retrieved}{Relevant} نشان داده مي شوند.

دو مقياس اصلي براي تشخيص كيفيت بازيابي متن وجود دارد:

* دقت: در صد اسناد بازيابي شده كه در واقع به پرس و جو مرتبط هستند (براي مثال پاسخهاي «درست») اين مقياس به صورت زير تعريف مي شود:

(1) 

* ياد آوري: در صد اسنادي كه به پرس و جو مرتبط هستند و در واقع بازيابي شده اند. اين مقياس به صورت زير تعريف مي شود:

(2) 

3-7 انواع موتورهاي جستجو

موتور جستجو ابزاري است که امکان جستجوي اطلاعات در وب را فراهم مي‌کند. موتورهاي جستجو در واقع راه حل ساده‌اي براي يافتن اطلاعات در اينترنت ارائه مي‌دهند. موتور جستجو مستقيماً شبکة گستردة جهاني را جستجو نمي کنند بلکه پايگاهي از صفحات وب کاملاً متني انتخاب شده از بيليونها صفحة وب روي سرويس دهنده را جستجو مي کند. موتورهاي جستجو به دو دستة اصلي کاوشگران وب (Web Crawler) و دروازه‌هاي وب (Web Portal) تقسيم‌بندي مي‌شوند که اين دو نوع موتور جستجو بر اساس خصوصيات ويژه‌اي که دارند، هر يک موارد کاربرد مخصوص به خود را دارند. در اين بخش اين موتورهاي جستجو مورد بررسي قرار خواهند گرفت و در نهايت تفاوتهاي اصلي آنها را برخواهيم شمرد و کاربردهاي آنها را خواهيم ديد.

3-7-1 کاوشگر وب

کاوشگرهاي وب از همان آغاز پيدايش وب به منظور جستجوي صفحات وب به وجود آمدند و اولين کاوشگر با نامMatthew در بهار 1993 نوشته شد. کاوشگر وب به اين صورت عمل مي‌کند که از چند آدرس پيش فرض وب شروع کرده و صفحات مربوط به آنها را از وب دريافت مي‌کند. سپس پيوندهاي هر صفحه را استخراج کرده و آنها را در يک صف قرار مي‌دهد تا در دفعات بعد آنها را بررسي کند. به اين ترتيب تمام کلمات و عبارات صفحات در يک پايگاه داده شاخص‌بندي مي‌شوند. اين پايگاه داده مي‌تواند بعداً به منظور پاسخگويي به پرس و جوهايي براي عبارات موجود در صفحه مورد جستجو قرار گيرد. هر کلمه يا عبارتي که در صفحات وب وجود دارد، وقتي در پايگاه داده ذخيره شد، براي کاربر قابل جستجو خواهد بود و بدينوسيله مي‌توان هر کلمه يا عبارتي را در تمام اينترنت جستجو کرد. پايگاه دادة کاوشگرها توسط برنامه اي به نام spider ساخته مي شود.

اطلاق واژة کاوشگر به اين عمليات به اين دليل است که در طي اين عمليات تمامي پيوندهاي درون هر صفحه براي ارجاعات بعدي استخراج مي‌شوند. البته اگر چه گفته مي شود که آنها وب را براي پيدا کردن صفحات مورد نظر کاوش مي کنند اما آنها در واقع در يکجا مي مانند و صفحات را با دنبال کردن لينکهايي که از قبل در پايگاهشان دارند، پيدا مي کنند. از لحاظ تئوري کاوشگر وب مي‌تواند از يک صفحه شروع کند، تمامي پيوندهاي درون آن را استخراج کند و آن پيوندها را هم به نوبت جستجو کند تا اينکه تمامي صفحات اينترنت را مورد بررسي قرار دهد. اين استراتژي مشکل کاملاً واضحي دارد که آن را به اين صورت بيان مي‌کنيم که نمي‌توان از يک نقطة شروع به همة بخشها وصفحات اينترنت دست پيدا کرد، چون بعضي از صفحات از بعضي ديگر هيچ ارجاعي ندارند و کاوشگر اينقدر هوشمند نيست که بداند کدام مسيرهاي ديگر را بايد بررسي کند تا تمام اينترنت را کاوش کند. در واقع اگر صفحه اي از هيچ صفحة ديگري پيوند نداشته باشد،spider نمي تواند آن را پيدا کند. تنها راه موجود براي اينکه يک صفحه توسط کاوشگر مورد جستجو قرار گيرد، اين است که براي او فرستاده شود.

کاوشگرهاي وب معمولاً به صورت خودکار عمل مي‌کنند و کمتر توسط انسانها نگهداري مي‌شوند. اين خصوصيات اين موتورهاي جستجو را بعنوان وسيله‌اي بسيار مناسب براي يافتن اطلاعات خاص اما روشي نيمه کارا در يافتن اطلاعات عمومي و مشترک مطرح مي‌کند.

3-7-2 دروازه وب

براي دروازه وب تعاريف مختلفي ارائه شده است و در رابطه با ارائة يک تعريف کلي براي آن اختلاف نظرهاي زيادي وجود دارد ،اين اختلاف نظرها حتي در رابطه با نامي که براي اين ابزار در نظر گرفته شده است هم وجود دارد يعني خيلي ها معتقدند که واژة “Portal” که در اين مقاله آن را به معناي «دروازه» گرفته‌ايم، واژة مناسبي براي اين ابزار نيست و نبايد استفاده شود. اما در رابطه با خصوصيات ويژة آن و عملياتي که بايد انجام دهد يک توافق کلي وجود دارد. لازم به ذکر است که صفحات وب ايستايي که پيوندهايي براي دسترسي به منابع ناهمگون در رابطه با يک موضوع خاص در اختيار قرار مي دهند، در واقع دايرکتوريها هستند و نبايد با دروزة وب اشتباه شوند. دروازه هاي وب داراي پيچيدگي و توانايي بيشتري در مقايسه با اين دايرکتوريها هستند. در واقع دروازة وب تنها يك موتور جستجو نيست بلكه از اجزاء مختلفي تشكيل شده است و امكانات مختلفي را در اختيار كاربران قرار مي دهد كه موتور جستجو يكي ازاين امكانات است.

بطور كلي دروازة وب را به اين صورت تعريف مي‌کنيم که «دروازة وب به پايگاه وبي گفته مي‌شود که اطلاعات را با توجه به موضوع آنها دسته‌بندي مي‌کند تا به اين صورت به يافتن آنچه به دنبال آن هستيم، کمک کند». به بياني ديگر يک دروازة وب مجموعه‌اي از صفحات وب است که به هم پيوند خورده‌اند و اين مجموعه به صورتي تشکيل شده است که سازمان اطلاعات، دستيابي به اطلاعات و فعل و انفعال با آن را در وب ساده مي‌کند. اين صفحات يک نقطة ورود واحد ايجاد مي‌کنند که کاربرانشان را قادر مي‌کند تا اطلاعات منابع مختلفي را به صورت يکجا ببينند. هر بار که صفحه‌اي از دروازه در وب نمايش داده مي‌شود به صورت خودکار گردآوري و قالب‌بندي مي‌شود تا با آخرين محتوا و قالب‌بندي که براي آن صفحه انتخاب شده است، هماهنگ باشد. دروازه‌هاي بزرگ وب در تلاشند تا تمام دسته‌بندي‌هاي مربوط به يک موضوع را فراهم آورند، اما هميشه موفق به انجام اين کار نمي‌شوند. دروازه‌ها هم امکان جستجو در بايگانيشان را مانند کاوشگرها فراهم مي‌کنند، اما فقط خلاصه و عناوين پايگاههاي وب را جستجو مي‌کنند و محتواي آنها را جستجو نمي‌کنند. دروازه‌هاي وب براي يافتن اطلاعات عمومي و کلي مناسب هستند، اما آنها قادر به ساماندهي همه چيز در وب نيستند و بنابراين يافتن اطلاعات خاص از طريق آنها به سادگي امکانپذير نيست. اين مسأله از اولين مسائلي است که هنگام انتخاب بين کاوشگر و دروازة وب براي يافتن اطلاعات بايد در نظر گرفته شود.

ليست نتايج در دروازه‌هاي وب بر اساس اولويت احتماليشان مرتب مي‌شوند و ممکن است اين ترتيب هميشه بر اساس ميزان ربط به موضوع و بهتر بودن نباشد، و هزينه‌اي که موتور جستجو براي آنها مي‌پردازد و مسائل ديگر هم در رتبه‌بندي صفحات مؤثرند. تعداد بسيار کمي از موتورهاي اين دسته بيش از 10 درصد وب را مي‌پوشانند و هيچکدام از آنها بيش از 20 درصد را هم نمي‌پوشانند. موتورهاي جستجو با دامنه و موضوع مشخص با احتمال بيشتري به طور کامل موضوعشان را پوشش مي‌دهند. از دروازه‌هاي مشهور مي‌توان Yahoo و Looksmart را نام برد.

3-7-3 مقايسة موتور جستجوهاي دروازه وب و کاوشگر وب

همانطور که در بخشهاي قبل گفته شد، دروازة وب و کاوشگر وب هر کدام داراي خصايص و ويژگيهاي معيني هستند. در اين بخش بر اساس جدولي اين دو دستة عمده از موتورهاي جستجو را با هم مقايسه خواهيم کرد. نتيجه اين مقايسه در جدول 3 نشان داده شده است.

3-7-4 در نهایت

افزايش توانايي تكنيكها و ابزارهاي مختلف در ايجاد و جمع‌آوري داده‌ها و اهميتي كه پايگاه‌هاي داده به دليل در دسترس بودن و قوي بودنشان در صنايع و تحقيقات مختلف دارند، همچنين شبکة گستردة جهاني که به عنوان يک منبع اطلاعاتي مهم بشمار مي‌رود، ما را با حجم عظيمي از داده و پايگاه‌هاي داده روبرو ساخته است. علاوه بر اين امروزه در کشف مؤثر اطلاعات از وب هم مشکلات بسيار زيادي مطرح است، که مي‌توان به حجم بسيار زياد اطلاعات در وب، پيچيدگي صفحات وب در مقابل اسناد متني معمولي، پويا بودن اطلاعات موجود در آن و برقراري ارتباط بين ميليونها کاربر در سراسر دنيا به عنوان مهمترين اين مشکلات اشاره کرد. با وجود اين حجم بسيار زياد اطلاعات، تنها بخش کوچکي از آن مي‌تواند براي شما مفيد باشد. اما چگونه مي‌توان بخشي از وب را که مورد نظر شماست، محاسبه کرد؟ چگونه مي‌توان صفحات وب با کيفيت بالا را براي يک موضوع پيدا کرد؟ اين مشکلات منجر به تحقيقاتي دربارة کشف و استفادة مؤثر وکارا از منابع در اينترنت شده است. نهايتاً وب کاوي با سه دسته عمليات عمده که عبارتند از: کاوش در محتواي وب، کاوش در ساختار وب و کاوش در استفاده‌هايي که از وب مي‌شود.

جدول3-3 بررسي تفاوتهاي کاوشگر وب و دروازة وب

|  |  |  |
| --- | --- | --- |
| **پارامتر ارزيابي** | **دروازة وب** | **کاوشگر وب** |
| نحوة ساخت | توسط انسان(با انتخاب صفحات) | توسط برنامة کامپيوتري(Spider) |
| نحوة دستيابي | از طريق يک سلسله مراتب موضوعي(صفحات بر اساس موضوع مرتب شده‌اند.) | بدون سلسله مراتب موضوعي(صفحات در جواب جستجوي بوسيلة يک الگوريتم کامپيوتري مرتب مي‌شوند) |
| شاخص‌بندي | فقط عنوان، توضيحات و چکيدة صفحات شاخص‌بندي شده و قابل جستجو هستند. | تمام متن صفحات (تمام کلمات صفحه) شاخص‌بندي شده و قابل جستجو هستند. |
| اندازه | اغلب کوچک و به ندرت بزرگ هستند. | اغلب عظيم هستند و اطلاعات زيادي را هم در پاسخ به يک جستجو برمي‌گردانند. |
| گزينش اطلاعات | اغلب صفحات به صورت انتخابي، بعد از ارزيابي اوليه‌اي در اختيار قرار مي‌گيرند. | اطلاعات از هر نوعي مي‌توانند باشند وبدون ارزيابي اوليه در اختيار قرار مي‌گيرند. |
| امکانات جستجو | امکان جستجوهاي ساده با تعداد کمي عملگر | امکان جستجوهاي پيچيده با تعداد زيادي عملگر |
| کاربرد | يافتن اطلاعات عمومي و کلي | يافتن اطلاعات جزيي و خاص در مورد يک موضوع |
| محدوديتها | - پوشش کم وب  - شاخص‌بندي محدود | - دقت كم نتايج برگردانده شده  - پوشش كم وب  - به روز نبودن نتايج |

در اين مقاله نحوة انتخاب عمليات مربوط به هر يك از كاربردهاي وب كاوي در جدول SPAM ارائه شد و روشها و ابزارهاي موجود براي اين كاربردها نيز در جدول T&A ارائه شد. با استفاده از جداول ارائه شده و همچنين استفاده از مقياسهاي ارزيابي روشهاي وب کاوي مي توان تکنيک مناسب را براي کاربرد مورد نظر انتخاب کرد.

موتور جستجو ابزاري است که امکان جستجوي اطلاعات در وب را فراهم مي‌کند. موتورهاي جستجو در واقع راه حل ساده‌اي براي يافتن اطلاعات در روي اينترنت ارائه مي‌دهند. موتورهاي جستجوي بر مبناي انديس زيادي وجود دارند كه در وب جستجو مي كنند، صفحات وب را انديس گذاري مي كنند و انديسهاي عظيمي بر مبناي كلمات كليدي مي سازند و ذخيره مي كنند. اين انديسها به يافتن صفحات وب كه شامل آن كلمات كليدي است كمك مي كند. اما موتورهاي جستجوي بر مبناي كلمات كليدي كم و بيش داراي كمبودهايي هستند:

1- يک عنوان ممکن است صدها يا هزاران سند را شامل شود. اين مسئله باعث مي شود که موتور جستجو تعداد زيادي سند برگرداند که بعضي از آنها اصلاً نامربوط باشند ويا از کيفيت کمي برخوردار باشند.

2- تعداد زيادي از اسنادي که کاملاً به عنوان مورد نظر وابسته هستند هم ممکن است داراي کلمات کليدي که آن موضوع را تعريف مي‌کند، نباشند. اين مسئله منجر به مشکل تکثر وتعدد معاني مي‌شود.

اين مشکلات نشان مي‌دهند که موتورهاي جستجوي فعلي براي کشف منابع در وب کافي نيستند. اما استفاده از ابزارها و روشهاي مختلف وب‌كاوي در بهبود نتايج اين موتورهاي جستجو تاثير بسزايي خواهد داشت و امروزه محققان مختلفي مطالعاتي را در راستاي بهبود نحوه عملكرد اين موتورها با استفاده از روشهاي ذكر شده دنبال مي كنند.

[1]

**فصل چهارم**

**نرم افزار های داده کاوی**

**4-1** **نرم افزار داده کاوی SPSS Clementine**

یکی از مطرح ترین نرم افزارها در زمینه داده کاوی است. این نرم افزار از سری نرم افزارهای معروف SPSS بوده و مانند نرم افزارهای آماری قبلی از امکانات بسیار زیادی در زمینه تحلیل داده ها برخوردار است.   
آخرین نسخه این نرم افزار 12 است که پس از انتشار این نسخه،نسخه بعدی با نام PASW Modeler منتشر شد. از مزایای این نرم افزار می توان به موارد زیر اشاره نمود:

- داشتن روش های بسیار متنوع برای تحلیل داده ها

- سرعت بسیار بالا در انجام محاسبات و استفاده از اطلاعات پایگاه داده ها

- داشتن محیط گرافیکی به منظور راحتی بیشتر کاربر برای انجام کارهای تحلیلی

در نسخه جدید امکان پاک سازی و آماده سازی داده ها به صورت کاملاً اتوماتیک انجام می شود. این نرم افزار تمامی نرم افزارهای پایگاه داده معروف مانند Microsoft Office و SQL و … را پشتیبانی می کند.

ماژول های موجود در این نرم افزار عبارتند از:

- PASW Association

- PASW Classification

- PASW Segmentation

- PASW Modeler Solution Publisher

این نرم افزار هم بر روی کامپیوتر شخصی و هم بر روی سرور قابل نصب است و از Windows های 32 و 64 بیتی نیز پشتیبانی می کند

**4-2 داده كاوي و OLAP:**

يكي از سوالهاي رايج در ميان متخصصان پردازش داده در مورد تفاوت ميان داده كاوي وOLAP(پردازش آناليزي on-line) .

Olap قسمتي از قالب ابزارهاي تصميم گيري است. پرس وجو هاي سنتي و ابزارهاي گزارش گيري كه چه چيزي در داخل يك پايگاه داده است.olap از اين فراتر ميرود و براي جواب دادن به علت درستي برخي موارد استفاده دارد.

نكته مهم آنكه داده كاوي كاربرد اين تكنيكها را براي مسائل تجاري مشابه بالا به طريقي كه اين تكنيكها را براي كاربر خبره دانش و آمارگير متخصص قابل دسترس سازد استفاده مي شود.

**4-2-1 داده كاوي موفق:**

دو نكته براي موفق بودن يك داده كاوي وجود دارد. اول اينكه يك فرموله سازي دقيق از مساله اي است كه شما بايد حل كنيد. دومين نكته استفاده از داده صحيح است. پس از انتخاب داده اي كه در دسترس شماست يا شايد خريد داده خارجي شما ممكن است نيازمند شويد آن را به روشهايي انتقال داده يا دسته بندي كنيد.

4-2-1-1 تحليل ارتباطات:

تحليل ارتباط يک رهيافت توصيفي براي اکتشاف داده است که مي تواند به مشخص سازي ارتباطات ميان مقادير در پايگاه داده کمک نمايد.دو رهيافت عام براي رسيدن به تحليل ارتباطي اکتشاف ارتباطي و اکتشاف توالي مي باشد.اکتشاف ارتباطات قوانيني را در مورد مواردي را که بايد با هم در يک رويداد ظاهرشوند مانند تراکنش خريد را مي ِابد.تحليل سبد عرضه يک نمونه شناخته شده از کشف ارتباط مي باشد.کشف توالي بسيبار شبيه کشف ارتباط است با توجه به اين نکته که در اينجا توالي يک ارتباط است که در طول يک بازه زماني صورت مي گيرد.

ارتباطات به صورت A=>B نوشته مي شود که به A مقدم يا طرف سمت چپ و به B تالي يا طرف سمت راست مي گويند.براي مثال در قانون ارتباطي "اگر مردم يک چکش بخرند آنگاه مي توانند ميخ بخرند" جمله مقدم "خريد چکش" و جمله تالي "خريد ميخ" مي باشد.

براحتي ميتوان نسبت تراکنشهايي را که شامل مورد يا ليستي ازموارد خاص مي باشد با شمردن آنها تعيين کرد (که در اطنجا موارد ميخ ها و چکش هامي باشد) را تعيين کرد.تعداد موجود از يک نوع ارتباط خاص که در يک پايگاه داده به نظر مي رسد را موجودي يا شيوع آن مورد مي گويند.اگر براي مثال گفته شود که از هر 1000تراکنش 15 تاي آن شامل "ميخ و چکش" مي باشد موجودي اين ارتباط 1.5%خواهد بود.يک موجودي کم(مثلا يک در ميليون) مي تواند بيانگر اين باشد که ان ارتباط خاص در پايگاه داده چندان مهم نيست.

براي کشف قوانين معنا دار ما بايد به فراواني متناسب دفعات اتفاق موارد و ترکيباتشان نيز بنگريم.باداشتن تعداد دفعات اتفاق مورد A مورد B چند بار اتفاق مي افتد؟به عبارت ديگر سوال اين است که ببينيم "هنگامي که مردم يک چکش مي خرند چه تعداد از اين افراد ميخ هم مي خرند؟ عبارت ديگر براي اين پيش بيني شرطي اطمينان نام دارد.

فرض کنيد پايگاه داده فرضي مان رابه صورت زير و با جزئيات بيشتر براي بيان اين مفاهيم در نظر بگيريم:

تمام تراکنشهاي سخت افزار :1000

تعداد تراکنشهايي که شامل "چکش " مي باشد:50

تعداد تراکنشهايي که شامل "ميخ" مي باشد:80

تعداد تراکنشهايي که شامل "تخته " مي باشد:20

تعداد تراکنشهايي که شامل " ميخ و چکش"مي باشد:15

تعداد تراکنشهايي که شامل " ميخ و تخته " مي باشد:10

تعداد تراکنشهايي که شامل " چکش و تخته" مي باشد: 10

تعداد تراکنشهايي که شامل " چکش و تخته و ميخ " مي باشد:5

حال قادر به محاسبه ايم:

موجودي "ميخ و چکش"=1.5%

موجودي " ميخ و چکش وتخته"=0.5%

درصد اطمينان "چکش=>ميخ" = 30%

درصد اطمينان " ميخ=> چکش" = 19%

درصد اطمينان " چکش و ميخ=>تخته" = 33%

درصد اطمينان " تخته=> چکش و ميخ " =25%

بنابراين ما مي بينيم که احتمال اينکه يک خرنده چکش ميخ هم بخرد(30%) بيشتر از احتمال آن است که فردي که ميخ مي خرد چکش هم بخرد(19%).ارتباط چکش و ميخ به اندازه اي بزرگ است که يک قانون با معني باشد.

Lift(نسبتا پيشرفت) يکي از معيارهاي اندازه گيري قدرت يک ارتباط است.هر چه lift بزرگتر باشد تاثير اتفاقات A بر احتمال اينکه B اتفاق بيفتد بيشتر است.lift بصورت نسبت

(اطمينان A=>B) تقسيم بر فراواني B محاسبه مي شود:

براي مثال ما:

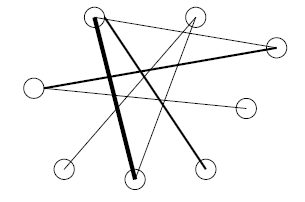
Lift "چکش=>ميخ" :3.75

Lift" چکش و ميخ =>تخته ":16.5

الگوريتمهاي ارتباط اين قوانين را با معادل مرتب سازي داده هنگام شمارش دفعاتي که مي توانند درصد اطمينان و موجودي را محاسبه كنند مي يابد. اثراتي كه هر يك از اين قوانين مي توانند داشته باشند يكي از معيارهاي تفاوت اين الگوريتم هاست. اين معيار مهم است زيرا كه نتايج تركيبي بسيار زيادي از تعداد بي شماري از قوانين بدست مي آيد حتي براي سبد هاي خريد. برخي از الگوريتمها يك پايگاه داده از قوانين, فاكتورهاي ايمن, و فراهم آوردن امكان جستجو(براي مثال تمام ارتباطاتي كه در آن كلمه بستني در قوانين به عنوان نتيجه آمده و فاكتوري برابر 80%را دارند نشان بده)را ايجاد مي نمايند.

اغلب تصميم گيري در مورد كار با قوانيني كه شما كشف كرده ايد دشوار است.به عنوان مثال در يك نقشه خريد براي مشتريان در يك فروشگاه قراردادن تمام اجناس مرتبط منطقي به صورت فيزيكي در كنار يكديگر ممكن است ارزش كامل سبد خريد را كاهش دهد – مشتريان ممكن است در مجموع ارزش كمتري خريد كنند چون آنها بر خلاف نقشه خريد مورد نظر شما در حين راه رفتن در مغازه اجناس مورد دلخواه خود را خريد مي كنند. در چنين حالتي تقريب و تحليل ارتباطات معمولا براي بدست آوردن هر گونه سودي از قوانين مرتبط با هم مورد نياز خواهد بود.

روشهاي گرافيكي مي توانند در نمايش ساختار ارتباطات نقش داشته باشند. در شكل زير هر يك از دواير يك مقدار يا يك رويداد را نمايش مي دهد. خطوط ارتباطي ميان اين دايره ها يك ارتباط را نشان مي دهند. خطوط كلفت تر ارتباطات قوي تر و فراوان تري را نمايش مي دهند.



شکل 4-1 گراف ساختار ارتباطات

4-2-1-2 سلسله مراتبي از انتخاب ها

هدف داده كاوي توليد دانش جديدي است كه كاربر بتواند بر اساس آن كار خود را جلو برد. اين كار بوسيله ساختن مدلي از جهان واقعي بر پايه داده هايي كه از منابع گوناگون بدست مي آيد صورت گيرد كه اين منابع مي تواند شامل تراكنشهاي هماهنگ, تاريخ مربوط به هر مشتري, اطلاعات نمايش گرافيكي, داده كنترل فرآيند و پايگاه داده هاي مرتبط خارجي مانند اطلاعات اعتبار اداري و ... باشد. نتيجه مدل سازي يك سري توضيحات در مورد الگوها و ارتباطات داده اي كه مي تواند به صورت مطمئني جهت پيش بيني آينده مورد استفاده قرار گيرد.

براي جلوگيري از سرگرداني در مراحل مختلف داده كاوي ايجاد تصويري از سلسله مراتبي از انتخابات و تصميم ها كه نياز مند آن هستيد در ذهن قبل از شروع كار به شما كمك خواهد كرد:

### **- هدف كار**

**- نوع پيش بيني**

**- نوع مدل انتخابي**

**- الگوريتم**

**- محصول**

اولين گام مشخص نمودن هدف كار مي باشد :

هدف نهايي از جستجوي اين داده چيست؟ براي مثال جهت يافتن الگوهاي مفيدي در داده خود براي اين كه به شما كمك كند مشتريان خود را حفظ كنيد شما بايد يك مدل براي پيش بيني سودبخشي به مشتري و مدل ديگري براي شناسايي مشترياني كه آنجا را ترك كرده اند طراحي كنيد. دانش شما از احتياجات و اهداف سازمانتان شما را به سمت فرموله كردن اهداف مدلهايتان راهنمايي خواهد كرد.

گام بعدي تصميم در مورد انتخاب نوعي پيش بيني كه از همه مناسب تر است مي باشد:

(1) طبقه بندي: تعيين اين كه اين مورد خاص در كدام كلاس يا دسته قرار مي گيرد.

(2) حدس زدن اينكه يك متغير چه مقدار عددي خواهد داشت(اگر متغيري باشد كه با زمان تغيير كند اين كار حدس سريهاي زماني ناميده مي شود).در مثال بالا شما مي توانيد از اين حدس براي پيش بيني مقدار سوددهي و طبقه بندي براي پيش بيني اينكه كدام مشتريان ممكن است خريد شما را ترك كنند استفاده كنيد.

حالا نوبت به نوع مدل مي رسد:

كه عبارت است از يك شبكه عصبي براي انجام حدس فوق الذكر و يك درخت تصميم براي طبقه بندي. مدلهاي آماري سنتي نيز براي انتخاب از مدلهاي معمولي خطي , تحليل تفكيكي و حدس منطقي وجود دارد. مهمترين نوع اين مدلها براي داده كاوي در بخش بعد (الگوريتمها و مدلهاي داده كاوي)توضيح داده مي شود.

الگوريتمهاي زيادي براي ساخت مدلهايتان در دسترس هستند. شما مي توانيد با استفاده از توابع شعاعي يا انتشاري شبكه عصبي بسازيد. براي درخت تصميم شما مي توانيد از ميان طرق cart, c5.0, Quest,CHAID يكي را انتخاب كنيد. برخي از اين الگوريتم ها در مدلها و الگوريتمهاي داده كاوي توضيح داده شده است.

هنگام انتخاب يك محصول داده كاوي بايد توجه داشت كه اين محصولات پياده سازيهاي مختلفي از يك الگوريتم خاص دارند حتي اگر اين الگوريتم براي همه آنها نام يكساني داشته باشد. اين تفاوتها در پياده سازي مي تواند بر روي مشخصه هاي قابل استفاده مانند استفاده از حافظه و ذخيره داده و همچنين بر روي مشخصه هاي كارايي مانند سرعت و دقت تاثير بگذارند.

بسياري از اهداف تجاري به بهترين شكل به وسيله ساخت انواع مختلفي از مدلها با استفاده از الگوريتمهاي مختلف به دست مي آيند. شما ممكن است تا زماني كه راه هاي مختلفي را امتحان نكنيد قادر نباشيد تعيين كنيد كدام نوع مدل بهترين است.

**4-2-2 طبقه بندي**

مسائل طبقه بندي به شناسايي خصوصياتي منجر مي شوند كه مشخص مي نمايند هر مورد به كدام گروه تعلق دارد.اين الگو هم مي تواند براي فهم داده موجود و هم براي پيش بيني اينکه هر نمونه جديد چگونه کار مي کند استفاده شود.براي مثال شما ممکن است بخواهيدپيش بيني کنيد که آيااشخاص براي پاسخگويي به درخواست يک ميل مستقيم که ممکن است به يک دستگاه تلفن با مسافت زياد آسيب برساند مي توانند گروه بندي شوند يا براي يک عمل جراحي بايد گروه بندي شوند.

داده کاوي مدلهاي طبقه بندي را بوسيله امتحان کردن داده طبقه بندي شده(موارد) و نهايتا يافتن يک الگوي پيش گو ايجاد مي کند.اين موارد موجود مي تواند از يك پايگاه داده تاريخي ناشي شود مانند اطلاعات افرادي كه تحت معالجه دارويي خاصي هستند و يا به سمت يك خدمت با مسافت دور جذب شده اند.يا اينكه از تجربه هايي كه طي آن يك نمونه از تمام پايگاه داده در جهان واقعي تست شده باشد و نتايج آن براي ايجاد يك گروه بند استفاده شده باشند منتج شود. براي مثال يك نمونه از ليستي از پيامها به عنوان پيشنهاد فرستاده خواهد شد و نتايج پيام رساني براي ساخت يك مدل طبقه بندي جهت بكار گرفته شدن در تمام پايگاه داده استفاده خواهد شد.[1]

#### 4-2-3 حدس بازگشتي

حدس بازگشتي از داده هاي موجود براي پيش بيني اين كه مقادير داده هاي ديگر چه خواهد بود استفاده مي كند. در ساده ترين حالت حدس مذكور از تكنيكهاي آماري مانند حدس خطي استفاده مي كند. متاسفانه بسياري از مسائل جهان واقع تصويري خطي از مقادير قبلي نيستند. براي نمونه مقادير فروش, ارزش فروش, ارزش سهام و نرخ ورشكستگي محصول براي پيش بيني سخت مي باشد زيرا آنها ممكن است بر فعل و انفعالات پيچيده حاصل از چندين متغير پيش بيني كننده متكي باشند. بنابراين تكنيكهاي پيچيده تري ممكن است براي پيش بيني متغيرهاي آينده ضروري باشند. انواع مدل يكسان اغلب مي توانند هم براي حدس بازگشتي وهم براي طبقه بندي استفاده شوند. براي مثال الگوريتم درخت تصميم CART (درختهاي حدس وطبقه بندي )هم براي ساخت درختهاي حدس و هم براي ساخت درختهاي طبقه بندي به كار مي رود. شبكه هاي عصبي هم مي توانند هر دو نوع مدل نام برده شده را ايجاد نمايند.

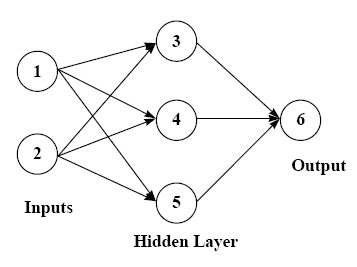
#### 4-2-4 سري هاي زماني

سري هاي زماني پيش بيني كننده مقاديري را كه هنوز مقدارشان مشخص نيست بر اساس يك سري از پيشگوهاي متغير با زمان پيش بيني مي كنند. مانند حدس بازگشتي اين روش هم از نتايج معلوم قبلي براي اعمال پيشگويي هاي بعدي اش بهره مي برد. مدلها بايد خواص منحصر بفرد زمان علي الخصوص سلسله مراتب دوره هاي زماني مانند دوره هاي فصلي تاثيرات تقويمي مانند تعطيلات محاسبات تاريخي و ملاحظات خاص مانند تطبيق گذشته با حال را ذخيره نمايند.

#### 4-2-5 شبكه هاي عصبي

شبكه هاي عصبي به طور خاصي مورد استفاده اند چرا كه آنها ابزاري موثر براي مدلسازي مسائل بزرگ و پيچيده كه ممكن است در آنها صدها متغير پيش بيني كننده كه فعل و انفعالات زيادي دارند وجود داشته باشد.(شبكه هاي عصبي زيستي بطور غير قابل مقايسه اي پيچيده تر هستند.)شبكه هاي عصبي مي توانند در مسائل طبقه بندي يا حدسهاي بازگشتي(كه در آنها متغير خروجي پيوسته است) استفاده شوند.

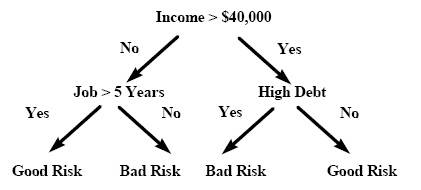
يك شبكه عصبي با يك لايه داخلي شروع مي شود كه در آن هر گره به يك متغير پيشگو منسوب مي گردد. اين گره هاي ورودي به يك تعداد از گره ها در لايه پنهان متصل مي شوند.گره ها در لايه پنهان مي توانند به گره هايي در يك لايه پنهان ديگر يا به يك لايه خروجي متصل شود. لايه خروجي خود شامل يك يا بيشتر متغيرهاي جواب مي باشد.



شکل 4-2 يك شبكه عصبي با يك لايه پنهان

#### 4-2-6 درخت هاي انتخاب

درخت هاي انتخاب راهي براي نمايش يك سري از قوانين كه به يك كلاس يا مقدار منجر مي شود مي باشند. براي مثال شما ممكن است بخواهيد درخواستهاي وام را برحسب ريسك اعتبار خوب يا بد طبقه بندي كنيد. شكل بعد يك مدل ساده از يك درخت انتخاب به همراه توضيح در مورد تمام بسته هاي پايه آن يعني گره انتخاب, شاخه ها و برگهاي آن كه اين مساله را حل مي كند نشان مي دهد.



شکل 4-3 درخت انتخاب

اولين بسته گره بالايي تصميم يا ريشه مي باشد كه يك بررسي جهت برقراري شرط خاصي مي نمايد. گره ريشه در اين مثال “Income>$40,000” مي باشد. نتايج اين بررسي منجر مي شود كه درخت به دوشاخه تقسيم گرددكه هر يك نشان دهنده جوابهاي ممكن است.در اين مورد بررسي شرط مذكور مي تواند داراي جواب خير يا بله باشد در نتيجه دو شاخه داريم.

براساس نوع الگوريتم هر گره مي تواند دو يا تعداد بيشتري شاخه داشته باشد. براي مثال CART درختهايي با تنها دوشاخه در هر گره توليد مي كند.چنين درختي يك درخت دودويي مي باشد.

مدلهاي مختلف درخت تصميم بطور عمومي در داده كاوي براي كاوش داده و براي استنتاج درخت و قوانين آن كه براي پيش بيني مورد استفاده قرار مي گيرد استفاده مي شوند. يك تعداد از الگوريتمهاي مختلف مي توانند براي ساخت درختهاي تصميم شامل CHAID, CART,Quest و C5.0 بكار روند.

اندازه درخت مي تواند از طريق قوانين متوقف شونده كه رشد درخت را محدود مي كنند كنترل شود. [2]

#### 4-2-7 استنتاج قانون

استنتاج قانون روشي براي بدست آوردن يك سري از قوانين براي طبقه بندي موارد مي باشد. اگرچه درختهاي تصميم مي توانند يك سري قوانين توليد كنند روشهاي استنتاج قانون يك مجموعه از قوانين وابسته كه ضرورتا درختي تشكيل نمي دهند را توليد مي نمايد. چون استنتاج كننده قوانين لزوما انشعابي در هر سطح قرار نمي دهد و مي تواند گام بعدي را تشخيص دهد گاهي اوقات مي تواند الگوهاي مختلف و بهتري را براي طبقه بندي بيابد. برخلاف درختان قوانين توليدي ممكن است تمام حالتهاي ممكن را پوشش ندهند.

##### 4-3 ساختن يك پايگاه داده داده كاوي

اين گام به همراه دو گام بعدي هسته آماده سازي داده را تشكيل مي دهند. در مجموع گامهاي گفته شده وقت و كار بيشتري از ساير گامها مي برند. ممكن است شما گامهاي تكراري در آماده سازي داده و ساختن مدل داشته باشيد چرا كه در هر مرحله ممكن است به نكته اي برسيد كه شما را بر آن دارد داده خود را بهبود بخشيد. اين گامهاي آماده سازي داده مي تواند 50% تا 90% وقت و كار از تمام فرآيند كشف دانش را به خود اختصاص دهد.

داده اي كه مي خواهد كاوش شود بايد در يك پايگاه داده ذخيره شود. بر اساس مقدار داده, پيچيدگي داده و استفاده هايي كه قرار است از آن شود يك فايل معمولي و يا يك SpreadSheet براي اين كار كافي است.

به احتمال زياد شما مي خواهيد داده موجود در انباره داده را تغيير دهيد. به علاوه شما ممكن است بخواهيد فيلدهاي جديدي كه از فيلدهاي موجود محاسبه شده است را به انبار داده خود بيافزاييد.اين يكي از دلايل استفاده از يك پايگاه داده جداگانه است.

دليل ديگر براي اين كار آن است كه انبار داده هاي يكي شده ممكن است به آساني انواع جستجوهايي را كه شما براي فهم داده به آنها نياز داريد انجام ندهد. مانند پرس و جوهايي كه داده را خلاصه مي كند, گزارشات چند بعدي و بسياري از انواع ديگر از گرافها يا مصورات.

و دليل آخر اينكه شما ممكن است بخواهيد اين داده را در يك سيستم مديريت پايگاه داده به همراه يك طراحي فيزيكي متفاوت از انبار داده خود ذخيره كنيد. مردم به طور روز افزوني در حال انتخاب پايگاه داده هاي خاص منظوره اي هستند كه اين نيازهاي داده كاوي را به نحو مناسبي حمايت كند. به هرحال اگر داده موجود در انبار داده شما اجازه مي دهد كه مراكز منطقي داده اي ايجادكنيد و اگر شما مي توانيد تقاضاي داده كاوي را ارضا نماييد پايگاه داده شما به خوبي وظيفه خود را انجام مي دهد.

مراحل لازم براي ساخت يك پايگاه داده داده كاوي به شكل زير مي باشد:

1. جمع آوري داده ها
2. توضيح داده ها
3. انتخاب داده ها
4. تعيين كيفيت داده ها و پاك كردن آن
5. تثبيت و يكپارچگي
6. ساختن فوق داده (داده هايي كه خود بيانگر توضيحي در مورد داده هاي موجود مي باشند.)
7. باركردن پايگاه داده مربوط به داده كاوي
8. نگهداري پايگاه داده مربوط به داده كاوي

اين كارها ممكن است لزوما به همين ترتيب گفته شده انجام نگردند.

##### 4-3-1 جستجوي داده

به بخش توضيح داده براي داده كاوي كه توضيح مختصري راجع به اشكال, تجزيه و تحليل ارتباط و ديگر وسايل جستجوي داده مي باشد نگاهي بياندازيد.

هدف شناسايي مهمترين فيلدها در پيش بيني نتيجه و تعيين اينكه كدام يك از داده هاي بدست آمده مفيد مي باشد است.

در يك مجموعه داده اي با صدها يا حتي هزاران ستون جستجوي داده مي تواند كار و زمان بر باشد. يك واسط مناسب و جواب كامپيوتر سريع در اين فاز مهم و حياتي مي باشند زيرا هنگامي كه شما براي دريافت پاسخ برخي گراف ها مجبور باشيد 20 دقيقه صبر كنيد ماهيت جستجوي شما به كلي تغيير خواهد كرد.

##### 4-3-2 آماده سازي داده براي مدل سازي

اين آخرين گام آماده سازي داده قبل از ساخت مدلهاست. چهار قسمت مهم در اين مرحله وجود دارد:

1. انتخاب متغيرها
2. انتخاب سطرها
3. ساختن متغيرهاي جديد
4. تغيير شكل متغيرها

#### 4-4 ساختن مدل داده كاوي

مهمترين مساله براي يادآوري در مورد ساخت مدل آن است كه اين كار يك فرآيند تكراري است. شما براي جستجو به مدلهاي جايگزين جهت يافتن سودمندترين آنها جهت حل مسائلتان نياز داريد. آنچه كه شما در جستجوي يك مدل مناسب ياد مي گيريد مي تواند شما را به بازگشتن به عقب و انجام برخي تغييرات در داده مورد استفاده خود و حتي بهبود بيان ساله راهنمايي كند.

هنگامي كه شما در مورد نوع پيش بيني كه مي خواهيد انجام دهيد تصميم گرفتيد بايد يك نوع مدل براي ساخت تصميم خود انتخاب كنيد.

آماده سازي و آزمايش مدل داده كاوي احتياج به اين دارد كه داده به حداقل دو گروه شكسته شود: يكي براي آماده كردن مدل و ديگري جهت تست مدل مربوطه. اگر شما از آماده سازي و تست متفاوتي استفاده ننمائيد دقت مدل خواهد بود.

پايه اي ترين روش تست داده تاييد اعتبار ساده مي باشد. براي انجام اين كار چون درصدي از پايگاه داده را به عنوان يك تست پايگاه داده كنار بگذاريد و به هر صورت از آن در برآورد و ساخت مدل استفاده ننمائيد. اين درصد معمولا بين 5 تا 33 مي باشد.

#### 4-4-1 ارزيابي و تفسير

بعد از ساخت يك مدل شما بايد نتايج آن را ارزيابي نموده و همچنين اهميت آن را نيز توضيح دهيد.

#### 4-4-2 ايجاد معماري مدل و نتايج

هنگامي كه يك مدل ساخته و تاييد اعتبار مي شود مي تواند در دو راه اصلي مورد استفاده قرار گيرد. راه اول براي تحليل گر است كه اعمالي را بر اساس ديد ساده از مدل و نتايج آن معرفي مي كند. راه دوم بكاربردن مدلها در مجموعه داده اي مختلف است. اين مدل مي تواند براي مشخص نمودن ركوردها بر اساس گروه بنديشان و يا مقدار دهي يك امتياز مثلا احتمال انجام يك عمل استفاده گردد.

هنگام به دست آوردن يك كاربرد پيچيده داده كاوي اغلب اگر چه بخش بحراني اما كوچك پروژه نهايي به حساب مي آيد. براي مثال دانشي كه از داده كاوي كشف مي شود مي تواند با دانش متخصصان داده و تراكنشهاي ورودي تركيب شود. در يك سيستم تشخيص فرآيند الگوهاي موجود فرآيند مي توانند با الگوهاي كشف شده تلفيق شوند. هنگامي كه موارد مفروض اين فرآيند براي ارزيابي به بررسي كنندگان فرستاده مي شوند بررسي كنندگان ممكن است نياز داشته باشند كه به ركوردهايي در پايگاه داده كه مربوط به قسمتهاي ادعا شده توسط يك سازنده است دسترسي پيدا كنند.

به طور كلي مراحلي كه توضيح داده شد براي انجام هر فرآيند داده كاوي لازم به نظر مي رسد.[3]

**4-5 نرم افزارWeka**

4-5-1 درباره وکا

تا به امروز نرم افزارهاي تجاري و آموزشي فراواني براي داده كاوي در حوزه هاي مختلف داده ها به دنياي علمو فناوري عرضه شده­اند. هريك از آنها با توجه به نوع اصلي داده هايي كه مورد كاوش قرار مي­دهند، روي الگوريتمهاي خاصي متمركز شده­اند. مقايسه دقيق و علمي اينابزارها بايد از جنبه هاي متفاوت و متعددي مانند تنوع انواع و فرمت داده هاي ورودي،حجم ممكن براي پردازش داده ها، الگوريتمها پياده سازي شده، روشهاي ارزيابي نتايج،روشهاي مصور سازي  ، روشهاي پيش پردازش  داده ها، واسطهاي كاربر پسند  ، پلت فرم  هاي سازگار براي اجرا،‌ قيمت و در دسترس بودن نرم افزار صورت گيرد. از آن ميان،‌نرم افزار Weka با داشتن امكانات بسيار گسترده،‌ امكان مقايسه خروجي روشهاي مختلف با هم، راهنماي خوب، واسط گرافيگي كارآ، سازگاري با ساير برنامه­هاي ويندوزي، و ازهمه مهمتر وجود كتابي بسيار جامع و مرتبط با آن [ Data Mining, witten et Al. 2005 ]، معرفي مي­شود.

ميزكار  Weka ، مجموعه­اي از الگوريتم­‏هاي يادگيري ماشيني و ابزارهاي پيش پردازش داده­‏ها مي‏­باشد. اين نرم­افزار به گونه­اي طراحي شده است كه مي‏­توان به سرعت، روش­‏هاي موجود را به صورت انعطاف­پذيري روي مجموعه­‏هاي جديد داده، آزمايش نمود. اين نرم­افزار، پشتيباني‏‏هاي ارزشمندي رابراي كل فرآيند داده كاوي ­‏هاي تجربي فراهم مي‏­كند. اين پشتيباني‏ها، آماده سازي داده­‏هاي ورودي، ارزيابي آماري چارچوب­‏هاي يادگيري و نمايش گرافيكي داده­‏هاي ورودي و نتايج يادگيري را در بر مي­گيرند. همچنين، هماهنگ با دامنه وسيع الگوريتم­‏هاي يادگيري، اين نرم­افزار شامل ابزارهاي متنوع پيش پردازش داده­هاست. اين جعبه ابزار  متنوع و جامع، از طريق يك واسط متداول در دسترس است، به نحوي كهكاربر مي‏­تواند روش­‏هاي متفاوت را در آن با يكديگر مقايسه كند و روش­‏هايي را كه براي مسايل مدنظر مناسب­تر هستند، تشخيص دهد.

نرم­افزار Weka در دانشگاه Waikato واقع در نيوزلند توسعه يافته است و اسم آن از عبارت "Waikato Environment for knowledge Analysis" استخراج گشته است. همچنين Weka ، نام پرنده­اي با طبيعت جستجوگر است كه پرواز نمي‏­كند و در نيوزلند، يافت مي‏­شود. اين سيستم به زبان جاوانوشته شده و بر اساس ليسانس عمومي و فراگير  GNU انتشار يافته است. Weka تقريباًروي هر پلت فرمي اجرا مي‏­شود و نيز تحت سيستم عامل­‏هاي لينوكس، ويندوز، ومكينتاش، و حتي روي يك منشي ديجيتالي شخصي  ، آزمايش شده است.

اين نرم­افزار، يك واسط همگن براي بسياري از الگوريتم­‏هاي يادگيري متفاوت، فراهم كرده است كه از طريق آن روش­‏هاي پيش پردازش، پس از پردازش  و ارزيابي نتايج طرح هاي يادگيري روي همه مجموعه هاي داده­‏ موجود،قابل اعمال است.

نرم افزار Weka ، پياده سازي الگوريتم­‏هاي مختلف يادگيري را فراهم مي‏­كند و به آساني مي‏­توان آنها را به مجموعه هاي داده خود اعمال كرد.

همچنين، اين نرم­افزار شامل مجموعه متنوعي از ابزارهاي تبديل مجموعه­‏ هاي داده­‏ها، همانندالگوريتم­‏هاي گسسته سازي مي‏­باشد. در اين محيط مي‏­توان يك مجموعه داده را پيشپردازش كرد، آن را به يك طرح يادگيري وارد نمود، و دسته­بندي حاصله و كارآيي­اش رامورد تحليل قرار داد. (همه اين كارها، بدون نياز به نوشتن هيچ قطعه برنامه­اي ميسراست).

اين محيط، شامل روش­‏هايي براي همه مسايل استاندارد داده كاوي مانند رگرسيون، رده‏بندي، خوشه­بندي، كاوش قواعد انجمني و انتخاب ويژگي مي‏­باشد. با در نظر گرفتن اينكه، داده­‏ها بخش مكمل كار هستند، بسياري از ابزارهاي پيشپردازش داده­‏ها و مصورسازي آنها فراهم گشته است. همه الگوريتم­‏ها، ورودي­‏هاي خودرا به صورت يك جدول رابطه­اي  به فرمت ARFF دريافت مي‏­كنند. اين فرمت داده­‏ها،مي‏­تواند از يك فايل خوانده شده يا به وسيله يك درخواست از پايگاه داده­اي توليدگردد.

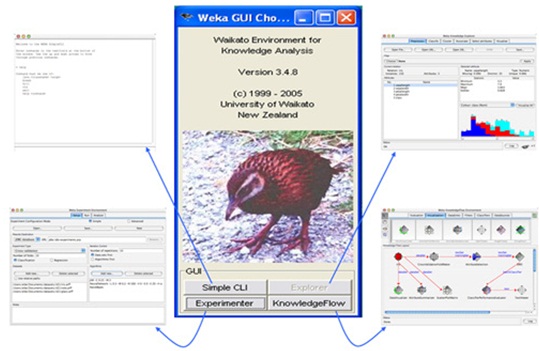
يكي از راه­‏هاي به كارگيري Weka ، اعمال يك روش يادگيري به يكمجموعه داده و تحليل خروجي آن براي شناخت چيزهاي بيشتري راجع به آن اطلاعاتمي‏­باشد. راه ديگر استفاده از مدل يادگيري شده براي توليد پيش‏بيني­‏هايي در موردنمونه­‏هاي جديد است. سومين راه، اعمال يادگيرنده­‏هاي مختلف و مقايسه كارآيي آنهابه منظور انتخاب يكي از آنها براي تخمين مي‏­باشد. روش­‏هاي يادگيري Classifier ناميده مي‏­شوند و در واسط تعاملي  Weka ، مي‏­توان هر يك از آنها را از منو  انتخاب نمود. بسياري از classifier ‏ها پارامترهاي قابل تنظيم دارند كه مي‏­توان ازطريق صفحه ويژگي‏‏ها ياobject editor به آنها دسترسي داشت. يك واحد ارزيابي مشترك،براي اندازه‏گيري كارآيي همه classifier به كار مي‏­رود.

پياده سازي­‏هاي چارچوب­‏هاي يادگيري واقعي، منابع بسيار ارزشمندي هستند كه Weka فراهم مي‏­كند. ابزارهايي كه براي پيش پردازش داده­‏ها استفاده مي‏­شوند. filterناميده مي‏­شوند. همانند classifier ‏ها، مي‏­توان filter ‏ها را از منوي مربوطه انتخاب كرده و آنهارا با نيازمندي­‏هاي خود، سازگار نمود. در ادامه، به روش به كارگيري فيلترها اشارهمي‏­شود.

علاوه بر موارد فوق، Weka شامل پياده سازي الگوريتم­‏هايي براي يادگيري قواعد انجمني، خوشه­بندي داده­‏ها در جايي كه هيچ دسته­اي تعريف نشده است،و انتخاب ويژگي‏هاي مرتبط در داده­‏ها مي‏­شود.

 روش استفاده از Weka

شکل4-1،راههای انتخاب واسط‌های مختلف Weka را نشان میدهد.آسان­ترين راه استفاده از Weka ، از طريق واسطي گرافيكي است كه Explorer خوانده مي‏شود. اين واسط گرافيكي، به وسيله انتخاب منوها و پر كردن فرم­‏هاي مربوطه،دسترسي به همه امكانات را فراهم كرده است. براي مثال، مي‏­توان به سرعت يك مجموعه داده را از يك فايل ARFF خواند و درخت تصميم‏گيري آن را توليد نمود. امادرخت­‏هاي تصميم‏گيري يادگيرنده صرفاً ابتداي كار هستند. الگوريتم­‏هاي بسيار ديگري براي جستجو وجود دارند. واسط Explorer كمك مي‏­كند تا الگوريتم­‏هاي ديگر نيز آزمايش شوند.

****

شکل 4-1 در وضعیت انتخاب واسط

اين واسط با دراختيار گذاشتن گزينه­‏ها به صورت منو، با وادار كردن كاربر به اجراي كارها با ترتيب صحيح، به وسيله خاكستري نمودن گزينه­‏ها تا زمان صحيح به كارگيري آنها، و با دراختيار گذاشتن گزينه­‏هايي به صورت فرم­‏هاي پرشدني، كاربر را هدايت مي‏­كند. راهنماي ابزار مفيدي، حين عبور ماوس از روي گزينه­‏ها، ظاهر شده و اعمال لازم مربوطه را شرح مي‏­دهد. پيش­فرض­‏هاي معقول قرار داده شده، كاربر را قادر مي‏­سازندتا با كمترين تلاشي، به نتيجه برسد. اما كاربر بايد براي درك معني نتايج حاصله،راجع به كارهايي كه انجام مي‏­دهد، بينديشد.   Wekaدو واسط گرافيكي ديگرنيز دارد. واسط knowledge flow به كاربر امكان مي‏­دهد تا چنيش­‏هايي براي پردازش داده­‏هاي در جريان، طراحي كند. يك عيب پايه­اي Explorer . نگهداري هر چيزي درحافظه اصلي آن است. (زماني كه يك مجموعه داده را باز مي‏­كنيم، Explorer ، كل آنرا، در حافظ باز مي‏­كند) نشان مي‏­دهد كه Explorer ، صرفاً براي مسايل بااندازه­‏هاي كوچك تا متوسط، قابل اعمال است. با وجود بر اين Weka شامل تعدادي الگوريتم­‏هاي افزايشي است كه مي‏­تواند براي پردازش مجموعه هاي داده بسيار بزرگ مورد استفاده قرار گيرد. واسط knowledge flow امكان مي‏­دهد تا جعبه  ­‏هاي نمايانگر الگوريتم­‏هاي يادگيري و منابع داده­‏ها را به درون صفحه بكشيم و با اتصال آنها به يكديگر، تركيب و چينش دلخواه خود را بسازيم. اين واسط اجازه مي‏­دهد تاجريان داده­اي از مؤلفه­‏هاي به هم متصل كه بيانگر منابع داده، ابزارهاي پيش پردازش، روش­‏هاي ارزيابي و واحدهاي مصوّر سازي هستند تعريف شود. اگر فيلترها والگوريتم­هاي يادگيري، قابليت يادگيري افزايشي را داشته باشند، داده­‏ها به صورت افزايشي بار شده و پردازش خواهند شد.

سومين واسط Weka ، كه Experimenter خوانده مي‏­شود، كمك مي‏­كند تا به اين سؤال عملي و پايه­اي كاربر حين استفاده ازتكنيك­‏هاي رده‏بندي و رگرسيون، پاسخ دهد: "چه روش­‏ها و پارامترهايي براي مسأله داده شده، بهتر عمل مي‏­كنند؟"

عموماً راهي براي پاسخگويي مقدماتي به اين سؤال وجود ندارد و يكي از دلايل توسعه Weka ، فراهم نمودن محيطي است كه كاربران Weka را قادر به مقايسه تكنيك­‏هاي گوناگون يادگيري بنمايد. اين كار، مي‏­تواند بهصورت تعاملي در Explorer انجام شود. با اين وجود، Experimenter با ساده كردن اجراي رده‏بندي كننده­‏ها و فيلترها با پارامترهاي گوناگون روي تعدادي از مجموعه­‏هاي داده، جمع­آوري آمار كارآيي و انجام آزمايش­‏هاي معنا، پردازش را خودكار مي‏­كند. كاربرهاي پيشرفته، مي‏­توانند از Experimenter براي توزيع بار محاسباتي بين چندين ماشين، استفاده كنند. در اين روش، مي‏­توان آزمايش­‏هاي آماري بزرگي را راه­اندازي نموده و آنها را براي اجرا، رها نمود.

وراي اين واسط­‏هاي تعاملي، عملكردپايه­اي Weka قرار دارد. توابع پايه­اي Weka ، از طريق خط فرمان  ­‏هاي متني قابلدسترسي هستند. زماني كه Weka ، فعال مي‏­شود، امكان انتخاب بين چهار واسط كاربريوجود دارد: Explorer ، knowledge ، Experimenter و واسط خط فرمان.

اكثركاربران، حداقل در ابتداي كار Explorer را به عنوان واسط كاربري انتخاب مي‏­كنند.   
**4-5-2 قابليتهاي Weka**

مستندسازي در لحظه، كه به صورت خودكار از كداصلي توليد مي‏­شود و دقيقاً ساختار آن را بيان مي‏­كند، قابليت مهمي است كه حيناستفاده از Weka وجود دارد.

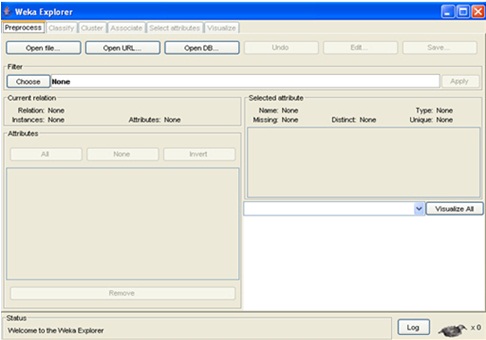
نحوه استفاده از اين مستندات و چگونگي تعيين پايه­‏هاي ساختماني اصلي Weka ، مشخص كردن بخش­‏هايي كه از روش­‏هاي يادگيري باسرپرست استفاده مي‏­كند، ابزاري براي پيش پردازش داده­‏ها بكار مي‏­رود و اينكه چه روش­‏هايي براي ساير برنامه­‏هاي يادگيري وجود دارد، در ادامه تشريح خواهد شد. تنهابه ليست كاملي از الگوريتم­‏هاي موجود اكتفا مي‏­شود زيرا Weka به طور پيوسته تكميل مي‏­شود و به طور خودكار از كد اصلي توليد مي‏­شود. مستندات در لحظه هميشه به هنگام شده مي‏­باشد. اگر ادامه دادن به مراحل بعدي و دسترسي به كتابخانه از برنامه جاواشخصي يا نوشتن و آزمايش كردن برنامه­‏هاي يادگيري شخصي مورد نياز باشد، اين ويژگي بسيار حياتي خواهد بود. در اغلب برنامه­‏هاي كاربردي داده كاوي، جزءيادگيري ماشيني، بخش كوچكي از سيستم نرم­افزاري نسبتاً بزرگي را شامل مي‏­شود. درصورتي كه نوشتن برنامه كاربردي داده كاوي مد نظر باشد، مي‏­توان با برنامه­نويسي اندكي به برنامه­‏هاي Weka از داخل كد شخصي دسترسي داشت. اگر پيدا كردن مهارت درالگوريتم­‏هاي يادگيري ماشيني مدنظر باشد، اجراي الگوريتم­‏هاي شخصي بدون درگيرجزييات دست و پا گير شدن مثل خواندن اطلاعات از يك فايل، اجراي الگوريتم­‏هاي فيلترينگ يا تهيه كد براي ارزيابي نتايج يكي از خواسته­‏ها مي‏­باشد. Weka داراي همه اين مزيت­‏ها است. براي استفاده كامل از اين ويژگي، بايد با ساختارهاي پايه­ايداده­‏ها آشنا شد.

**4-5-3 دريافت Weka**

نرم افزار Weka ، در آدرس<http://www.cs.waikato.ac.nz/me/weka>، در دسترس است. از اينطريق مي‏­توان نصب كننده  متناسب با يك پلت فرم معين، يا يك فايل Java jar را كه درصورت نصب بودن جاوا به راحتي قابل اجرا است، دانلود نمود.

**4-5-3-1 مروري بر Explorer**

واسط گرافيكي اصلي براي كاربران،‌ Explorer است كه امكان دسترسي به همه امكانات Weka را از طريق انتخاب منوها و پر كردن فرمها فراهم مي­آورد. شكل 2،‌ نماي Explorer ‌ را نشان مي­دهد. در اين واسط، شش پانل  مختلف وجود دارد كه ازطريق نوار  بالاي صفحه قابل انتخاب هستند و با وظايف داده كاوي پشتيباني شده توسط Weka ‌ متناظر مي­باشند.

****

شکل4-2.واسط گرافیکی Explorer

به طور خلاصه، كاركرد تمام گزينه­‏ها به شرح ذيل است.

  Preprocessانتخاب مجموعه داده و اصلاح  آن از راه­‏هاي گوناگون

  Classifyآموزش  برنامه­‏هاي يادگيري كه رده‏بندي يا رگرسيون انجام مي‏­دهند و ارزيابي آنها.

  Clusterيادگيري خوشه­‏ها براي مجموعه هاي داده

  Associate:يادگيري قواعد انجمني براي داده­‏ها و ارزيابي آنها

  Select attributes: انتخاب مرتبط­ترين جنبه  ها در مجموعه هاي داده

  Visualize: مشاهده نمودارهاي مختلف دوبعدي داده ­‏ها و تعامل با آنها

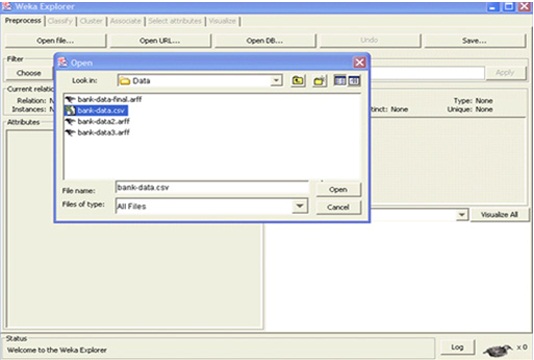
  Weka Exphorerامكان رده بندي دارد، چنانچه به كاربران اجازه مي‏­دهد به صورت تعاملي اقدام به ساخت درخت تصميم‏گيري كنند. Weka نمودار پراكندگي داده­‏ها را نسبت به دو ويژگي انتخاب شده،فراهم مي‏­آورد. وقتي زوج ويژگي­اي كه رده­‏ها را به خوبي جدا مي‏­كند، پيدا شد،امكان ايجاد دو شاخه با كشيدن چند ضلعي اطراف نقاط داده­‏ها بر نمودار پراكندگي وجود دارد.

  هر نوار، دسترسي به دامنه كاملي از امكانات رافراهم مي­كند. در پايين هر پانل، جعبه status و دكمه log قرار دارد. جعبه status پيغام­‏هايي است كه نشان مي‏­دهد چه عملياتي در حال انجام داده شدن است. مثلاً اگر Explores مشغول خواندن يك فايل باشد، جعبه status آن را گزارش مي‏­دهد. كليك راستدر هر جا داخل اين جعبه يك منو كوچك با دو گزينه مي‏­آورد، نمايش ميزان حافظه دردسترس Weka و اجراي Java garbage collector ..

لازم است توجه شود كه garbage collector به طور ثابت به عنوان يك عمل پيش زمينه در هر حال اجرا مي‏­شودكليك دكمه log ، گزارش عملكرد متني كارهايي كه Weka تاكنون در اين بخش انجام دادهاست با برچسب زماني ارايه مي‏­كند.

زمانيكه Weka در حال عمليات است،پرنده كوچكي كه در پايين سمت راست پنجره است، بالا و پايين مي‏­پرد. عدد پشت × نشان مي‏­دهد كه به طور همزمان چند عمليات در حال انجام است. اگر پرنده بايستد در حاليكه حركت نمي‏­كند، او مريض است! اشتباه داده است و بايد Explorerاز نو اجرا شود.   
  1. خواندن و فيلتر كردن فايل­‏ها

  در بالاي پانل Preprocess درشكل 4-3 ، دكمه­‏هايي براي باز كردن فايل، URL ‏ها و پايگاه هاي داده­‏ وجود دارد. درابتدا تنها فايل­‏هاي با پسوند arff . در browser فايل نمايش داده مي‏­شود. براي ديدن ساير فايل­‏ها يايد گزينه  Format در جعبه انتخاب فايل تغيير داده شود.

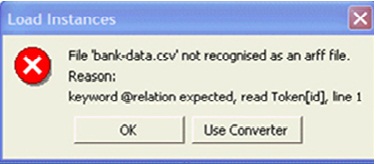
****

شکل 4-3 باز کردن فایل

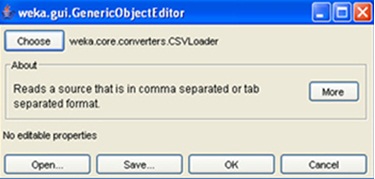
. تبديل فايل­‏ها به فرمت ARFF

نرم افزار Weka داراي سه مبدل فرمت فايل  مي‏­باشد، براي فايل­‏هاي صفحه گسترده با پسوند CSV ، با فرمت فايل C4.5 با پسوند names . و data و براي نمونه­‏هاي سري با پسوند bsi .

اگر Weka قادر به خواندن داده­‏ها نباشد، سعي مي‏­كند آن را به صورت ARFF تفسير كند. اگر نتواند پیغامنشان داده شده در شکل 4 -4 (الف) ظاهر میشود. باانتخاب گزینه Use Convertor، پیغام شکل 4-4 (ب) ظاهر میشود.

****

شکل4-4 (الف) پیغام خطا

****

شکل4-4(ب) ویرایشگر

این، یک ویرایشگر عمومی اشیاء است که در Weka برای انتخاب و تنظیم اشیا بکار میرود. به عنوان مثال وقتی پارامتری برای Classifier تنظیم میشود، جعبه ای با نوع مشابه بکار برده میشود. CSV Loader برای فایلهای با پسوند .CSVبه طور پیش فرض انتخاب میشود. دکمه More اطلاعات بیشتری در مورد آن میدهد.

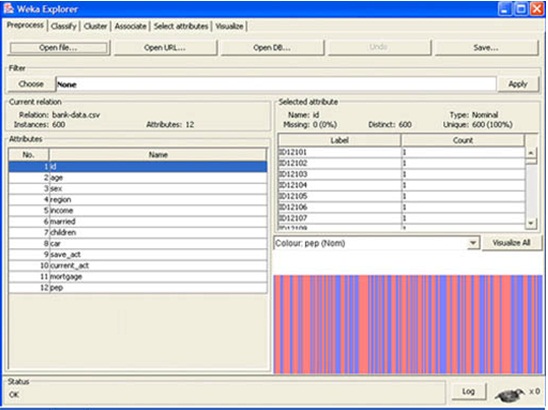
همیشه مطالعه مستندات ارزشمنداست، در این حالت نشان میدهد که ردیف نخست صفحه گسترده، نام ویژگی را تعیین میکند. برای استفاده از این مبدل باید بر Ok کلیک شود. برای مورد مختلف لازم است بر choose کلیک شود تا از لیست انتخاب انجام شود.

  گزینه اول، Arffloader است و فقط به دلیل ناموفق بودن به این نقطه میرسیم.CSVLoader پیش فرض است و در صورت نیاز به فرض دیگر، choose کلیک میشود. دومین گزینه، مربوط به فرمت C4.5 است که دو فایل برای مجموعه داده وجود دارد یکی اسمها و دیگـری داده های واقعـی میباشد. چهارمین برای نمونه های سریالی، برای بازخوانی مجموعه داد‌ه‌ای است که به صورت شیئ سریالی شده جاوا ذخیره شده است. هر شیء در جاوا میتواند در این شکل ذخیره و بازخوانی شود. به عنوان یک فرمت بومی جاوا، سریعتر از فایل ARFF خوانده میشود چرا که فایل ARFF باید تجزیه و کنترل شود. وقتی یک مجموعه داده بزرگ مکررا بازخوانی میشود، ذخیره آن در این شکل سودمند است.

  ویژگیهای دیگر ویرایشگر عمومی اشیا در شکل 4-4 (ب)، save و open است که به ترتیب برای ذخیره اشیای تنظیم شده و بازکردن شیئی که پیش از این ذخیره شده است، به کار میرود. اینها برای این نوع خاص شیئ مفید نیستند. لکن پانلهای دیگر ویرایشگر عمومی اشیاء، خواص قابل ویرایش زیادی دارند. به دلیل مشکلاتی که ممکن است حین تنظیم مجدد آنها رخ دهد، میتوان ترکیب اشیاء ایجاد شده را برای استفاده‌های بعدی، ذخیره کرد.

  تنها منبع مجموعه‌های داده برایWeka ، فایلهای موجود روی کامپیوتر نیستند. میتوان یک URL را باز کرد تا Weka از پروتکل HTTP برای دانلود کردن یک فایل Arff از شبکه استفاده کند. همچنین میتوان یک پایگاه داده‌ها را باز نمود ( open DB ـ هر پایگاه دادهای که درایور اتصال به مجموعه های داده به زبان جاوا JDBC را دارد.) و به وسیله دستور select زبان SQL ، نمونه‌ها را بازیابی نمود. داده‌ها میتوانند به کمک دگمه save به همه فرمتهای ذکر شده، ذخیره شوند. جدای از بحث بارگذاری و ذخیره مجموعه‌های داده، پانل preprocess به کاربر اجازه فیلتر کردن داده‌ها را میدهد. فیلترها، اجزای مهم Weka هستند.

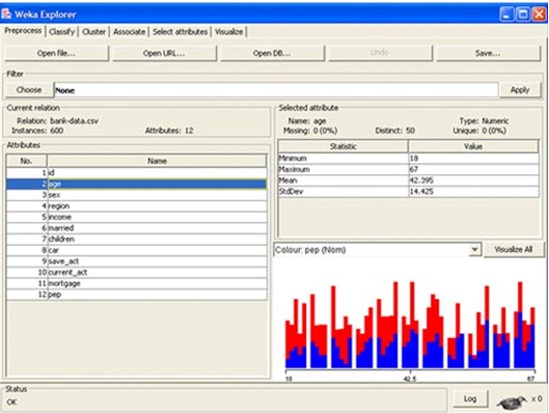
بعد از اینکه فایل بارگذاری شد، Weka فیلدها را تشخیص میدهد و حین بررسی آنها، اطلاعات آماری پایهای را برای هر کدام از صفات محاسبه میکند. همان طور که در شکل 5 نشان داده شده است، لیست صفات تشخیص داده شده، در سمت چپ، پایین و اطلاعات پایگاه داده مربوطه در بالایآن نشان داده میشود.

****

شکل 4-5 بانک اطلاعاتی [**bank-data.csv**](http://maya.cs.depaul.edu/%7Eclasses/ect584/WEKA/data/bank-data.csv)

با کلیک برروی هر کدام از صفات، میتوان اطلاعات آماری اصلی آن را در سمت راست مشاهده نمود. به عنوان مثال شکل 4-6 از انتخاب فیلد age نتیجه شده است.

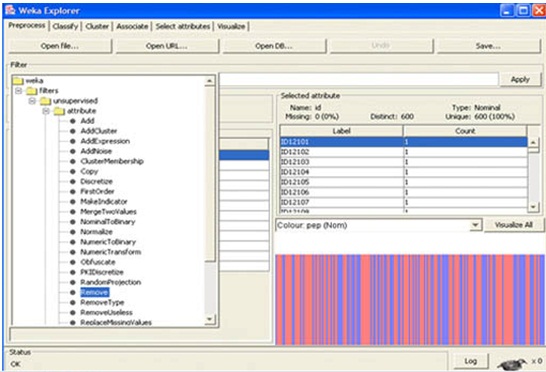
نمودار ترسیم شده در سمت راست، پایین، بر اساس دو فیلد است. فیلد دوم به صورت پیشفرض، آخرین فیلد در پایگاه داده است که میتوان آن را به دلخواه تغییر داد.

****

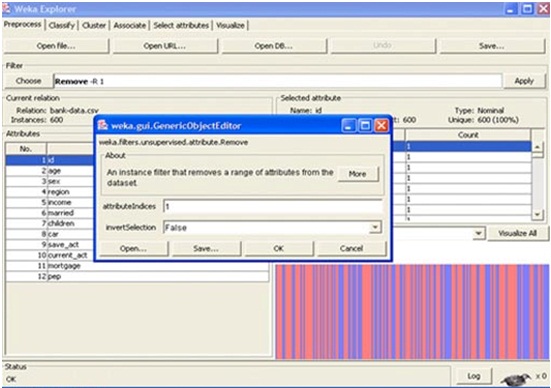
شکل4-6. اطلاعات آماری فیلد age

**4-5-3-2 به کارگیری فیلترها**

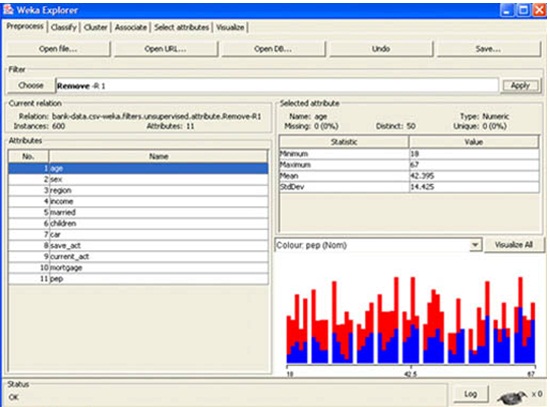
با کلیک دگمهchoose (گوشه بالا و سمت چپ) در شکل 3 میتوان به لیستی از فیلترها دست یافت. میتوان از فیلترها برای حذف ویژگیهای مورد نظر از یک مجموعه داده و یا انتخاب دستی ویژگیها استفاده نمود. مشابه این نتیجه را میتوان به کمک انتخاب ویژگیهای مورد نظر با تیک زدن آنها و فشار دادن کلیه Remove به دست آورد.

****

شکل 4-7 (الف). انتخاب فیلتر Remove

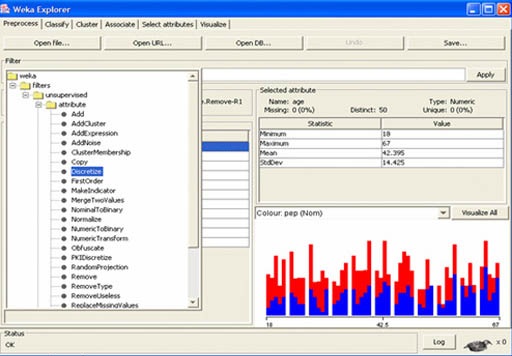


شکل 4-7(ب). وارد کردن شماره فیلد مورد نظر

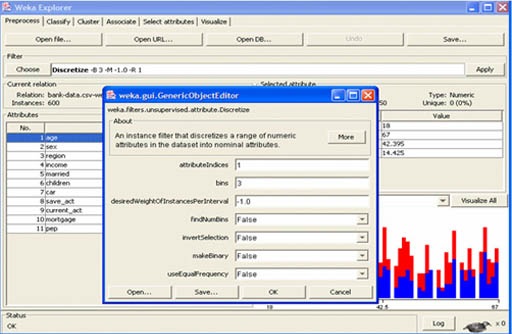


شکل 4-7 (ج).انتخاب گزینه Apply و حذف فیلد id

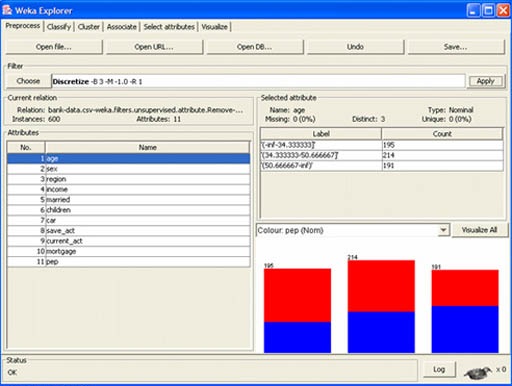
یکی دیگر ازفیلترهای موجود، Discretize است که با استفاده از آن میتوان مقادیریک صفت پیوسته را به تعداد دلخواه بازه گسسته تبدیل کرد. شکل7 مراحل لازم برای شکستن مقادیر صفت age به 3 بازه را نشان میدهد.

****

شکل4-8 (الف). انتخاب فیلتر Discretize



شکل4-8 (ب). وارد کردن شماره فیلد مورد نظر و انجام تنظیمات



شکل 4-8(ج).انتخاب گزینه Apply

**4-5-4 الگوريتم­‏هاي يادگيري**

زماني كه يك الگوريتم يادگيري با استفاده از دگمه choose در پانل classify انتخاب مي‏­شود، نسخه خط فرماني رده بند در سطري نزديك به دگمه ظاهرمي‏­گردد. اين خط فرمان شامل پارامترهاي الگوريتم است كه با خط تيره مشخص مي‏­شوند. براي تغيير آنها مي‏­توان روي آن خط كليك نمود تا ويرايشگر مناسب شيء، باز شود. جدول شكل 6، ليست اسامي رده بندهاي Weka را نمايش مي‏­دهد. اين الگوريتم­‏ها به ردهبندهاي Bayesian ، trees ، functions rules ، lazy و دسته نهايي شامل روش­‏هاي متفرقه تقسيم شده­اند.

  Simple Linear Regresion مدل رگرسيون خطي يك ويژگي مشخص را ياد مي‏­گيرد. آنگاه مدل با كمترين خطاي مربعات را انتخاب مي‏­كند. در اين الگوريتم، مقادير از دست رفته ومقادير غيرعددي مجاز نيستند .

  Linear Regression رگرسيون خطي استانداردكمترين خطاي مربعات را انجام مي‏­دهد مي‏تواند به طور اختياري به انتخاب ويژگي بپردازد، اين كار مي‏­تواند به صورت حريصانه  با حذف عقب رونده انجام شود، يا باساختن يك مدل كامل از همه ويژگي‏‏ها و حذف يكي يكي جمله­‏ها با ترتيب نزولي ضرايب استاندارد شده آنها، تا رسيدن به شرط توقف مطلوب انجام گيرد.

  Least Med sq يك روش رگرسيون خطي مقاوم است كه ميانه (به جاي ميانگين ) مربعات انحراف از خطرگرسيون را كمينه مي‏­كند. اين روش به طور مكرر رگرسيون خطي استاندارد را به زيرمجموعه­‏هايي از نمونه­‏ها اعمال مي‏­كند و نتايجي را بيرون مي‏­دهد كه كمترين خطاي مربع ميانه را دارند.

  SMO teg الگوريتم بهينه سازي حداقل ترتيبي راروي مسايل رگرسيون اعمال مي‏­كند. ( Scholkopf, 1998 ، Smola )   
  Pace Regression ، با استفاده از تكنيك رگرسيون pace ، مدل­‏هاي رگرسيون خطي توليدمي‏­كند (2002 ، Wang و Witten ). رگرسيون pace ، زماني كه تعداد ويژگي‏‏ها خيلي زياد است، به طور ويژه­اي در تعيين ويژگي‏‏هايي كه بايد صرف‏نظر شوند، خوب عملمي‏­كند. در واقع در صورت وجود نظم و ترتيب خاصي، ثابت مي‏­شود كه با بي­نهايت شدن تعداد ويژگي‏‏ها، الگوريتم بهينه عمل مي‏­كند.

  RBF Network ، يك شبكه باتابع پايه­اي گوسي شعاعي را پياده سازي مي‏­كند. مراكز و عرض­‏هاي واحدهاي مخفي به وسيله روش ميانگين K  تعيين مي‏­شود. سپس خروجي­‏هاي فراهم شده از لايه­‏هاي مخفي ،با استفاده از رگرسيون منطقي در مورد رده­‏هاي اسمي و رگرسيون خطي در مورد رده­‏هايعددي، با يكديگر تركيب مي‏­شوند. فعال سازي­‏هاي توابع پايه پيش از ورود به مدل­‏هاي خطي، با جمع شدن با عدد يك، نرماليزه مي‏­شوند. در اين الگوريتم مي‏­توان، K تعداد خوشه­‏ها، بيشترين تعداد تكرارهاي رگرسيون­‏هاي منطقي براي مسأله­‏هايرده­‏هاي رسمي، حداقل انحراف معيار خوشه­‏ها، و مقدار بيشينه رگرسيون را تعيين نمود. اگر رده­‏ها رسمي باشد، ميانگين K به طور جداگانه به هر رده اعمال مي‏­شود تا K خوشه مورد نظر براي هر رده استخراج گردد.

**4-5-5 رده بندهاي Lazy**

ياد گیرنده­‏هاي lazy نمونه­‏هاي آموزشي را ذخيره مي‏­كنند و تا زمان رده بندي هيچ كار واقعي انجام نمي‏­دهند.

  IB1 يك يادگيرنده ابتدايي بر پايه نمونه است كه نزديك­ترين نمونه­‏هاي آموزشي به نمونه­‏هاي آزمايشي داده شده را ازنظر فاصله اقليدسي پيدا كرده و نزديكترين رده­اي مشابه رده همان نمونه­‏هاي آموزشي را تخمين مي‏­زند.

  IBK يك رده بند با K همسايه نزديك است كه معيار فاصلهذكر شده را استفاده مي‏­كند. تعداد نزديكترين فاصله­‏ها (پيش فرض 1= K ) مي‏­تواندبه طور صريح در ويرايشگر شيء تعريف شود. پيش‏بيني­‏هاي متعلق به پيش از يك همسايه مي‏­تواند بر اساس فاصله آنها تا نمونه­‏هاي آزمايشي، وزن­دار گردد.

دوفرمول متفاوت براي تبديل فاصله به وزن، پياده سازي شده­اند. تعداد نمونه­هاي آموزشي كه به وسيله رده بند نگهداري مي‏­شود، مي‏­تواند با تنظيم گزينه اندازه پنجره محدودگردد. زماني كه نمونه­‏هاي جديد اضافه مي‏­شوند، نمونه­‏هاي قديمي حذف شده تا تعدادكل نمونه­‏هاي آموزشي در اندازه تعيين شده باقي بماند.

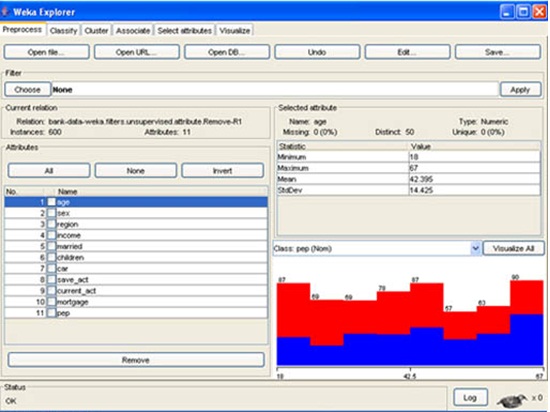
  Kstar ، يك روش نزديكترين همسايه است كه از تابع فاصله­اي عمومي شده بر اساس تبديلات استفاده مي‏­كند.

LWL يك الگوريتم كلي براي يادگيري وزن دار شده به صورت محلي است.اين الگوريتم با استفاده از يك روش بر پايه نمونه، وزن­‏ها را نسبت مي‏­دهد و ازروي نمونه­‏هاي وزن­دار شده، رده بند را مي‏­سازد. رده بند در ويرايشگر شيء LWL انتخاب مي‏­شود. Nave Bayes براي مسايل رده بندي و رگرسيون خطي براي مسايل رگرسيون،انتخاب­‏هاي خوبي هستند. مي‏­توان در اين الگوريتم، تعداد همسايه­‏هاي مورد استفاده را كه پهناي باند هسته و شكل هسته مورد استفاده براي وزن دار كردن را (خطي، معكوس،يا گوسي) مشخص مي‏­كند، تعيين نمود. نرمال سازي ويژگي‏‏ها به طور پيش فرض فعال است.[ Data Mining, witten et Al. 2005 ].

[4]

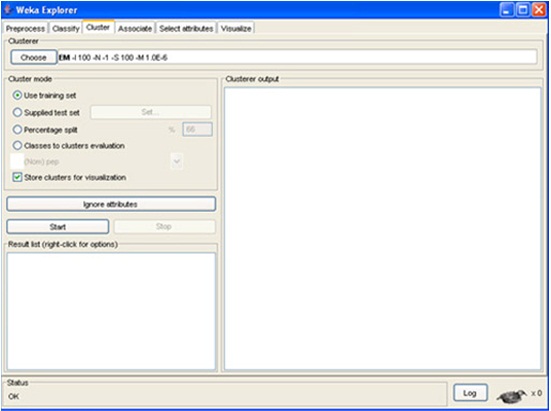
**4-5-6 خوشه[[12]](#footnote-10)**

فایل نمونه مورد استفاده در این قسمت bank-data.csv است که در مرحله preprocess فیلد id را از آن حذف میکنیم )شکل 4-9)

****

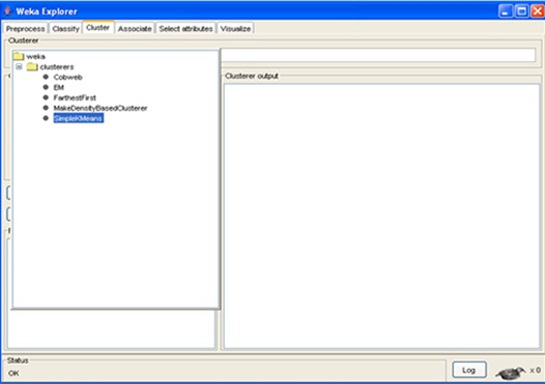
شکل 4-9. بارگذاری فایل نمونه

بعد از باز کردن فایل مورد نظر و با کلیک بر روی پانل cluster پنجرهای مطابق شکل 4-10 باز میشود**.**

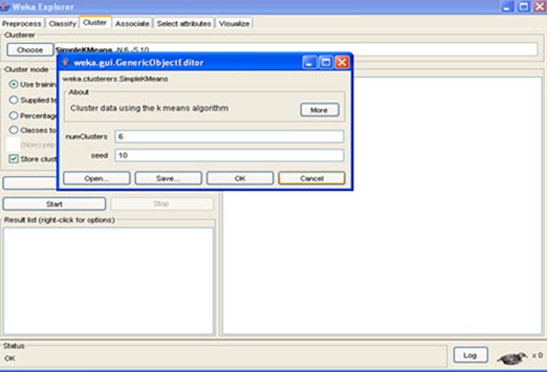
****

شکل 4-10. انتخاب پانل cluster

Weka الگوریتمهای خوشه‌بندی (clustering) متنوعی را پیاده‌سازی میکند. با کلیک دگمه choose در پانل cluster میتوان الگوریتم خوشه‌بندی مورد نظر را انتخاب نمود (شکل 4-11). دراین مثال، الگوریتم SimpleKMeans را انتخاب میکنیم. زمانی که یک الگوریتم خوشه‌بندی انتخاب میشود، نسخه خط فرمانی الگوریتم خوشه‌بندی در سطری نزدیک به دگمه ظاهر میگردد. این خط فرمان شامل پارامترهای الگوریتم است که با خط تیره مشخص میشوند. برای تغییر آنها میتوان روی آن خط کلیک نمود تا ویرایشگر مناسب شیء باز شود (شکل4-12). در این مثال تعداد کلاسترها را 6 تا تنظیم میکنیم.

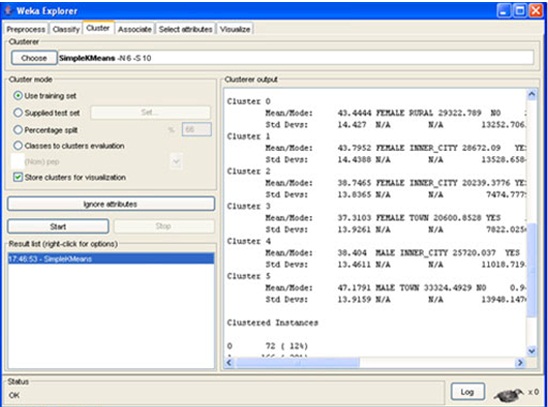
****

شکل 4-11 انتخاب الگوریتم خوشه‌بندی



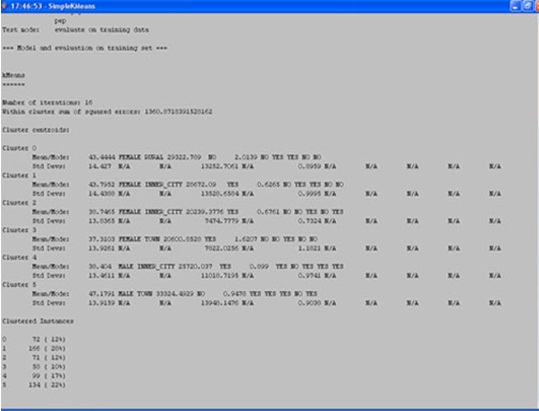
شکل 4-12. تنظیم پارامترهای الگوریتم خوشه‌بندی

باکلیک بر روی دکمه start مدل مورد نظر تولید میشود (شکل4-13)

****

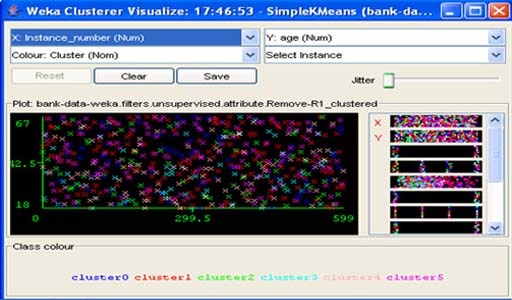
شکل 4-13. مدل حاصل از اجرای الگوریتم خوشه‌بندی

با راست کلیک بر روی مجموعه جواب در پانل Result list در سمت چپ-پایین و انتخاب گزینه‌ی "View in separate window "میتوان نتیجه را در پنجرهای جداگانه مشاهده نمود (شکل 4-14). همانطور که میبینید، اطلاعات آماری مربوط به هرکلاستر، از جمله مرکز ثقل هر کلاستر، تعداد و درصد اعضای هر کلاستردر این پنجره قابل مشاهده است.

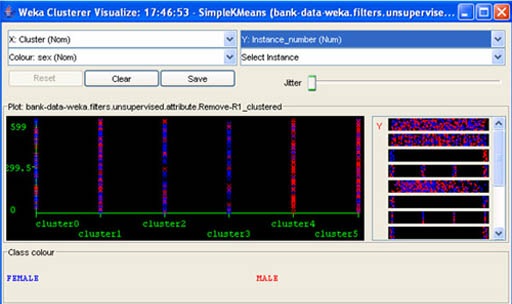
****

شکل 4-14. نتیجه حاصل از اجرای الگوریتم خوشه‌بندی

روش دیگر برای کسب اطلاعات در مورد هرکلاستر، مصورسازی است. با راست کلیک بر روی مجموعه جواب در پانل Result list در سمت چپ و انتخاب گزینه ی " Visualize cluster assignments" پنجرهای مطابق شکل4-15(الف) باز میشود. انتخابهای مختلف برای هرکدام از سه بعد نمودارحاصل (محور X، محور Y، رنگ) نمودارهای مختلفی را نتیجه میدهد که میتوان از آنها اطلاعات مورد نظر را بدست آورد.به عنوان مثال در شکل4-15(ب)، محور Xها نماینده شماره کلاستر، محور Yها نماینده شماره نمونه در بانک اطلاعاتی، و رنگها نماینده جنسیت هستند (قرمز: مرد، آبی: زن). همانطورکه مشاهده میشود کلاستر2 بیشتر توسط زنان احاطه شده است و کلاستر4 توسط مردان.

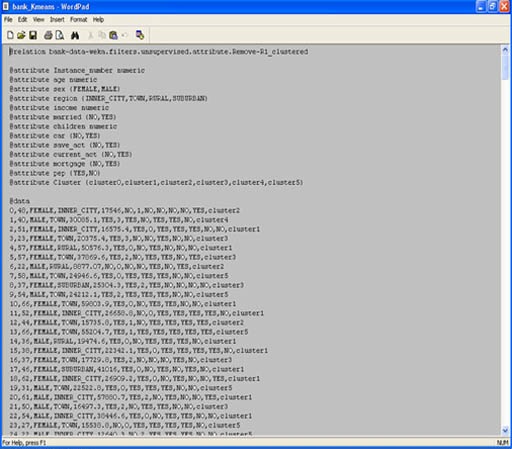
****

شکل4-15 (الف)



شکل4-15(ب) مصورسازی نتیجه حاصل از clustering

علاوه بر این ممکن است علاقه‌مند باشیم که بدانیم هر نمونه در بانک اطلاعاتی، به کدام کلاستر اختصاص داده شده است. برای این منظور در پنجره شکل 4-15(ب)، گزینه save را انتخاب میکنیم و فایل مورد نظر را با نام " bank-data-Kmeans" ذخیره میکنیم. فایل حاصل رامیتوان از طریق یک نرم‌افزار پردازش متن مثل Notepad یا word باز کرد. بخش ابتدایی این فایل در شکل4-16 نشان داده شده است. همانطورکه مشـاهده میکنید، Weka ویژگی جدیدی به نام Cluster را به مجموعه ویژگیهای موجود اضافه کرده است.

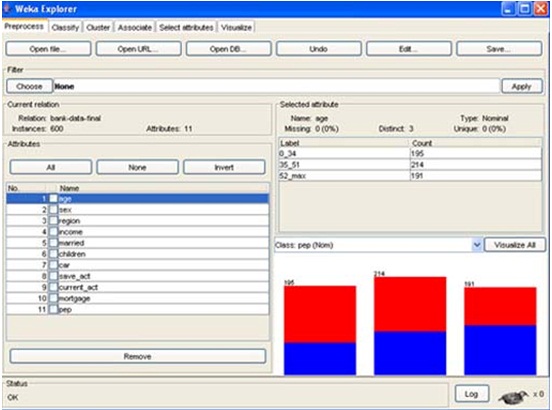
****

شکل4-16. نتیجه اختصاص نمونه‌ها به کلاسترها

**4-5-7 وابسته[[13]](#footnote-11)**

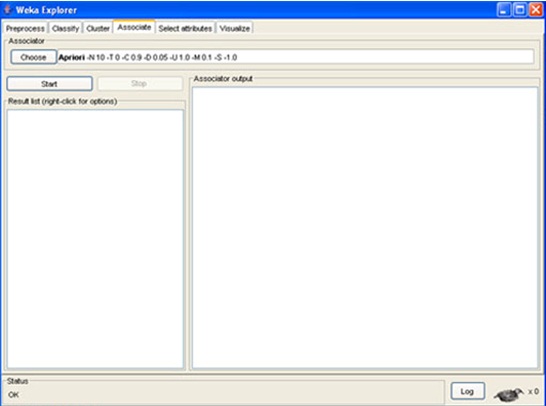
در این قسمت میتوان قوانین Association حاکم بربانک اطلاعاتی را با استفاده از الگوریتم‌های مختلف بدست آورد.

فایل نمونه مورد استفاده در این بخش bank-data-final.arff است که در مرحله preprocess فیلد id از آن حذف شده و مقادیرصفات age و income به صورت گسسته در آمده است (شکل4-17)

****

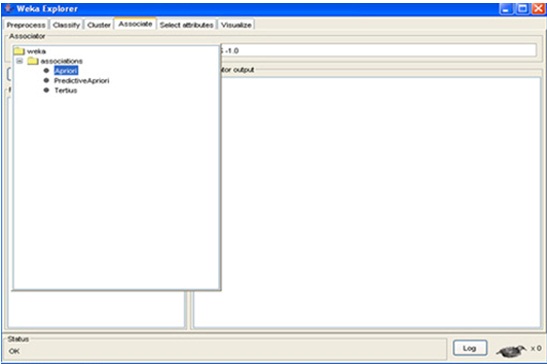
شکل 4-17. بارگذاری فایل نمونه

بعد از باز کردن فایل مورد نظر و با کلیـک بر روی پانل Associate پنجره‌ای مطـابق شکـل4-18 باز میشود**.**

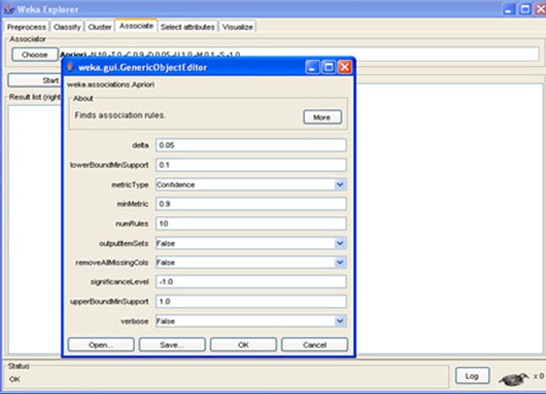
****

شکل 4-18. انتخاب پانل Associate

Weka الگوریتمهای association متفاوتی را پیاده‌سازی میکند. با کلیک دگمه choose در پانل Associate میتوان الگوریتم association مورد نظر را انتخاب نمود (شکل 4-19). دراین مثال، الگوریتم Apriori را انتخاب میکنیم. زمانی که یک الگوریتم انتخاب میشود، نسخه خط فرمانی آن الگوریتم در سطری نزدیک به دگمه ظاهر میگردد. این خط فرمان شامل پارامترهایالگوریتم است که با خط تیره مشخصمیشوند. برای تغییر آنها میتوان روی آن خط کلیک نمود تا ویرایشگر مناسب شیء باز شود (شکل4-20). در این مثال همان پارامترهای پیشفرض را میپذیریم.

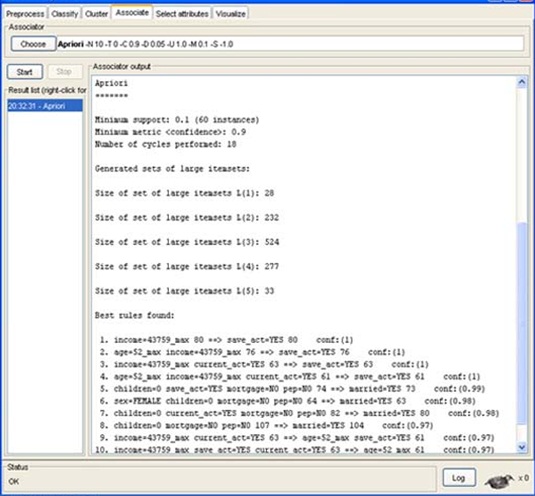
****

شکل 4-19. انتخاب الگوریتم association



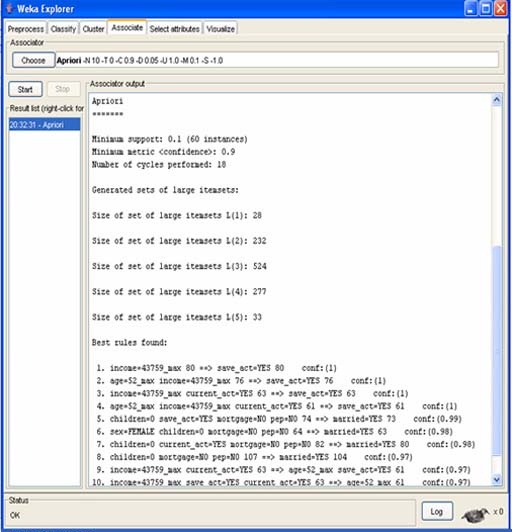
شکل 4-20 تنظیم پارامترهای الگوریتم association

با کلیک بر روی گزینه More در شکل4-20، میتوان توضیحات لازم را درمورد هر کدام از پارامترها بدست آورد. شکل4-21 این توضیحات را نشان میدهد**.**

****

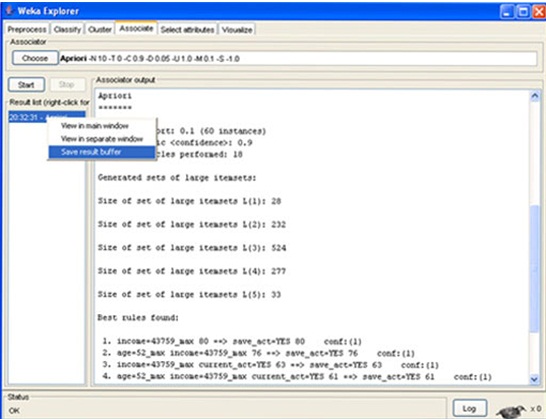
شکل 4-21. توضیحات بیشتردر مورد پارامترهای الگوریتم Apriori

بعد از تنظیم پارامترهای مورد نظر، با کلیک بر روی گزینه start، قوانین بدست آمده که بر اساس metric Type مرتب شده‌اند در صفحه اصلی نمایش داده میشوند.(شکل4-22)

****

شکل 4-22. نتیجه اجرای الگوریتم Apriori

با راست کلیک بر روی مجموعه جواب در پانل Result list در سمت چپ و انتخاب گزینه‌ی مورد نظر، میتوان نتایج بدست آمده را در پنجره‌ای جداگانه نمایش داد و یا قوانین بدست آمده را به فرمت دلخواه (.csv, .arff, .txt) ذخیره کرد. شکل4-23 گزینه‌های موجود را نشان میدهد.

****

شکل 4-23. ذخیره قوانین بدست آمده

###### فصل ششم

###### منابع

1. Source: W.Frawley and G. Piatetsky. Knowledge Discovery I DataBases.ISSN 0738-4602 [↑](#footnote-ref-1)
2. Source: W.Frawley and G. Piatetsky. Knowledge Discovery I DataBases.ISSN 0738-4602 [↑](#footnote-ref-2)
3. Source: R.Grossman [↑](#footnote-ref-3)
4. Knowledge Discovery From Database [↑](#footnote-ref-4)
5. Decision Support System [↑](#footnote-ref-5)
6. پيوست 1 [↑](#footnote-ref-6)
7. Metadata [↑](#footnote-ref-7)
8. آدرس پست الکترونیکی: [info@peivand.org](mailto:info@peivand.org)  
   وب سایت: <http://www.peivand.org/> [↑](#footnote-ref-8)
9. Information and Communication Technology)ICT( [↑](#footnote-ref-9)
10. [1]

    -1CHRIS RYGIELSKI, “DATA MINING TECHNIQUES FOR CUSTOMER RELATIONSHIPMANAGEMENT”, TECHNOLOGY IN SOCIETY, 2002.

    2- HILL L., “CRM: EASIER SAID THAN DONE”, INTELLIGENT ENTERPRISE,

    1999

    3- Microsoft Visual Studio .Net Documentation

    4- Client/Server Survival Guide by Robert Orfali, Dan Harkey, Jeri Edwards

    5- شاه‌سمندي، پرستو «داده‌كاوي در مديريت ارتباط با مشتري» (1384)، مجله تدبير شماره 156.

    6- Jeffery W. Seifert , Analyst in information science and Technology Policy, ‘ Data Mining : An Overview ‘ December 2004.

    *[2]*

    1-دانشنامه آزاد ویکی پدیا

    2-ماهنامه عملی آموزشی تدبیر شماره 156

    *.3Hamshahri Newspaper*

    *.4Barbara Mento and Brendan Rapple, SPEC Kit 274: Data mining and data warehousing, Association of Research Libraries, Washington, DC (2003, July)*

    *.5*[*http://www.infotechera.com/*](http://www.infotechera.com/)

    *.6*[*http://www.ece.ut.ac.ir/dbrg/index.htm*](http://www.ece.ut.ac.ir/dbrg/index.htm)

    *.7http://www.irandoc.ac.ir/index.htm*

    *.8http://www.arts.uci.edu/dobrain/gems.980415b.htm*

    1جمالي،آرمان-شهرالكترونيكي،بسترورودبه رقابت­هاي عصرسيبرنتيكhttp /www.editorial.com

    [2] Two Crows Corporation. *Introduction to data mining and knowledge discovery*, (3rd ed.) (p. 4). Potomac, MD7 Two Crows Corporation, 1999.

    [3] Rahman Hakikur , *Social and Political Implications of Data Mining: Knowledge Management in E-Government,* Published in the United States of America byInformation Science Reference (an imprint of IGI Global), Hershey • New York, ISBN 978-1-60566-230-5, 2009.

    [4]Kazienko Przemyslaw, Adamski Michal., *“AdROSA—Adaptive Personalization of Web Advertising”*, *Information Sciences* (177) ,2269 – 2295(2007).

    [5] Zhou Ping, Le Zhongjian,*A Framework for Web Usage Mining in Electronic Government*, Integration and Innovation Orient to E-Society, Volume 2, pp. 487-496, inIFIP International Federation for Information Processing, December 27,2007**.**

    [6] Chan, P., Fan W., Prodromidis, A. and Stolfo, S. *Distributed data mining in credit card fraud detection*, *IEEE Intelligent Systems*, pages 67–74, Nov/Dec 1999. [↑](#endnote-ref-1)
11. نوشته ویوک.ان. پاتکار(V. N. Patkar)  
    ترجمه مریم صراف زاده و افسانه حاضری

    دنیای صنعت برق

    www.sbargh.ir

     [↑](#endnote-ref-2)
12. Cluster [↑](#footnote-ref-10)
13. Association [↑](#footnote-ref-11)