بخش سوم

مدیریت داده‌کاوی

فصل ششم: انباره داده‌ها

فصل هفتم: متدلوژی اجرا و پیاده‌سازی پروژه‌های داده‌کاوی

فصل ششم

انباره داده‌ها

سازمانها درک کرده‌اند که سیستم‌های انباره‌داده ابزارهای ارزشمندی در رقابتهای امروزه هستند. بنگاههای بسیاری، میلیونها دلار برای ساخت انباره‌داده صرف کرده‌اند. طبق تعریف اینمون[4] یکی از پیشتازان معماری در ساخت سیستمهای انباره‌داده، انباره‌داده مجموعه‌ای موضوع‌گرا، یکپارچه، از زمانهای مختلف و غیرفرار به منظور پشتیبانی از فرآیند تصمیم‌سازی است. داده‌کاوی چیزی فراتر از پردازش بر روی یک پایگاه داده معمولی می‌باشد. مثالهای زیر این تفاوت را نشان می‌دهند. یک پرس و جوی ساده و پیدا کردن تمامی افراد با نام «علی» در یک پایگاه داده بسیار ساده است ولی در مقابل پیدا کردن افرادی که کارت اعتباری آنها وضعیت مناسبی ندارد و در مرز ورشکستگی می‌باشند، خیلی ساده نیست. پیدا کردن افرادی که بیش از ۱۰۰/۰۰۰ تومان خرید داشته‌اند ساده است ولی در مقابل پیدا کردن افرادی که عادتهای خرید مشابهی دارند و یکسری اقلام خاصی را با هم خرید می‌کنند، کار ساده‌ای نیست. پیدا کردن افرادی که در یک تاریخ خاص از یک فروشگاه خاص شیر خریده‌اند، بسیار ساده است ولی در مقابل پیدا کردن افرادی که غالباً شیر خریداری می‌کنند خیلی ساده نیست.

با توجه به مثالهای مطرح شده، کاملاً مشخص است که داده‌هایی که در داده‌کاوی و الگوریتم‌های آن استفاده می‌شوند، تفاوت عمده‌ای با داده‌های عادی در پایگاه داده‌ها دارند. جهت فراهم کردن این نوع داده‌ها که در انباره‌داده‌ها قرار می‌گیرند، باید یکسری پردازش‌های خاص روی آنها صورت پذیرد. این نوع پردازش‌ها به نام‌های پاکسازی داده‌ها و یکپارچه‌سازی داده‌ها معروفند، که در فصل دوم کاملاً توضیح داده شده‌اند.

۶-۱-داده‌کاوی و انباره‌داده‌ها

در دهه ۹۰ میلادی پدیده انباره‌داده‌ها ظهور یافت. قبل از انبارسازی داده‌ها سیستمهای کامپیوتری جهت ذخیره، جمع‌آوری، تغییر و تصحیح داده‌ها طراحی شده بودند. این سیستمهای اولیه به سیستمهای عملیاتی یا میراثی[1] موسوم هستند. گرچه جمع‌آوری و ذخیره داده‌ها کارهای مفیدی به حساب می‌آیند، اما دسترسی به آنها و تحلیل داده‌های عملیاتی به راحتی امکان پذیر نیست و علت نیز عدم یکپارچگی داده‌ها می‌باشد. هر برنامه کاربردی داده‌ها را بنا بر نیاز خود تفسیر می‌کند.

مدیران برای تصمیم‌گیری نیازی به کوهی از اطلاعات جزئی روزانه ندارند، آنها نیازمند چکیده اطلاعات برای دوره‌های زمانی متفاوت می‌باشند و به همین علت است که داده‌های تاریخی دارای مفهوم و ارزش بیشتری می‌باشد. انباره‌داده‌ها برای شرکتهایی که مصمم به استفاده از داده‌کاوی هستند یک ضرورت می‌باشد، یکی از ماهیتهای وجودی انباره‌داده‌ها، یکپارچگی داده‌ها هنگام قرارگیری در انباره‌داده‌هاست. این بدین معنی است که دقت بسیاری به‌کار گرفته می‌شود تا یکنواختی و پیوستگی در درک اهداف عام سازمانی بوجود آید. اگر انباره‌داده‌ها وجود نداشته باشد، داده‌کاو باید زمان بسیار زیادی را صرف جمع‌آوری، پاکسازی و یکپارچه‌سازی داده‌ها کند. بدین ترتیب وقت بسیاری باید صرف شود تا تحلیل داده‌ها آغاز شود.

در انباره‌داده‌ها، داده‌های تاریخی جمع‌آوری و سازماندهی می‌شوند. وجود داده‌های تاریخی برای یافتن الگوها و روابطی که سازمان به دنبال آنهاست، برای داده‌کاوی یک ضرورت است. اگر چنانچه این داده‌های تاریخی وجود نداشته باشند، داده‌کاو باید به دنبال جمع‌آوری آنها باشد.

علت دیگر اهمیت انباره‌داده‌ها این است که انباره‌داده‌ها شامل داده‌های جزئی و داده‌های کلی، در کنار یکدیگر می‌باشد. بدون تردید، داده‌کاو به اطلاعات جزئی برای تحلیل نیازمند است، اما داده‌های خلاصه شده نیز به‌کار می‌آیند. از آنجا که در انباره‌داده انواع داده‌های خلاصه شده وجود دارد، داده‌کاو می‌تواند به سرعت داده‌های انباره‌داده را بررسی کند و این باعث کاهش تکرار تحلیلها توسط داده‌کاو می‌شود.

---

[1]- Legacy

۶-۲- مفاهیم انباره‌داده‌ها

ویژگیهای مهم یک انباره‌داده عبارتند از:

- موضوع محوری[۱]
- جامعیت[۲]
- پویاپذیری[۳] (مهم بودن عامل زمان)
- پایایی[۴] (غیرفرار و دایمی بودن)

موضوع محوری: داده‌ها طبق یک موضوع خاص سازماندهی می‌شوند، به‌عنوان مثال داده‌های مربوط به مشتریان، محصولات و یا داده‌های مرتبط با فروش، هر کدام جداگانه در نظر گرفته می‌شوند. اما در پایگاه داده‌های معمولی، داده‌ها بر اساس عملیات و پردازشهای روزانه ایجاد می‌شوند و موضوع آنها مرتبط با کل پردازش می‌باشد.

جامعیت: داده‌های انباره، از تجمع دیگر داده‌ها ساخته می‌شوند. این داده‌ها ممکن است مربوط به پایگاه داده‌های رابطه‌ای، فایلهای بدون ساختار و یا رکوردهای مرتبط با پردازش‌های برخط باشند. جهت یکسان‌کردن داده‌ها، روشهای پاکسازی به‌کار برده می‌شود. نوعی هماهنگی کلی در مورد داده‌های مختلفی که از سیستمهای متفاوت آمده‌اند، لازم است. به‌عنوان نمونه لازم است مقادیر عددی مربوط به «قیمت هتل»، «هزینه صبحانه»، «مالیات» و دیگر موارد مشابه که ممکن است از مکانهای متفاوت آمده باشند، یکسان شده و یک انباره‌داده با نام «مسافر» ایجاد شود.

پویاپذیری: افق زمانی برای انباره‌داده‌ها بسیار مهم‌تر ازداده‌های مرتبط با سیستمهای عملیاتی می‌باشد. در ساختار انباره‌داده‌ها عاملی به نام زمان در نظر گرفته می‌شود. این عامل می‌تواند به‌طور ضمنی و یا به وضوح بیان شود. اما در سیستمهای عملیاتی عامل زمان، عاملی کلیدی نیست.

۱- Subject Oriented
۲- Integrated
۳- Time Variant
۴- NON Volatile

پایائی: پایگاه داده‌ها شامل داده‌هایی است که روزانه با آنها کار می‌شود و بخشهایی به آن اضافه و یا از آن حذف می‌شود، در مقابل انباره‌داده این ویژگی را ندارد. با توجه به همین امر واضح است که به‌روز شدن داده‌ها در انباره‌داده‌ها مقدور نمی‌باشد انباره‌داده‌ها نیازی به پردازشهایی از قبیل: تراکنشهای داده‌ای، بازیافت و مکانیزم‌های کنترل هم‌زمان ندارد. تنها اعمالی که در انباره‌داده‌ها صورت می‌پذیرد عبارتند از: بارگذاری (مقدار دهی) اولیه داده‌ها و دسترسی به داده‌ها.

پردازشی که بر روی انباره‌داده‌ها انجام می‌گیرد *OLAP* نامیده می‌شود (*OLAP* در مقابل *OLTP* آمده است که پردازش‌هایی است که بر روی داده‌های پایگاه داده‌ها انجام می‌گیرد). جدول (٦-١) *OLAP*[1] و *OLTP*[2] را با توجه به برخی پارامترهای مهم مقایسه کرده است.

جدول ٦-١) مقایسه پردازشها در انباره‌داده‌ها و پایگاه داده‌ها

ویژگیها	*OLAP*	*OLTP*
کاربران	کارشناسان خبره	اپراتورها
کارکرد	تصمیم گیری	کارهای روزمره
طراحی پایگاه داده	بر مبنای موضوع	بر مبنای کاربرد
داده	تاریخی، چندبعدی مجتمع	جاری، روزبه‌روز و با جزییات
کاربرد	در موارد خاص	تکراری
نحوه دسترسی	کاوش و کشف	خواندن ونوشتن
واحد کاری	جستجوهای پیچیده	پردازش ساده
تعداد رکوردها	میلیونها	دهها
تعداد کاربران	صدها	هزاران
اندازه پایگاه داده	۱۰۰ *TB_GB*	۱۰۰ *MB_GB*
شاخص	پردازش	جستجو و پاسخ

[1]- Online Analytical Process

[2]- Online Transaction Process

این پارامترها عبارتند از: «کاربران»، «کارکرد»، «طراحی پایگاه داده»، «داده»، «کاربرد»، «نحوه دسترسی»، «واحد کاری»، «تعداد رکوردهای در دسترس»، «تعداد کاربران»، «اندازه پایگاه داده» و «شاخص».

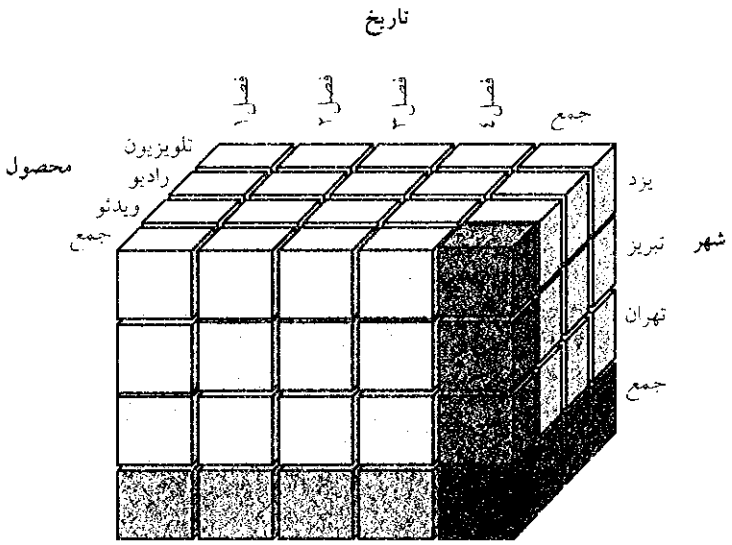
ساختار انباره‌داده

انباره‌داده‌ها بر مبنای ساختاری چند بعدی، که مدل داده‌ها را به شکل مکعب اطلاعاتی نشان می‌دهد، ساخته شده است. یک مکعب اطلاعاتی داده‌ها را در چندین بعد مختلف مدلسازی می‌کند. در زیر مثالی از یک مکعب اطلاعاتی به نام «فروش» بررسی شده است که شامل این ابعاد می‌باشد:

Item(نوع، نام تجاری، نام کالا)

Time (سال، فصل، ماه، هفته، روز)

یک مکعب اطلاعاتی اجازه می‌دهد که داده‌ها در ابعاد مختلفی مدلسازی شده و استفاده شوند. عموماً سازمانها با توجه به نیازهای آینده‌شان ابعاد مختلفی از داده‌ها را نگهداری می‌کنند. به‌عنوان مثال داده‌های فروش در سازمانها نیاز است با دیدگاههای زیر ذخیره شوند: در فرایند فروش مهم است که دقیقاً چه اقلامی فروخته شده‌اند. نام آنها چه بوده است نام تجاری آنها چیست و دیگر اطلاعات مرتبط. زمانهای فروش نیز مهم است.



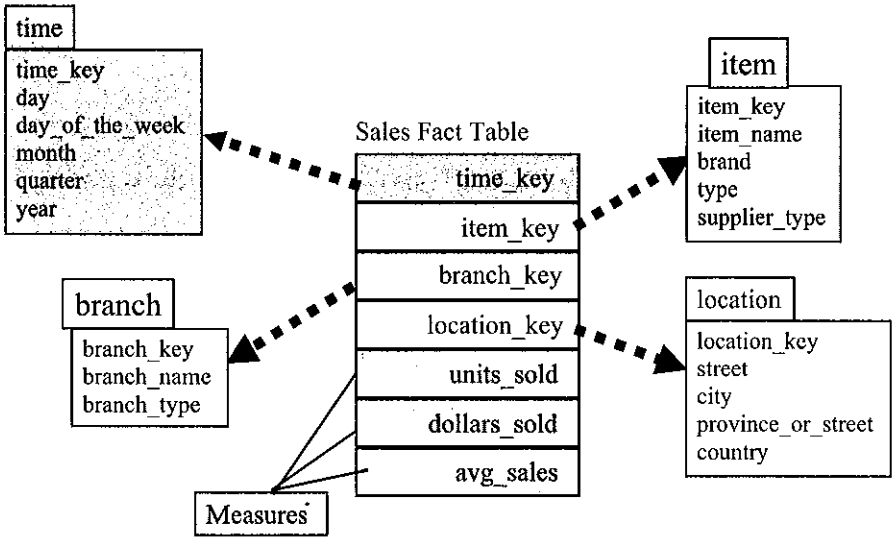
شکل ٦-١) ابعاد داده فروش

اینکه فروش روزانه، ماهانه، فصلی و سالانه هر کدام چه میزان بوده است. اگر فروش در مکانهای مختلف جغرافیایی صورت گرفته است، هر کدام چگونه بوده‌اند. شکل (٦-١) این ابعاد را نشان می‌دهد.

۶-۲-۱- مدل مفهومی انباره‌داده‌ها

در ادامه به چند مفهوم اساسی در انباره‌داده‌ها اشاره می‌کنیم:

مدل ستاره‌ای[1]: یک جدول اصلی که متصل به مجموعه‌ای از جداول دیگر باشد مدل ستاره‌ای نامیده می‌شود. شکل (٦-٢) یک مدل ستاره‌ای می‌باشد که جدول اصلی فروش[2] را به جداول دیگر از جمله اقلام فروش[3]، مکان[4] و زمان[5] و شعب فروش[6] متصل می‌کند.



شکل ٦-٢) مدل ستاره‌ای

---

[1]- Star Schema

[2]- Sales

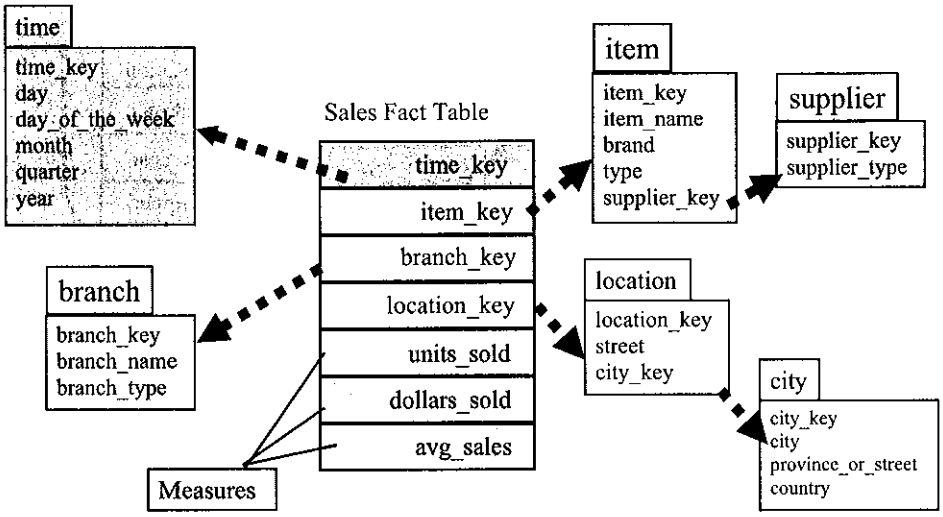
[3]- Item

[4]- Location

[5]- Time

[6]- Branch

اگر در مدل ستاره‌ای جداول جانبی با شکستن به چند جدول، نرمال شوند. این مدل تبدیل به مدل برف‌دانه[1] می‌شود. به‌عنوان مثال در شکل (٦-٣) جدول مکان به دو جدول شکسته شده است و بخشی از اطلاعات آن در جدول دیگری به نام شهر[2] وارد شده است.



شکل ٦-٣) مدل برف‌دانه‌ای

حال اگر این اطلاعات اصلی به همراه جداولش به اطلاعات دیگری نیز مرتبط باشند و بخواهیم این ارتباطات را نیز نمایش دهیم از مدل صورت فلکی[3] استفاده می‌کنیم. شکل (٦-٤) ارتباطات جدول اصلی فروش را با جدول اصلی دیگری به نام حمل و نقل[4] نشان می‌دهد.

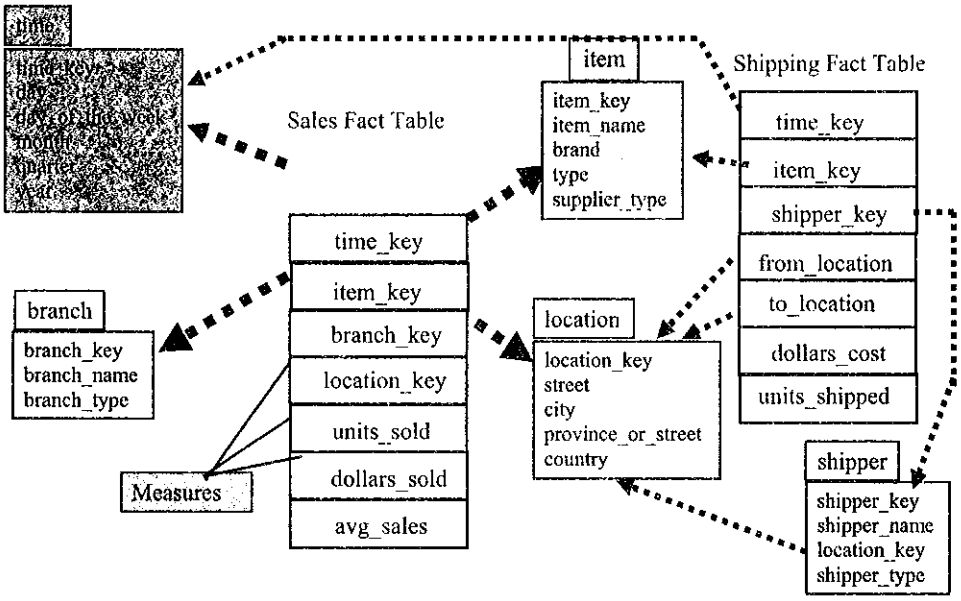
---

[1]- Snow-Flake Schema

[2]- City

[3]- Fact Constellation

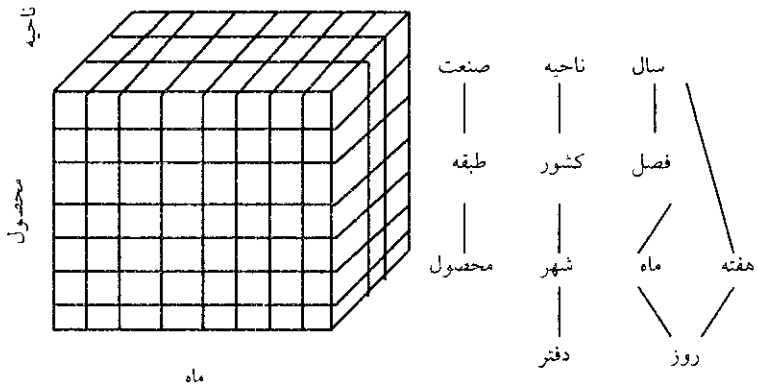
[4]- Shipping



شکل ٦- ٤) مدل صورت فلکی

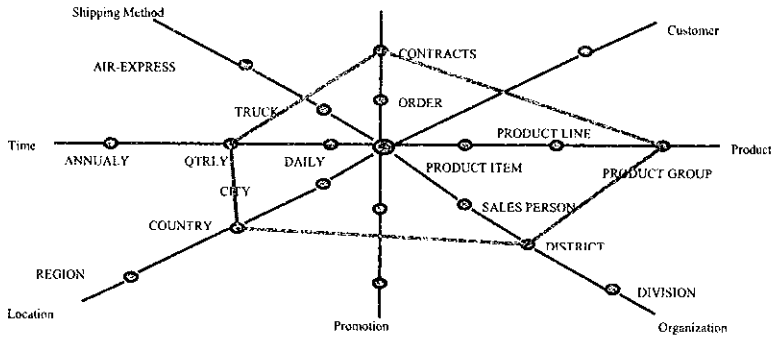
داده‌های چند بعدی

«فروش» یک مفهوم چندبعدی است که از بخشهای محصول، زمان فروش و محل فروش تشکیل شده است و می‌توان دادهٔ چند بعدی فروش را به شکل زیر نمایش داده در شکل (٦-٥)، سلسله مراتب ابعاد داده‌ها نمایش داده شده‌اند.



شکل ٦-٥) ابعاد مختلف داده‌های فروش

اگر ابعاد داده‌ها بیشتر از سه بعد باشد، می‌توان با مدل شبکه ستاره‌ای[1] آن را نمایش داد. شکل (٦-٦) داده‌های مرتبط با ابعاد فروش را نشان می‌دهد. روی هرمحور در این مدل یک بعد نمایش داده می‌شود. به‌عنوان مثال بعد زمان روی یک محورنشان داده شده و علاوه بر آن سلسله مراتب زمان نیز که عبارتند از روزانه، فصلی و سالانه روی این محور نشان داده می‌شوند به‌عنوان مثال اگر یک محصول خاص، فروش فصلی داشته و در یک مکان جغرافیایی خاص فروخته شود، در این صورت نقاط روی این نمودار که حاوی این اطلاعات می‌باشند با یک خط شکسته به هم متصل می‌شوند و بدین ترتیب داده‌های این محصول خاص با ابعاد مورد نظر نمایش داده می‌شود.



شکل ٦-٦) مدل شبکه ستاره‌ای

زبان *MQL* جهت پیاده‌سازی انباره‌داده‌ها

زبان *MQL*[2] شبیه زبان *SQL*[3] می‌باشد و برای تعریف هر کدام از مفاهیم ارائه شده در بخش قبلی روشهای خاصی وجود دارد. جهت تعریف یک جدول اصلی از دستور *Define Cube* و

---

[1]- Star Net

[2]- Mining Query Language

[3]- Structured Query Language

برای تعریف جداول جانبی از دستور *Define Dimension* استفاده می‌شود. مثال زیر دستورات مربوط به شکل مدل ستاره‌ای را به زبان *MQL* نشان می‌دهد.

```
define cube sales_star [time, item, branch, location ]:
dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars) ,
    units_sold = count(*)
define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier_type)
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key,street, city, province_or_state, country)
```

دستورات زیر تعریفی از مدل برف دانه‌ای با استفاده از زبان *MQL* است.

```
define cube sales_snowflake [time, item, branch, location]:
dollars_sold = sum(sales_in_dollars), avg_sales = avg(sales_in_dollars),
    units_sold = count(*)
define dimension time as (time_key, day, day_of_week, month, quarter, year)
define dimension item as (item_key, item_name, brand, type, supplier(supplier_key,
    supplier_type))
define dimension branch as (branch_key, branch_name, branch_type)
define dimension location as (location_key,street,city(city_key, province_or_state,
    country))
```

۶-۲-۲- فرایند طراحی انباره‌داده

برای طراحی یک انباره‌داده مؤثر، اولین مرحله، درک و تحلیل نیازهای کسب و کار و ساخت چارچوب تحلیل کسب و کار می‌باشد. ساخت یک سیستم اطلاعاتی پیچیده و بزرگ می‌تواند مانند ایجاد یک ساختمان بزرگ و پیچیده در نظر گرفته شود که مالک، معمار و سازنده آن دیدهای مختلفی داشته واین نظرات برای تشکیل یک چارچوب پیچیده ترکیب می‌شوند. دیدگاه‌های مختلفی در طراحی انباره‌داده وجود دارد که عبارتند از:

- **دید بالا به پایین**[1]: روش بالا به پایین عبارت است از روشی که در آن ابتدا یک طرح کلی ایجاد شده و سپس به جزئیات پرداخته می‌شود.
- **دید پایین به بالا**: ابتدا نمونه‌های کوچک ساخته شده و سپس نمونه گسترش داده می‌شود.
- **دید پرس و جوی کسب وکار**[2]: یک شکل کلی از داده‌های انباره‌داده بر مبنای نگرش کاربر نهایی ایجاد می‌شود.

از دیدگاه مهندسی نرم افزار دو روش عمده جهت طراحی انباره‌داده‌ها وجود دارد که عبارتند از مدل آبشاری و مدل مارپیچی.

- **مدل آبشاری**[3]: این مدل یک روش ساخت‌یافته و نظام‌مند بوده که در ایجاد انباره‌داده، گام به گام جلو می‌رود.
- **مدل مارپیچی**[4]: این روش یک نوع مدلسازی سریع است که در ابتدا یک مدل کوچک ساخته شده و سپس با بررسی مجدد آن را بهبود می‌دهند.

۶-۲-۳- معماری انباره‌داده

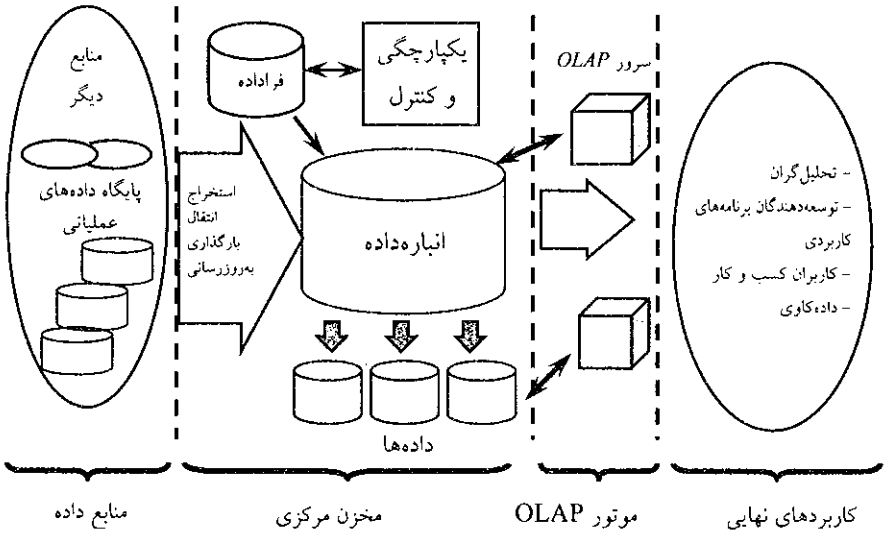
رویکرد چند لایه در انباره‌داده‌ها نیازمند این است که داده‌ها به شکلهای مختلف درآیند. این رویکرد باعث به‌وجود آمدن یک سیستم جامع برای مدیریت داده به منظور تصمیم‌گیری می‌شود. مهم‌ترین اجزاء این سیستم همان‌طور که در شکل (٦-٧) نشان داده شده است عبارتند از:

- سیستمهای منبع[5]، جایی که داده‌ها از آنجا می‌آیند و همان سیستمهای عملیاتی می‌باشند.
- استخراج، انتقال و بارگذاری داده میان منابع مختلف داده.
- مخزن مرکزی[6]، محل اصلی ذخیره‌سازی داده در انباره‌داده‌ها است و یک پایگاه دادهٔ رابطه‌ای با مدل منطقی می‌باشد.
- مخزن فراداده[1]، توضیح می‌دهد که چه چیزهایی وجود داشته و در کجا موجود هستند.

---

[1]- Top-Down
[2]- Business Query
[3]- Water Fall
[4]- Spiral
[5]- Source System
[6]- Central Repository

- فراداده، دسترسی سریع و اختصاصی را برای کاربران نهایی و برنامه‌های کاربردی فراهم می‌کند.
- بازخور عملیاتی، سیستمهای پشتیبان تصمیم را با سیستمهای عملیاتی یکپارچه می‌کند.
- کابران نهایی، مهم‌ترین دلیل اصلی توسعه انباره‌های داده در مرحله اول می‌باشند. آنها از داده‌ها و دانش استخراج شده از آنها استفاده می‌کنند.



شکل ٦-۷) معماری انباره‌داده‌ها

تقریباً همهٔ مؤلفه‌هایی که ذکر شد در تمامی انباره‌های داده وجود دارند. داده همانند آب می‌باشد که از منابع سیستم سرچشمه گرفته و در انباره‌داده جاری شده تا به کاربران نهایی ارائه شود. این مؤلفه‌ها در بسترهای سخت‌افراز، نرم‌افراز و شبکه سوار شده‌اند، این زیرساختارها، باید به اندازه کافی قوی باشند تا نیازمندی‌های کاربران نهایی و همچنین نیازمندیهای پردازش و رشد داده را پوشش دهند. در ابتدا با چهار عمل اصلی استخراج[۲]، به‌روزرسانی[۳]، بارگذاری[۱] و انتقال[۲]،

---

[۱]- Metadata Repository

[۲]- Extract

[۳]- Refresh

داده‌ها از پایگاه داده‌های معمولی جمع‌آوری شده و به انباره‌داده فرستاده می‌شوند. داده‌های موجود در انباره‌داده‌ها مستقیما با دیگر سیستمها در ارتباط نیستند و اگر یک سیستم عملیاتی بخواهد از داده‌های انباره‌داده استفاده کند از بازارچه داده‌ها[۳] استفاده می‌کند. بازارچه داده‌ها بخشی است که داده‌های مربوط به یک برنامهٔ کاربردی خاص را به‌طور موقت از انباره‌داده‌ها دریافت کرده و در اختیار کاربر می‌گذارد. در واقع به‌جز ذخیره و بازیابی اطلاعات هیچ عملیات دیگری بر روی انباره‌داده‌ها امکان‌پذیر نیست.

همان‌طور که اشاره شد بازارچه داده‌ها یک سیستم تخصصی است که کلیه داده‌های مورد نیاز یک بخش یا یک برنامه کاربردی را فراهم می‌کند. بازارچه داده‌ها معمولاً در سیستمهای گزارش‌دهی مورد استفاده قرار می‌گیرند. این قبیل بازارچه داده‌ها معمولاً از فنآوری *OLAP* استفاده می‌کنند. نیازی نیست که تمامی اطلاعات بازارچه داده مستقیماً از مخزن مرکزی آمده باشند، درواقع مخزن مرکزی یکی از منابع داده‌ای بازارچه داده‌ها می‌باشد.

فراداده‌ها، داده‌های مربوط به داده‌ها هستند. فراداده‌ها وضعیت‌های مختلف داده‌ها را توصیف می‌کنند. مثالهایی ساده از توصیف فراداده‌ها عبارتند از:

- اطلاعات ساختاری داده‌ها چگونه ذخیره و سازماندهی شده‌اند.
- اطلاعات متریک: مقدار داده‌ها و نحوه توزیع آنها چگونه است.
- اطلاعات تجاری: داده‌ها چگونه استفاده می‌شوند.

این مخزن می‌تواند به‌عنوان یکی از مؤلفه‌های بانک اطلاعاتی تلقی شود. در واقع فراداده، ابزاری را در اختیار کاربران نهایی قرار می‌دهد تا به راحتی در انباره‌داده به جستجو بپردازند.

۶-۳- انواع انباره‌داده

از نقطه نظر معماری سه مدل انباره‌داده وجود دارد: انبارهٔ بنگاه[۴]، بازارچه داده و انباره مجازی[۱].

---

[۱]- Load

[۲]- Transform

[۳]- Data Mart

[۴]- Enterprise Warehouse

- **انبارهٔ بنگاه اقتصادی**: این مدل کلیه اطلاعات درباره موضوعات معین داخل سازمان را گردآوری می‌کند. معمولاً از یک یا چند سیستم عملیاتی و یا ارائه کنندگان اطلاعات، داده‌ها فراهم می‌شوند. این مدل شامل داده‌های کلی و داده‌های جزء می‌باشد و می‌تواند در اندازه‌های کوچکی از گیگابایت تا هزاران گیگابایت، ترابایت و یا بیشتر مرتب شود. یک انباره‌داده بنگاه می‌تواند تحت سیستمهای مین‌فریم[۲] سنتی، ابرسرورهای *unix* یا بسترهای معماری موازی پیاده‌سازی شود. این انباره نیازمند مدلسازی کسب و کار گسترده می‌باشد که طراحی و ساخت آن ممکن است سالهای زیادی به طول انجامد.
- **بازارچه داده**: شامل زیرمجموعه‌ای از داده‌های سازمانهای گسترده است که شامل داده‌های مرتبط با گروه ویژه‌ای از کاربران می‌باشد. به‌عنوان مثال یک بازارچه داده بازاریابی می‌تواند به موضوعاتی مانند مشتری، اقلام جنس و فروش محدود شود. داده‌های موجود در بازارچه داده تمایل به خلاصه شدن دارند. بازارچه‌های داده اغلب تحت سرورهای اداری ارزان قیمت که بر مبنای *windows/NT ،unix* یا *OS/۲* هستند، پیاده‌سازی می‌شوند. چرخه پیاده‌سازی بازارچه‌های داده بیشتر در مقیاس هفته برآورد می‌شود. بازارچه‌های داده بر اساس منبع داده به دو دسته زیر تقسیم می‌شوند:
  - **بازارچه‌های داده مستقل**: این نوع بازارچه‌های داده مرتبط با بیش از یک سیستم عملیاتی بوده و یا مرتبط با داده‌هایی هستند که به‌طور محلی درون یک بخش خاص یا محدوده جغرافیایی خاص تولید شده‌اند.
  - **بازارچه‌های داده وابسته (غیر مستقل)**: به‌طور مستقیم از انباره بنگاه استخراج می‌شوند.
- **انباره مجازی**: این انباره مجموعه‌ای از برشها بر اساس دیدگاهها[۳]ی مختلف بر روی پایگاههای داده عملیاتی می‌باشد. برای پردازش یک پرس‌وجوی مؤثر فقط برخی از

۱- Virtual Warehouse

۲- Main Frame

۳- View

دیدگاههای خلاصه به‌کار می‌رود. ساخت یک انباره مجازی آسان است اما نیازمند ظرفیت اضافه در سرورهای مربوطه می‌باشد.

۶-۴- انباره‌داده و سیستم‌های عملیاتی

انباره‌داده مخزنی از داده‌های یک بنگاه است که اغلب برای تحقیق و پشتیبانی تصمیمات از آن استفاده می‌شود. این انباره در مقابل سیستم عملیاتی سازمان[1] که با تراکنشهای روزمره سازمان سروکار دارد (مثلاً *OLTP*) قرار می‌گیرد. از آنجایی‌که سیستمهای عملیاتی اغلب به‌صورت مستقل طراحی می‌شوند، برای پردازش یکپارچه داده‌ها به مخزنی خاص نیاز دارند که کلیه داده‌های مرتبط را یکجا دربرگیرد.

مزیت دیگر انباره‌داده این است که فعالیتهای تحلیلی مانند *OLAP* را از سیستم عملیاتی جدا می‌کند. از انباره‌داده‌ها می‌توان برای داده‌کاوی، مصورسازی داده‌ها[2]، ارائه گزارشات پیشرفته و به‌کارگیری انواع ابزارهای *OLAP* استفاده کرد. همچنین اطلاعات از منبع اصلی مستقل می‌شوند که این مسئله در زمانی که اطلاعات اجرایی در حال تغییر هستند بسیار اهمیت دارد. علاوه بر این اگر ساختار ذخیره اطلاعات جهت یک نوع عملیات خاص، طراحی شده باشد (مثلاً ساختار ستاره‌ای برای عملیات فروش) جستجو در ابعاد مختلف نیز بسیار ساده‌تر خواهد شد.

با توجه به آنچه گفته شد می‌توان در موارد زیر انباره‌داده‌ها را از سیستمهای عملیاتی متمایز نمود:

- اهداف
- ساختار
- اندازه
- بهینه بودن عملکرد
- فنآوری‌های استفاده شده

---

[1]- Operational System

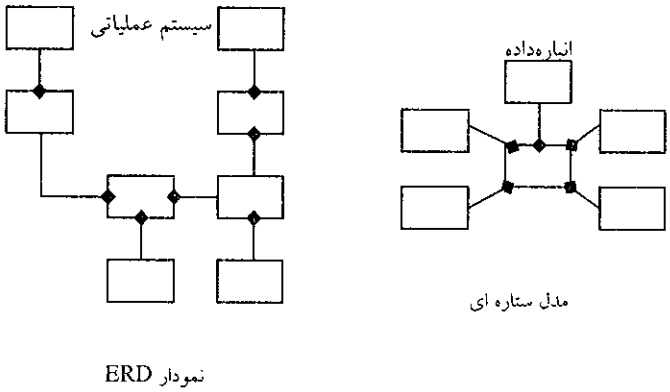
[2]- Data Visualization

جدول (٦-٢) این تفاوت‌ها را به تفصیل نشان می‌دهد:

جدول ٦-٢) مقایسه سیستمهای عملیاتی و انباره‌داده‌ها

سیستمهای عملیاتی	انباره‌داده‌ها
بر مبنای کاربرد	بر مبنای موضوع
کوچک *(several MB up to GB)*	بزرگ *(hundreds of GB up to TB)*
داده‌های جاری	داده‌های تاریخی
جداول نرمال	جداول غیرنرمال
به روز شدن همیشگی	به روز شدن دسته‌ای
جستجوهای ساده	جستجوهای پیچیده

ساختار اغلب انباره‌ها به‌صورت ستاره‌ای طراحی می‌شود زیرا اولاً موضوع‌گرا هستند و ثانیاً از آنجا که هر نوع ارتباطی بین موجودیتها ممکن است مورد مطالعه قرارگیرد، تا حد امکان لازم است موجودیتها بدون واسطه مرتبط شده باشند.



شکل ٦-٨) تفاوت ساختارها

شکل (٦-٨) تفاوت در ساختار دو نوع سیستم را نشان می‌دهد. عامل مهم دیگری که در انباره‌داده‌ها باید به آن توجه شود نوع و ماهیت داده‌ها است چرا که داده‌ها باید دقیق[1]، سازگار[1]،

[1]- Accurate

به‌موقع[۲]، یکپارچه، کامل[۳]، معتبر[۴] بوده و در راستای قواعد کسب و کار باشند و به‌علاوه باید به خوبی قابل درک[۵] باشند.

۶-۴-۱- کاربران نهایی انباره‌داده‌ها

کاربران نهایی در واقع آخرین و مهم‌ترین مؤلفه در هر انباره‌داده می‌باشند. این کاربران نهایی تحلیل‌گران، توسعه دهندگان برنامه‌های کاربردی و کاربران کسب و کار می‌باشند.

تحلیل‌گران

تحلیل‌گران نیاز دارند که به غالب داده‌ها به منظور استخراج مدلهای مختلف و تهیه گزارشها دسترسی داشته باشند. آنها از یکسری ابزارهای خاص از جمله بسته‌های آماری، ابزارهای داده‌کاوی و صفحات گسترده استفاده می‌کنند. معمولاً تحلیل‌گران به عنوان نخستین ذینفعان انباره‌های داده محسوب می‌شوند. تعداد افراد خبره‌ای که در این دسته قرار می‌گیرند بسیار کم است. کاری که آنها انجام می‌دهند از درجه اهمیت بسیار بالایی برخوردار بوده و بسیار پیچیده است. یک انباره‌داده، داده‌های پاکسازی شده را به‌طور یکجا جمع‌آوری می‌کند. این داده‌ها باید ویژگیهای زیر را دارا باشند تا بتوانند به راحتی مشکلات تحلیل‌گران را حل کنند:

- داده‌های سراسر بانک اطلاعاتی باید سازگار باشند.
- داده‌ها باید با زمان سازگار باشند.
- یک سیستم باید بتواند به پایین‌ترین سطح اطلاعات و تراکنشها دسترسی داشته باشد.

توسعه دهندگان برنامه‌های کاربردی

انباره‌های داده معمولاً طیف گسترده‌ای از برنامه‌های کاربردی را پشتیبانی می‌کنند. به‌منظور توسعه یک برنامه کاربردی پایدار انباره‌های داده نقش به‌سزایی دارند. اول اینکه برنامه‌ای را که آنها توسعه می‌دهند باید در برابر تغییرات در ساختار انباره‌داده ایمن باشد. ایجاد جداول جدید،

۱- Consistent
۲- Timely
۳- Complete
۴- Valid
۵- Well Understood

فیلدهای جدید و تغییرات ساختاری جداول باید حداقل تأثیر را بر روی برنامه‌های کاربردی موجود داشته باشد. وجود یکسری مشخصه ویژه بر روی داده‌ها به تحقق این امر کمک می‌کند.

علاوه بر آن داشتن دانشی درباره اینکه هر برنامه کاربردی از چه فیلدهایی استفاده می‌کند، می‌تواند مانع بن بست[1] شود. توسعه دهندگان سیستم نیاز دارند بدانند ارزش معتبر فیلدها چیست و علاوه بر آن ارزش هر فیلد به چه معناست. پاسخ این سؤالات هدف فراداده می‌باشد. فراداده مستنداتی را در ارتباط با ساختار داده ارائه می‌کند. از آنجا که کسب و کار واقعی نیازمند توسعه برنامه‌های کاربردی می‌باشد، درک نیاز توسعه‌دهندگان و دسترسی آنها به انباره‌داده‌ها از درجه اهمیت بالایی برخوردار است. انباره‌های داده به مرور دچار تغییر می‌شوند و برنامه‌های کاربردی نیز همچنان از آنها استفاده می‌کنند. این امر مهم‌ترین عامل موفقیت کنترل و مدیریت تغییرات می‌باشد.

کاربران کسب و کار

کاربران کسب و کار نیز از جمله استفاده‌کنندگان انباره‌داده می‌باشند. نیازمندیهای آنها موجب توسعه برنامه‌های کاربردی و معماری انباره‌داده می‌شود. در برخی از کسب و کارها کاربران نهایی تنها با گزارشهای تهیه شده از انباره‌های داده‌ها، یا صفحات گسترده سروکار دارند. اینکه افراد بر روی میز خود کامپیوتر داشته و قادر باشند به‌طور مستقیم به انباره‌داده دسترسی داشته باشند، از درجه اهمیت بالایی برخوردار است. کاربران با استفاده از ابزارهای موجود می‌توانند گزارشهای ترسیمی و تحلیلی بسیار جالب و با ارزشی متناسب با نیاز خود از انباره‌داده استخراج کنند. از طرف دیگر آنها همچنین قادر خواهند بود تا به درون مخزن انباره‌داده وارد شده و تا پایین‌ترین سطح، داده‌های موجود را بررسی کنند.

کاربردهای انباره‌داده

سه کاربرد مهم برای انباره‌داده شناخته شده است: یکی از این کاربردها، داده‌کاوی می‌باشد که این ارتباط قبلاً به تفصیل توضیح داده شد. کاربردهای دیگر انباره‌داده در پردازش اطلاعات و

---

[1]- Gridlock

پردازشهای تحلیلی می‌باشد. به‌عنوان مثال پرس‌وجوها و تحلیلهای آماری و ایجاد جداول، نمودارها و گزارشهای گرافیکی با استفاده از پردازش اطلاعات در انباره‌داده‌ها قابل ارائه می‌باشد.

انباره‌داده‌ها به تحلیل‌گران کسب و کار کمکهای شایانی می‌کند:

- داشتن انباره‌داده، یک مزیت رقابتی است چرا که با دسترسی سریع و به‌موقع به داده‌ها به غلبه بر رقبا کمک می‌کند.
- یک انباره‌داده می‌تواند بهره‌وری کسب و کار را توسعه دهد. چرا که قادر است اطلاعات سازمان را خیلی دقیق، با سرعت و به‌طور کارا تشریح کند.
- یک انباره‌داده بازاریابی، ارتباط با مشتریان را تسهیل می‌کند، چرا که داده‌های مرتبط با مشتریان و اقلام مختلف در کلیه بخش‌ها و کلیه فروشگاهها را در اختیار قرار می‌دهد. همچنین انباره‌داده‌ها می‌توانند توسط ردیابی روندها، الگوها و استثنائات در دوره‌های زمانی طولانی با یک روش سازگار و قابل اطمینان به کاهش هزینه‌ها کمک کنند.

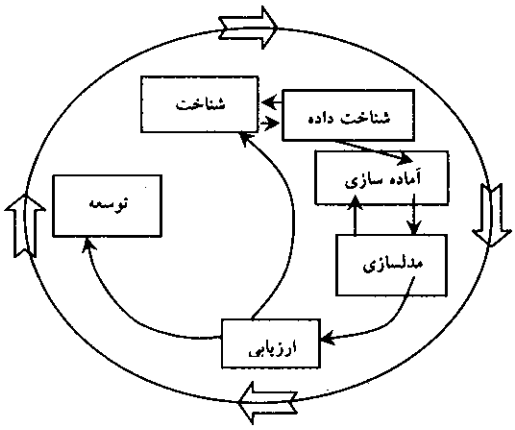
منابع

1) Han J. and Kamber M. (2006) Data Mining: Concepts and Techniques San Francisco ، CA: Morgan Kaufmann.

فصل هفتم

متدلوژی اجرا و پیاده‌سازی پروژه‌های داده‌کاوی

روشهای مختلفی برای پیاده‌سازی و اجرای پروژه‌های داده‌کاوی وجود دارد. یکی از روشهای بسیار قوی، متدلوژی *CRISP*[1] می‌باشد. این متدلوژی از گامهای شناخت سیستم، شناخت داده‌ها، آماده‌سازی داده‌ها، مدلسازی، ارزیابی و توسعه سیستم تشکیل شده است. هر کدام از این گامها به زیر بخشهایی تقسیم می‌شوند.

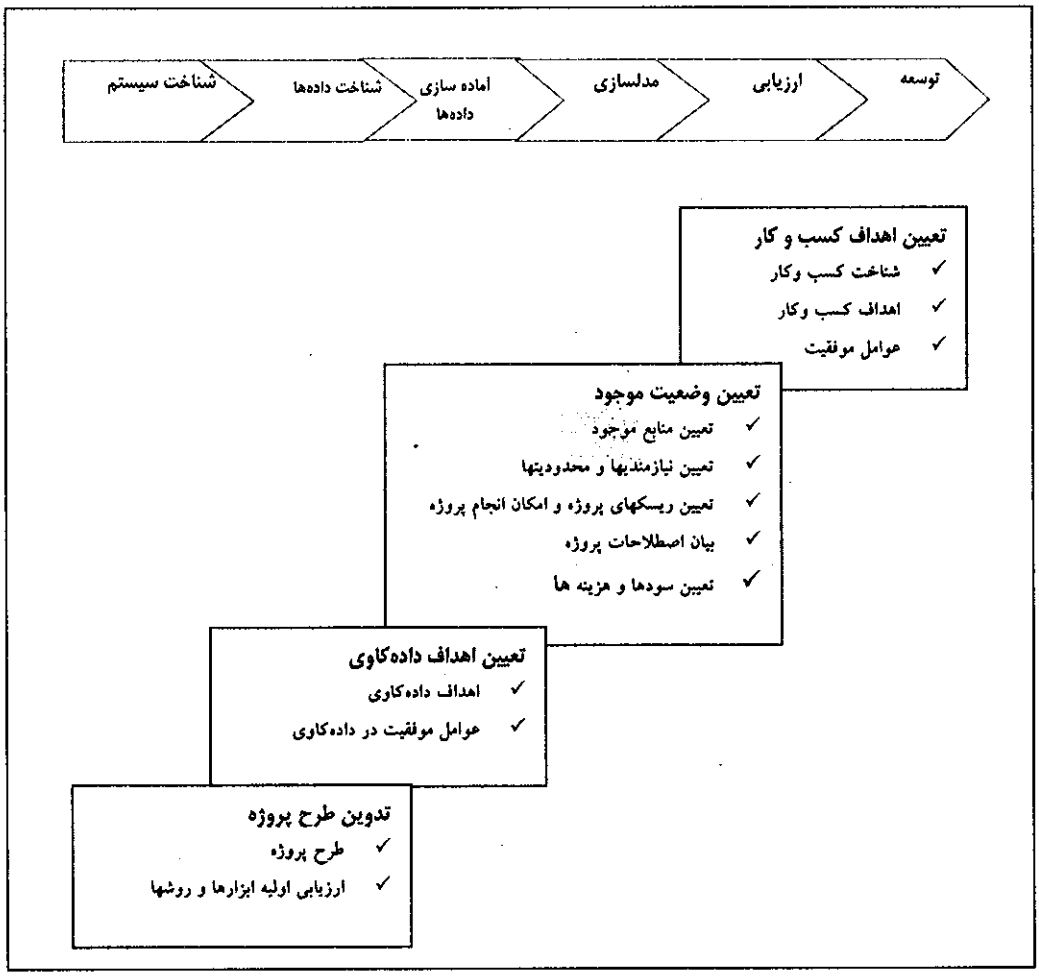


شکل ۷-۱) گامهای متدولوژی CRISP

[1]- Cross Industry Standard Process for Data Mining

۷-۱- گام شناخت سیستم

درگام شناخت سیستم ابتدا به شناخت کسب وکار مورد نظر پرداخته می‌شود. سپس اهداف مورد نظر و عوامل موفقیت کلیدی آن تعیین شده و دوباره اهداف کسب و کار بازنگری می‌شود. شناسایی فرصتها و عوامل موفقیت کلیدی یک کسب وکار قدم بسیار مهمی می‌باشد، چرا که باعث افزایش اطلاعات و انجام بهتر کارها می‌شود. در واقع در این گام به تعیین زمینه‌ها و عواملی می‌پردازیم که داده‌ها باعث افزایش ارزش در آنها می‌شوند.



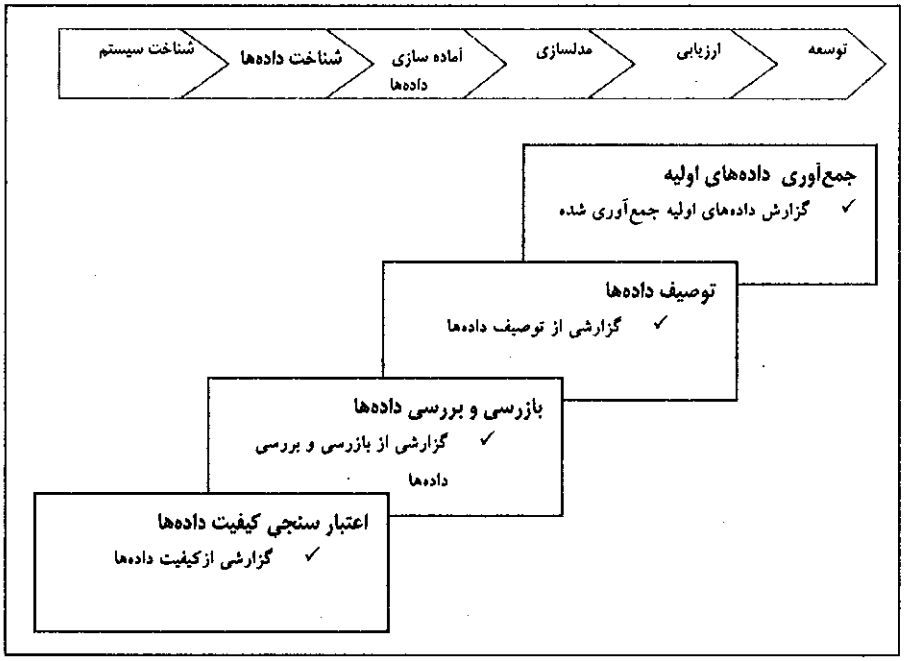
شکل ۷-۲) جزئیات مربوط به گام شناخت سیستم

پس از تدوین اهداف مورد نظر کسب وکار، می‌بایست به شناخت وضعیت موجود پرداخت. به‌منظور تعیین دقیق وضعیت موجود، به تعیین منابع موجود پرداخته و نیازمندیها و محدودیتهای موجود تعیین می‌شوند. شناخت ریسکهای بر سر راه پروژه و نیازمندیهای پروژه کمک می‌کنند تا طرح امکان‌سنجی پروژه تدوین شود. در این قدم به‌منظور تدوین طرح امکان‌سنجی، تعیین منابع پروژه ضروری می‌باشد. منابع پروژه عبارتند از: منابع انسانی، منابع مالی، تجهیزات و دیگر منابع. به همین دلیل در این گام تعیین سودها و هزینه‌های پروژه امری اجتناب‌ناپذیر می‌باشد. تعیین اهداف داده‌کاوی و تدوین طرح پروژه نیز از دیگر بخشهایی است که در این گام باید به آنها پرداخت. به عنوان مثال در یک کسب و کار خاص اهداف زیر برای پروژه داده‌کاوی تعیین می‌شود:

- برنامه‌ریزی بازاریابی محصولات و خدمات جدید
- قیمت‌گذاری محصولات و خدمات جدید
- شناخت مشتریان ناراضی و جلوگیری از رویگردانی آنها
- تعیین بازار هدف

۷-۲- گام شناخت داده‌ها

پس از شناخت کسب و کار به سراغ شناخت داده‌ها می‌رویم. شناخت داده‌ها عبارت است از جمع‌آوری داده‌های اولیه، توصیف داده‌ها، بازرسی و بررسی داده‌ها و اعتبارسنجی کیفیت داده‌ها. کارآیی داده‌کاوی مستقیماً مرتبط با داده‌های مورد استفاده دارد. هر اندازه داده‌ها دقیق‌تر جامع‌تر و با کیفیت بهتری باشند خروجی داده‌کاوی کارآتر خواهد بود. بنابراین انتخاب وجمع‌آوری داده‌های درست، توصیف آنها، یکپارچه‌سازی قالب آنها به‌منظور استفاده در داده‌کاوی، از اهمیت بسیار بالایی برخوردار می‌باشد. علاوه بر این بازرسی و بررسی داده‌ها به‌منظور تعیین میزان کیفیت آنها بسیار مهم می‌باشد. در شکل (۷-۳) به این گامها و خروجیهای هر مرحله اشاره شده است. پروژه داده‌کاوی باید در راستای اهداف داده‌کاوی صورت گیرد. بنابراین جمع‌آوری و یکپارچه کردن داده‌های مرتبط با این اهداف در برآورده کردن اهداف و شاخصهای مرتبط با آن بسیار مهم و حیاتی است.

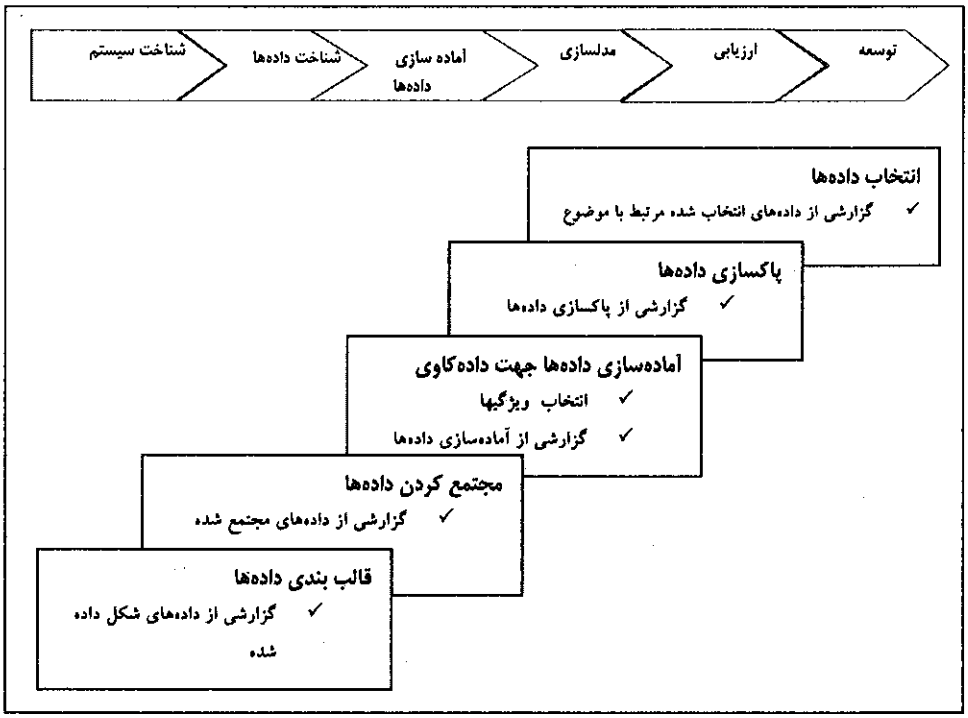


شکل ۷-۳) جزئیات مربوط به گام شناخت داده‌ها

۷-۳- گام آماده سازی داده‌ها

گام آماده‌سازی داده‌ها عبارت است از: انتخاب داده‌ها، پاکسازی داده‌ها، آماده‌کردن داده‌ها جهت داده‌کاوی، مجتمع کردن آنها و قالب بندی داده‌ها. برای اجرای هرکدام از این زیر بخشها، فعالیتهای دیگری نیز ضروری است که در شکل (۷-٤) آمده است.

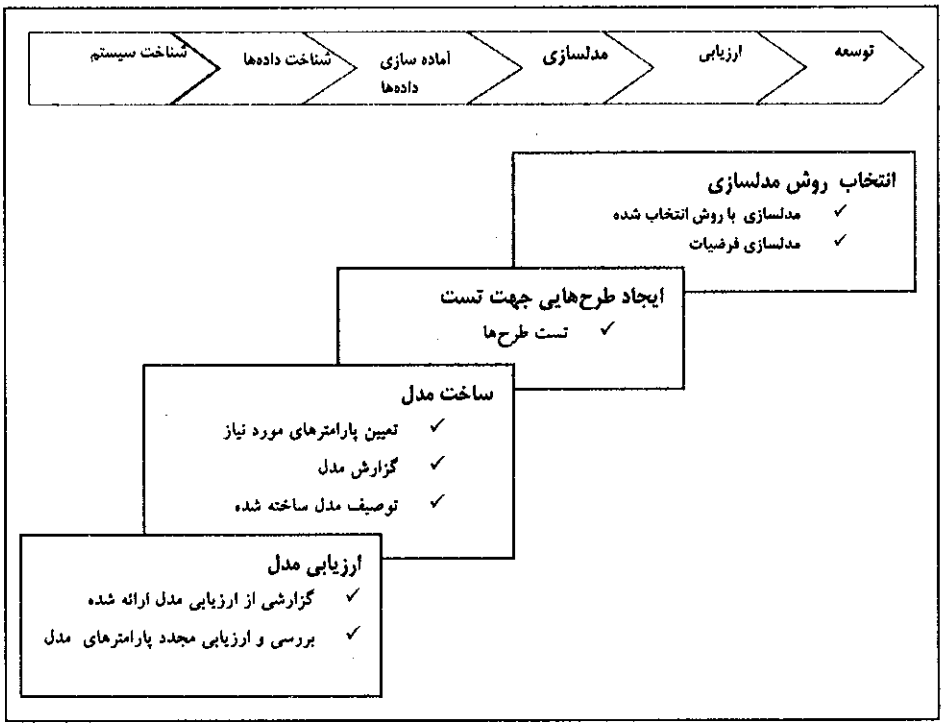
جمع‌آوری و محافظت از داده‌ها گام بسیار مهمی می‌باشد. اصولاً چون قالب و نوع داده‌ها در طول زمان تغییر می‌کند، ممکن است قالب بسیاری از داده‌های موجود متفاوت باشد. همچنین به‌علت اینکه داده‌ها از منابع مختلف داخلی و خارجی جمع‌آوری شده و یکپارچه می‌شوند، باز هم ممکن است قالب داده‌ها با هم یکسان نبوده و یا حتی برخی از داده‌های قبلی از بین رفته و دور ریخته شده باشند و بخشهایی از داده‌ها موجود باشد. در داده‌کاوی اهمیت داده‌های قدیمی به هیچ وجه کمتر از داده‌های جدید نمی‌باشد.



شکل ۷-٤) جزئیات مربوط به گام آماده‌سازی داده‌ها

۷-۴- گام مدلسازی

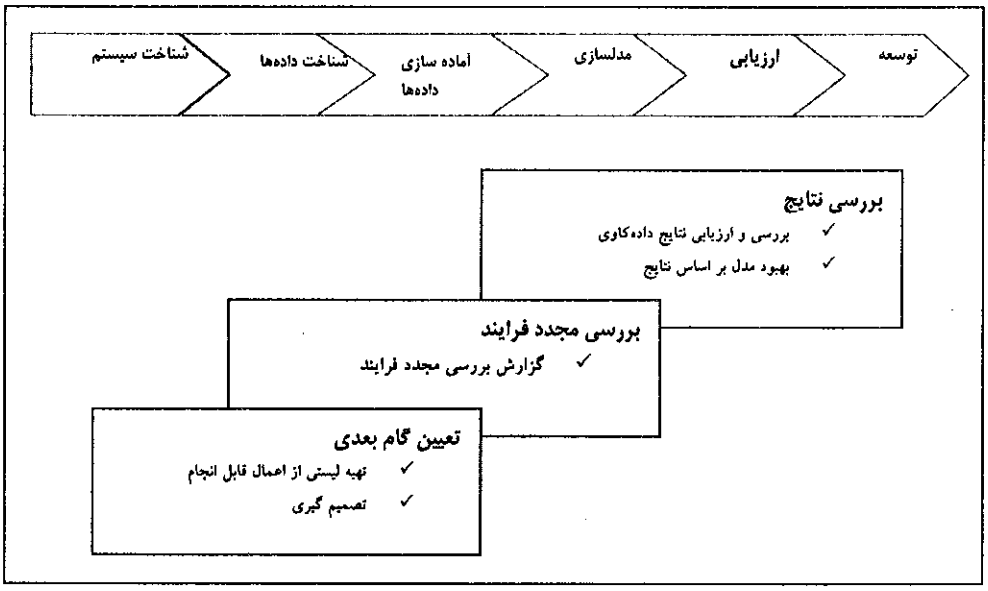
پس از شناخت داده‌ها و آماده‌سازی آنها، حال می‌توان به مدلسازی پرداخت. در اولین قدم از مدلسازی می‌بایست روش مناسب را انتخاب کرد. انتخاب روش مناسب بسیار تعیین کننده می‌باشد. پارامترهای مورد نیاز مدل نیز پس از تعیین روش مورد استفاده، مشخص می‌شوند. پس از انتخاب مدل و تعیین پارامترها، بخشهای کوچکی از پروژه تعریف شده و پس از اجرا شدن، در هر مرحله به دقت تست می‌شوند تا کیفیت مدل ایجاد شده تضمین شود. در این مرحله اگر مدل مورد نظر دقت لازم را نداشت و یا کیفیت مطلوب را حاصل نکرد، ابتدا به تغییر پارامترهای مدل می‌پردازیم و مجدداً مدل را تست می‌کنیم. اگر هنوز کیفیت لازم را کسب نکرده بود، مدل را تغییر داده و مدل جدیدی می‌سازیم. برای انجام هر کدام از زیر بخشهای مدلسازی، فعالیتهای دیگری نیز ضروری است که در شکل (۷-۵) آمده است.



شکل ۷-۵) جزئیات مربوط به گام مدلسازی

۷-۵- گام ارزیابی

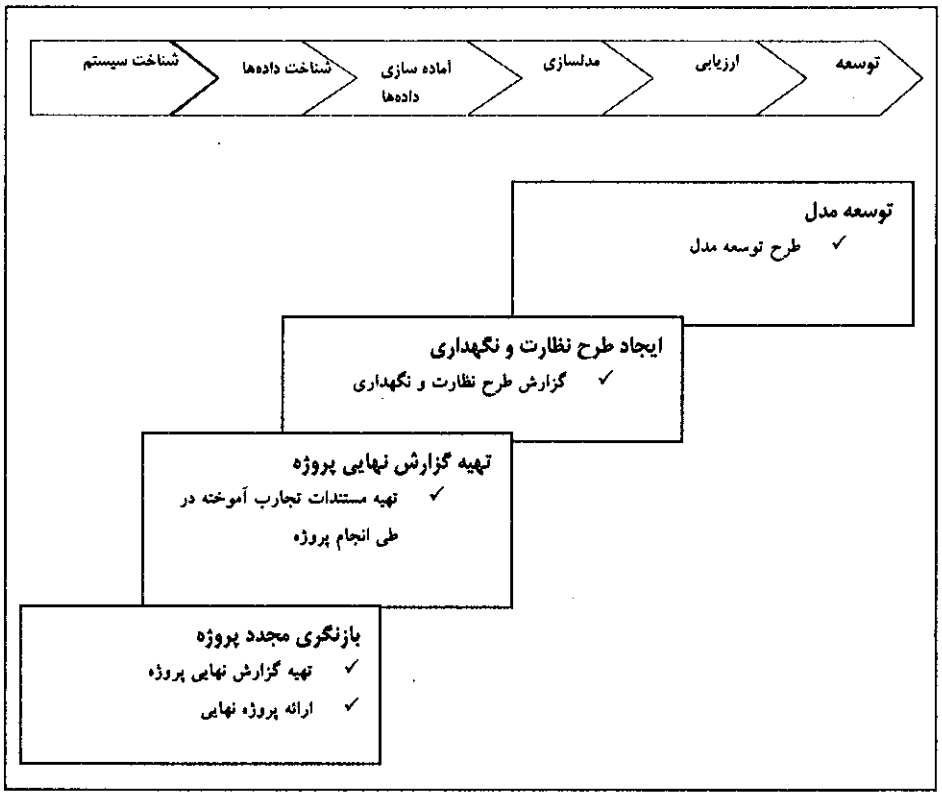
پس از مدلسازی، حال می‌بایست به ارزیابی نتایج حاصل از مدل پرداخت. نتایج ارزیابی باعث بهبود مدل شده و مدل را قابل استفاده می‌کند. در این گام اعتبار مدل بررسی شده وگزارشی از کل فرایند تهیه می‌شود. در انتها نیز لیستی از اقدامات اصلاحی قابل انجام تهیه شده و به‌عنوان راهکار ارائه شده و تصمیم‌گیریها بر این اساس انجام می‌شود.



شکل ۷-۶) جزئیات مربوط به گام ارزیابی

۷-۶- گام توسعه

پس از استخراج لیست اقدامات قابل انجام، دورنمایی از طرح توسعه ایجاد می‌شود. در این گام این طرح مجددا بررسی شده و مدون می‌شود. علاوه بر آن طرح نظارت و نگهداری پس از اتمام پروژه نیز در این گام تهیه می‌شود. در پروژه‌های داده‌کاوی تأکید زیادی بر روی مستندسازی تجارب آموخته در طی انجام پروژه وجود دارد و در این گام گزارش مرتبط با آن نهایی می‌شود. در این مرحله پروژه قابل ارائه نهایی بوده و گزارش نهایی آن نیز استخراج شده است.



شکل ۷-۷) جزئیات مربوط به گام توسعه

منابع

راهنمای نرم افزار SPSS Clementine

بخش چهارم

مباحث ویژه در داده کاوی

فصل هشتم: سریهای زمانی در داده‌کاوی

فصل نهم: شبکه‌های اجتماعی

فصل دهم: کاربرد داده‌کاوی در مدیریت ارتباط با مشتری

فصل هشتم

سریهای زمانی در داده‌کاوی

یک سری زمانی دنباله‌ای از مشاهدات بر روی یک متغیر مورد توجه است که در نقاط گسسته‌ای از زمان که معمولاً فاصله‌های مساوی دارند (روزانه – هفتگی – ماهانه – فصلی – سالانه)، رخ می‌دهد. تجزیه و تحلیل سریهای زمانی، متضمن توصیف فرآیند یا پدیده‌ای است که تولید دنباله می‌کند. جهت پیش‌بینی سریهای زمانی، لازم است که رفتار فرآیند را با یک مدل ریاضی که قابل تعمیم به آینده باشد، توصیف کرد. معمولاً لازم نیست مدل نماینده مشاهدات خیلی قدیمی یا فراتر از زمان مورد انتظار پیش‌بینی باشد[۱]. مهم‌ترین نکته در داده‌های سری زمانی، آن است که این داده‌ها دارای همبستگی هستند. از ضریب همبستگی برای تعیین همبستگی بین مقادیر X و Y استفاده می‌شود، اما وقتی خود متغیرهای مستقل، مقادیرشان به هم مرتبط باشد، به آن خودهمبستگی[1] گویند. با توجه به تعریف داده‌های سری زمانی، روشهای آماری مبتنی بر فرض مستقل بودن مشاهدات، مناسب نبوده و به جای آن می‌توان از معادلات خودهمبستگی در تحلیل سریهای زمانی استفاده کرد. [۲]

با توجه به اهمیت و نقش سریهای زمانی، در حوزه‌های مختلفی از آنها استفاده می‌شود که نمونه‌هایی از کاربردهای آن به شرح زیر است. [۲]

- بازرگانی و اقتصاد: مقادیر فروش و قیمتهای ماهیانه، قیمت سهام در روزهای مختلف

---

[1]- Auto Correlation

- علوم مهندسی: ثبت علائم الکتریکی، ولتاژ و سیگنالها.
- پزشکی: داده‌های الکتروکاردیوگرام و دیگر کاربردها.
- هواشناسی: درجه حرارت روزانه و میزان بارندگی سالیانه.
- کنترل کیفیت: ثبت مشاهدات مربوط به فرایند در نمودارهای کنترل.
- علوم اجتماعی: نرخهای زاد و ولد و مرگ و میر سالیانه.

۸-۱- داده‌کاوی سریهای زمانی[۱]

یک سری زمانی ساده‌ترین شکل داده‌های زمانی است. سری زمانی دنباله‌ای از اعداد حقیقی است که به‌صورت منظم در طول زمان گردآوری شده است. هر عدد، نشان‌دهنده مقدار یک متغیر مشاهده شده می‌باشد. همان‌طور که اشاره شد، داده‌های سری زمانی در حوزه‌های مختلفی مثل تحلیل بازار سهام، علوم ارتباطات، پزشکی، داده‌های مالی و غیره مطرح می‌شوند. همچنین داده‌های وب که میزان استفاده از وب سایتهای مختلف را ثبت می‌کنند (برای مثال تعداد کلیکها) را می‌توان با سریهای زمانی مدل کرد. در حقیقت، سریهای زمانی برای نمایش بخش بزرگی از داده‌های ذخیره شده در بانکهای اطلاعاتی تجاری به‌کار می‌رود که به تدریج به‌عنوان یک نوع داده متفاوت، اهمیت بیشتری یافته است. اهمیت داده‌های سریهای زمانی موجب تحقیقات زیادی در زمینه تحلیل این نوع داده‌ها شده است. ادبیات آماری در مورد سریهای زمانی بسیار وسیع است و به‌طور عمده به مسائلی مانند شناسایی الگوها و تحلیل روند (مانند رشد خطی فروش شرکت در طول یک سال)، تحلیلهای فصلی (مثلاً فروش زمستانی یک محصول تقریباً دو برابر فروش تابستانی است) و پیش‌بینی (مانند پیش‌بینی فروش فصل آینده) می‌پردازند. این موضوعات کلاسیک در تعدادی از مراجع آماری بررسی شده است. [۵]

هدف اصلی داده‌کاوی سریهای زمانی کشف الگوهای موجود در وقایع و داده‌های یک سری زمانی است. کشف این الگوهای ناشناخته، همگام با استفاده از دیگر روشهای مختلف داده‌کاوی مانند سیستمهای پایگاه داده، آمار، یادگیری ماشینی، شبکه‌های عصبی، تئوری مجموعه‌ها، منطق فازی و غیره در داده‌کاوی اتفاق می‌افتد. از مهم‌ترین کاربردهای داده‌کاوی سریهای زمانی،

۱- Time Series Data Mining

می‌توان به دسته‌بندی، خوشه‌بندی و کشف قواعد از داده‌ها اشاره کرد. در داده‌کاوی سریهای زمانی دو سؤال اساسی زیر مطرح می‌شود:

- چگونه می‌توان روابط همبستگی در درون سریهای زمانی را پیدا کرد؟

چگونه می‌توان سریهای زمانی با حجم انبوهی از داده‌ها را تحلیل کرده و الگوهای منظم، روند، تغییرات تصادفی، داده‌های مغشوش و غیره را از آنها استخراج کرد؟[٦]

در ادامه به بررسی جنبه‌های مختلفی از داده‌کاوی سریهای زمانی، با تمرکز بر شناسایی اجزاء سریهای زمانی و روشهای جستجوی تشابه در سریهای زمانی پرداخته می‌شود.

۸-۱-۱- اجزاء سریهای زمانی و تحلیل آنها

یک سری زمانی شامل یک متغیر وابسته (y) می‌باشد که تابعی از زمان است. چنین تابعی به شکل یک نمودار سری زمانی نمایش داده می‌شود. در مطالعه داده‌های سریهای زمانی، همواره دو هدف اساسی زیر دنبال می‌شود:[٦]

- مدلسازی سریهای زمانی با تأکید بر فرآیند ایجاد سریهای زمانی
- پیش‌بینی سریهای زمانی با تأکید بر پیش‌بینی مقادیر متغیرهای سری زمانی

تجزیه و تحلیل روند شامل شناسایی چهار جزء یا مشخصه اساسی هر سری زمانی می‌باشد.[۲]

- **روند خطی و غیرخطی:** تغییر دراز مدت در میانگین یا به‌عبارت دیگر حرکت دراز مدت تدریجی افزایشی یا کاهشی داده‌ها در طول زمان است.
- **تغییر سیکلی:** تغییرات دوره‌ای موجود در داده‌های یک سری زمانی است. معمولاً این نوع افزایشها و کاهشهای لحظه‌ای در داده‌ها، در دوره‌های بیشتر از یکسال اتفاق می‌افتد.
- **تغییرات فصلی:** تغییراتی است که به‌صورت فصلی در داده‌های سری زمانی اتفاق می افتد. این نوع الگوی تغییر در طول یکسال مشاهده می‌شود.
- **تغییرات باقیمانده‌ها یا تصادفی:** اگر سه جزء قبلی از یک سری زمانی حذف شود، سری باقیمانده حاصل می‌شود. که ممکن است تصادفی باشد.

با توجه به توضیحات ارائه شده، می‌توان معادله یک سری زمانی را به‌صورت یکی از دو حالت (۸-۱) و (۸-۲) نشان داد. [٦]

$$y = T_t + C_t + S_t + R_t \qquad (۱-۸)$$

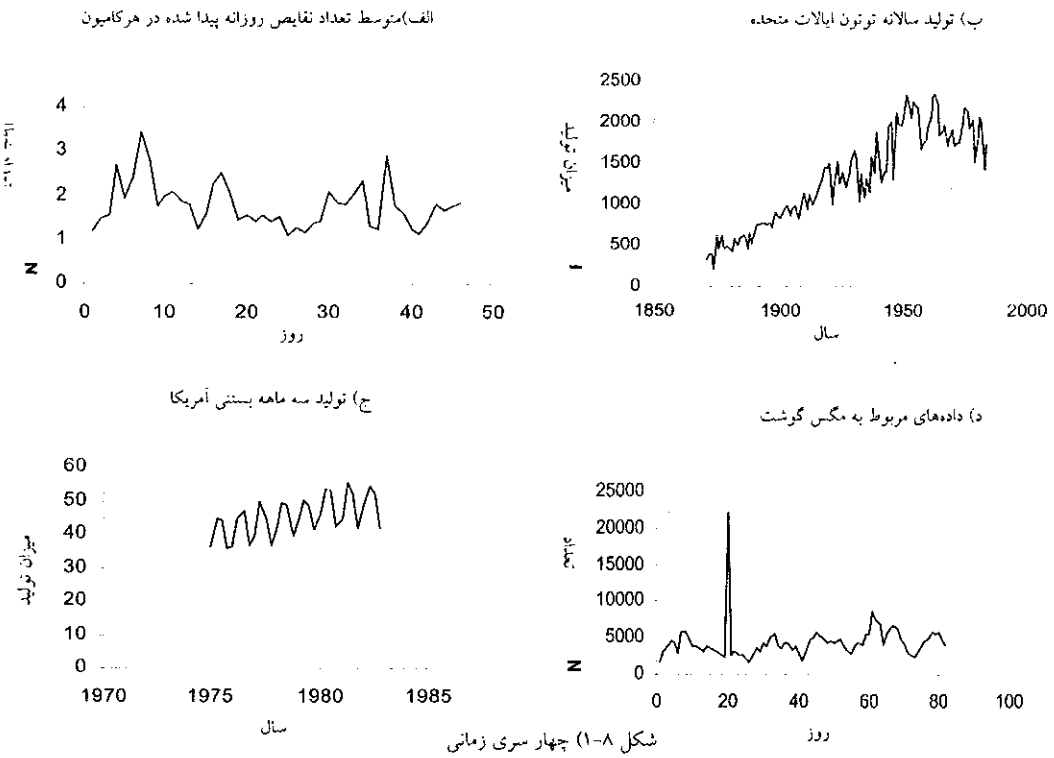
$$y = T_t * C_t * S_t * R_t \qquad (۲-۸)$$

یکی از مهم‌ترین جنبه‌های کاربرد سریهای زمانی، پیش‌بینی می‌باشد. پیش‌بینی سریهای زمانی با استفاده از یک معادله ریاضی، الگویی تاریخی در داده‌های سری زمانی ایجاد می‌کند. این روش برای پیش‌بینی کوتاه مدت یا بلند مدت مقادیر آینده مورد استفاده قرار می‌گیرد. روشهای مختلفی برای پیش‌بینی سریهای زمانی، مورد استفاده است که از بین آنها روش میانگین متحرک تلفیق‌شده با اتو رگرسیون[۱] که به مدل «باکس– جنکینز» نیز موسوم است از اهمیت ویژه‌ای برخوردار است[٦]. برای آشنایی بیشتر با روشهای پیش‌بینی سریهای زمانی، می‌توان به مراجع آمار و اقتصادسنجی مراجعه کرد.

معمولاً مطلوب است که سیستم پیش‌بینی بتواند تغییرات پایدار را مشخص و با تعدیل مدل پیش‌بینی، فرآیند جدید را تعقیب کند. در عین حال سیستم پیش‌بینی، تغییرات تصادفی و موقتی را تشخیص داده و در مقابل آنها واکنش نشان ندهد. در هنگام کار با مدلهای پیش‌بینی با توجه به مراحل مختلف سیکل پیش‌بینی (مثلاً عمر محصول)، لازم است که مدلهای پیش‌بینی مختلفی به‌کارگرفته شوند. مثلاً ممکن است گاهی اوقات فقط روند را حفظ کرده و بقیه علل تغییرات سری زمانی را حذف کنیم. [٤]

به‌طور کلی یک سری زمانی را ایستا[۲] گویند، هرگاه تغییر منظمی در میانگین و واریانس آن وجود نداشته و تغییرات دوره‌ای اکید حذف شده باشد. نظریه احتمال سریهای زمانی بیشتر با سریهای زمانی ایستا سر و کار دارد و به این دلیل است که در تجزیه و تحلیل سریهای زمانی، برای استفاده از نظریه ایستایی لازم است که سری ناایستا را به ایستا تبدیل کنیم. مثلاً می‌توانیم روند و تغییرات فصلی را از مجموعه داده‌ها حذف کرده و سپس به وسیله یک فرآیند تصادفی ایستا، تغییر در باقیمانده‌ها را الگوسازی کنیم. [۲]

۱- Arima
۲- Stationary



شکل (۸-۱) چهار سری زمانی

شکل (۸-۱) که چهار سری زمانی را نشان می‌دهد، خصوصیات اجزاء سری‌های زمانی را آشکار می‌کند. به نظر می‌رسد که متوسط تعداد نقایص روزانه پیدا شده در هرکامیون، در انتهای خط تولید کارخانه تولید کامیون که در شکل (۸-۱-الف) نشان داده شده است، در حول سطح ثابتی، نوسان می‌کند. همان‌طورکه گفته شد سریهای زمانی که این پدیده را نشان می‌دهند، ایستا در میانگین نامیده شده و حالتهای ویژه سریهای زمانی ایستا هستند.

تولید سالانه توتون ایالات متحده که در شکل (۸-۱-ب) نشان داده شده در حول سطح ثابتی تغییر نمی‌کند، بلکه در کل یک روند رو به بالا را نشان می‌دهد. علاوه بر این، واریانس این سری توتون، با اضافه شدن سطح سری، افزایش می‌یابد. سریهای زمانی که این پدیده را نشان می‌دهند، ناایستا در میانگین و واریانس گفته شده و مثالهایی از سریهای زمانی ناایستا هستند. تولید سه ماهه بستنی آمریکا در شکل (۸-۱-ج) طرح خاص دیگری را نشان می‌دهد که به‌واسطه تغییرات فصلی، طبیعتی تکراری دارد. سریهای زمانی که تغییرات فصلی را دربرمی‌گیرند، سریهای زمانی فصلی نامیده می‌شوند.

سریهای زمانی ناایستا را مانند آنهایی که در شکل‌های (۸- ۱- ب) و(۸-۱-ج) نشان داده شده‌اند، می‌توان با تبدیلات مناسبی به سری ایستا تبدیل نمود. سری زمانی چهارم که در شکل (۸-۱- د) نشان داده شده است، داده‌های مربوط به مگس گوشت است. این سری پدیده دیگری از ناایستایی به‌دلیل تغییر ساختاری ناشی از یک یا چند اغتشاش خارجی را منعکس می‌کند. این نوع ناایستایی را نمی‌توان با یک تبدیل استاندارد حذف نمود.

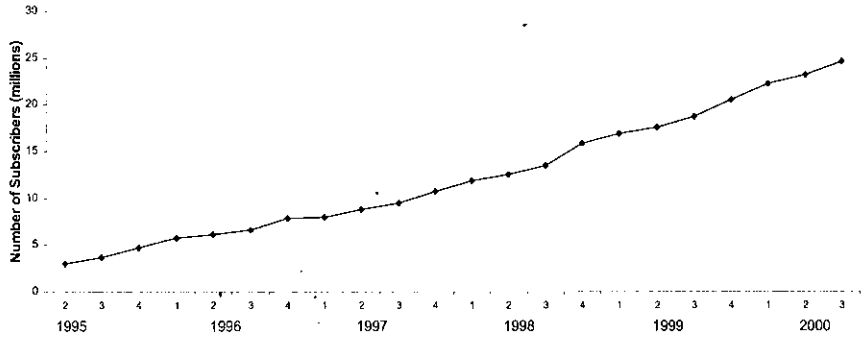
۸-۱-۲- شناسایی، تجزیه و حذف اجزاء سریهای زمانی

برای تجزیه و تحلیل مجموعه‌ای از داده‌ها، در اولین مرحله لازم است که نمودار مشاهدات را نسبت به زمان رسم کنیم. این کار غالباً مهم‌ترین خواص یک سری زمانی مانند روند، فصلی بودن و مشاهدات دورافتاده را آشکار می‌کند.

روشهای متعددی برای شناسایی، تعیین و یا حذف برخی از اجزاء سری‌های زمانی وجود دارد. با توجه به اینکه مهم‌ترین اجزاء یک سری زمانی، جزء روند و فصلی می‌باشد، در ادامه به بررسی روشهای شناسایی، هموارسازی، تجزیه و یا حذف این اجزاء پرداخته می‌شود. [۲]

۸-۱-۳- سریهای زمانی با روند خطی

اگر جزء روند در یک سری زمانی یک حالت افزایشی یا کاهشی مستقیم داشته باشد، می‌توان معادله روند را به شکل $y_t = a + bt + e_t$ نشان داد که e_t جزء تصادفی در این رابطه است. شکل (۸-۲) معادله یک روند خطی را نشان می‌دهد.



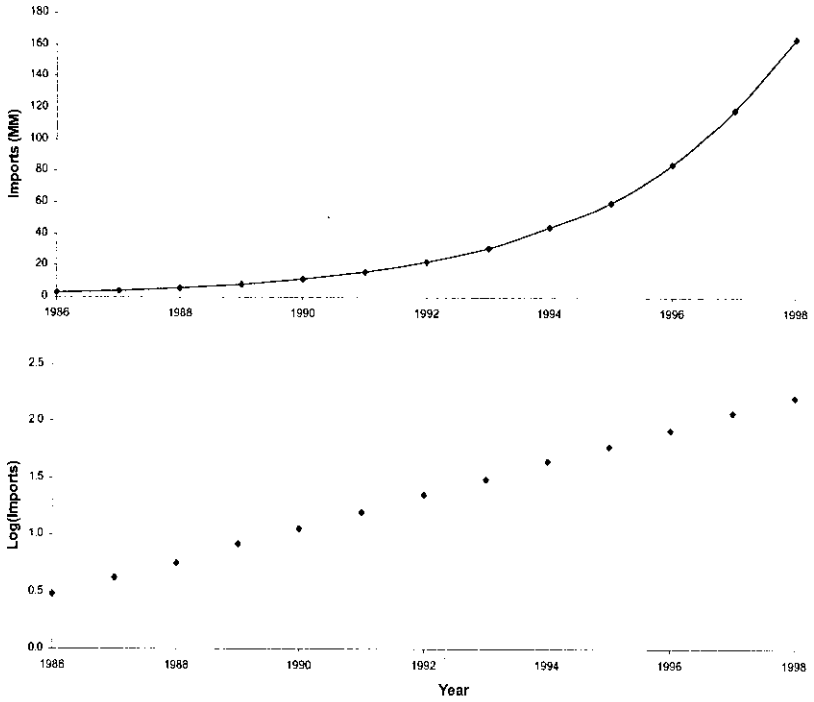
شکل ۸-۲) سری زمانی با روند خطی

سریهای زمانی با روند غیرخطی

گاهی اوقات می‌توان با در نظر گرفتن تبدیلات خاصی، مانند لگاریتم یا ریشه دوم داده‌ها، روند غیرخطی سری‌ها را به یک روند خطی تبدیل کرد. درصورتی‌که انحراف معیار با میانگین نسبت مستقیم داشته باشد، یک تبدیل لگاریتمی مطابق روابط (۸-۳) و (۸-٤) مناسب است.

$$log(y_t) = a + b + e_t \qquad (۸-۳)$$

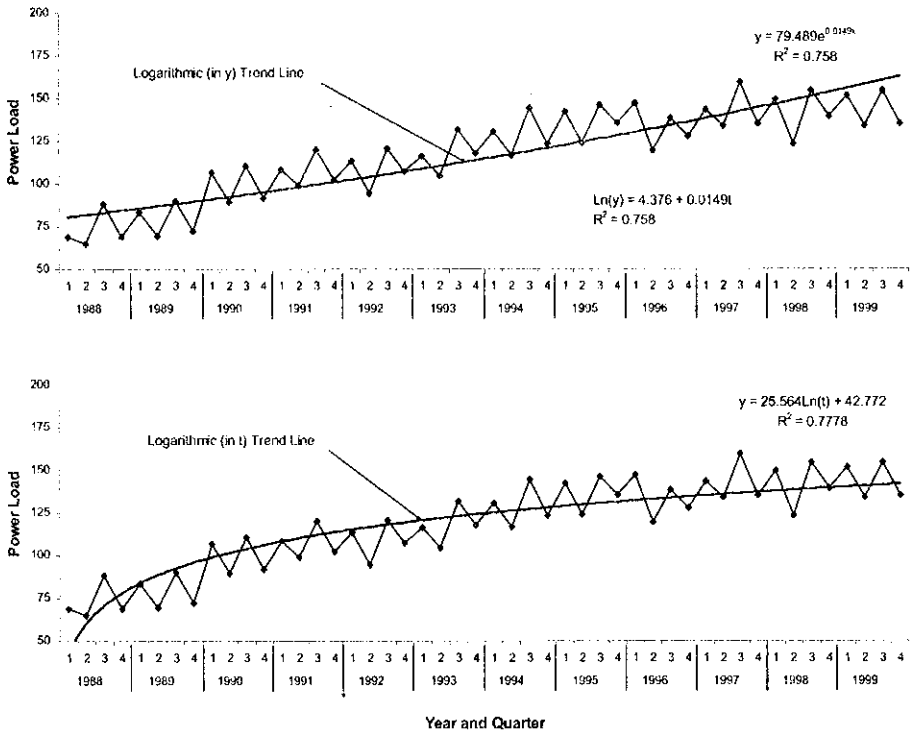
$$y_t = \exp(a + bt + e_t) \qquad (۸-٤)$$



به ترتیب با روند غیرخطی و تبدیل یافته آن شکل ۸-۳) سری زمانی

اگر یک سری زمانی با یک نرخ کاهشی بر حسب زمان ظاهرشود، ممکن است رابطه (۸-۵) مناسب باشد.

$$y_t = a + bln(t) + e_t \qquad (۸-۵)$$



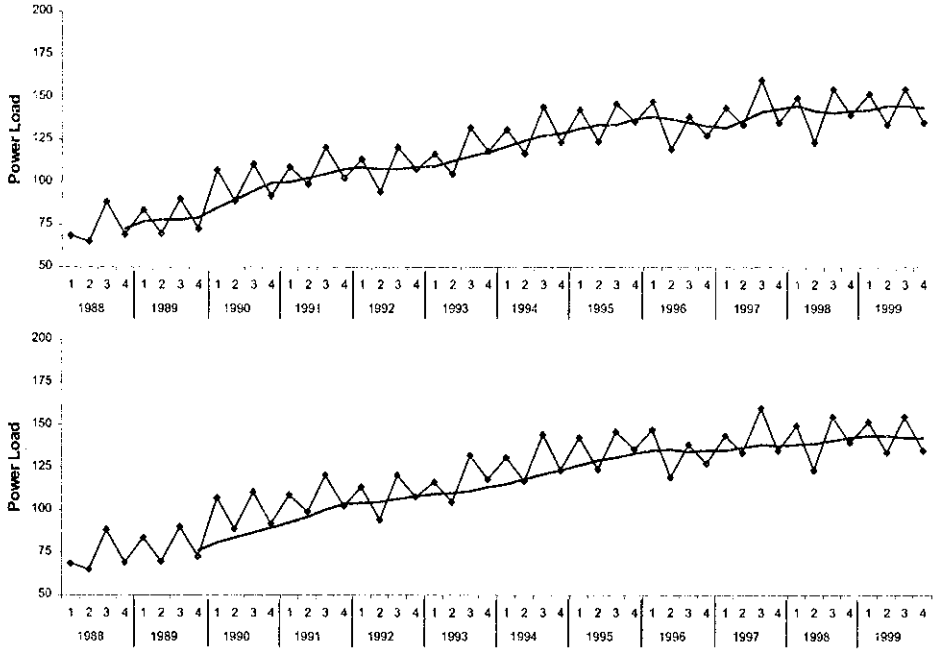
شکل ۸-٤) سری زمانی با تبدیل لگاریتمی روی متغیر وابسته (بالا) و زمان (پایین)

میانگین متحرک

یک روش دیگر برای ارزیابی روند در سری‌های زمانی محاسبه میانگین m مشاهده اخیر می‌باشد. این روش به میانگین متحرک موسوم است و مطابق رابطه (۸-٦) محاسبه می‌گردد.

$$\overline{y}_{ma(٤)} = \frac{(y_t + y_{t-١} + y_{t-٢} + y_{t-٣})}{٤} \qquad (٦-٨)$$

میانگین متحرک عمدتاً نوسانات موجود در داده‌ها را همواره می‌کند. این روش معمولاً زمانی به‌خوبی عمل می‌کند که داده‌ها یک روند خطی و الگوی منظمی از نوسانات داشته باشند.



شکل ۸-۵) روش میانگین متحرک درسری زمانی به ترتیب چهار نقطه و هشت نقطه

هموارسازی نمایی

مهم‌ترین عیب روش میانگین متحرک این است که وزن یا اهمیت داده‌های گذشته، به صورت یکسان در نظر گرفته می‌شود. برای برطرف کردن این اشکال از یک روش جامع‌تر که یک روش میانگین متحرک موزون است و در آن تخصیص وزن به دوره‌های گذشته به‌صورت یک تصاعد هندسی می‌باشد، استفاده می‌کنیم. این روش هموارسازی نمایی نام دارد و به مقادیر گذشته سری زمانی تا به حال، وزن داده می‌شود و البته وزن بیشتری برای داده‌های جدید در نظر گرفته شده و هرچه به سمت داده‌های قدیمی پیش می‌رویم، وزن آنها طی یک فرآیند نمایی، کاهش می‌یابد.

Let w=0.5

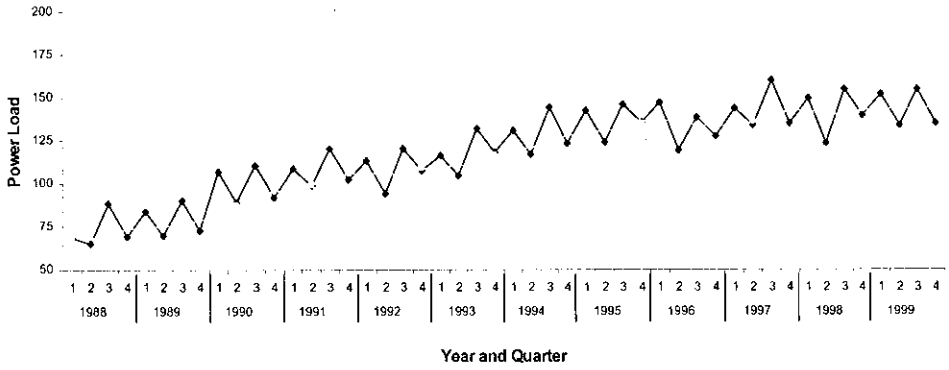
$S_1 = Y_1$
$S_2 = 0.5Y_2 + (1-0.5)S_1 = 0.5Y_2 + 0.5Y_1$
$S_3 = 0.5Y_3 + (1-0.5)S_2 = 0.5Y_3 + 0.25Y_2 + 0.25Y_1$
$S_4 = 0.5Y_4 + (1-0.5)S_3 = 0.5Y_4 + 0.25Y_3 + 0.125Y_2 + 0.125Y_1$

$S_1 = Y_1$

$S_t = wY_t + (1-w)S_{t-1}$
$= wY_t + w(1-w)Y_{t-1} + w(1-w)^2 Y_{t-2} + \cdots$

شکل ۸-٦) هموارسازی میانگین متحرک (راست) ومیانگین متحرک موزون نمایی (چپ)

هرچه مقدار ثابت هموارسازی (w) بیشتر باشد نشان دهنده آن است که داده‌های قدیمی اثر کمتری بر روی هموارسازی پیش‌بینی دارند و هر چه مقدار w کمتر باشد داده‌های سری زمانی هموارتر خواهند بود.

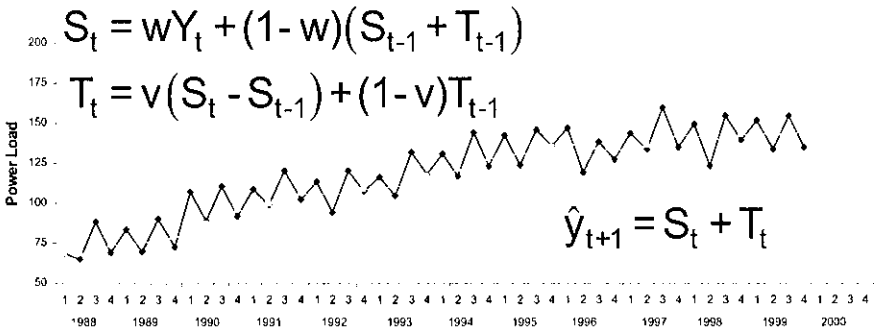


شکل ۸-۷) میانگین متحرک موزون نمایی (w=۰,۳٤)

انتخاب مقدار w را می‌توان براساس حداقل کردن شاخصهای میانگین قدر مطلق خطا و میانگین خطای مربع شده که دربحث پیش‌بینی به آنها اشاره شد، انجام داد. در مثال بالا $W = ۰/۳٤$، براساس حداقل کردن مقدارمیانگین خطای مربع شده به‌دست آمده است.

هموارسازی نمایی با تنظیم روند

مانند روش میانگین متحرک، هموارسازی نمایی ساده، نسبت به روند واکنش نشان نمی‌دهد. در این حالت استفاده از روشهای دیگر نظیر هموارسازی نمایی با تنظیم روند مناسب است.



شکل ۸-۸) پیش‌بینی با استفاده از میانگین متحرک موزون نمایی (w=۰,۳٤)

هموارسازی نمایی با تنظیم اثر روند و فصلی

همان‌طورکه گفته شد، در تجزیه و تحلیل سریهای زمانی گاهی اوقات لازم است اثرات روند فصلی یا سیکلی را به‌طور مجزا مطالعه کنیم. برای حذف کردن و یا اندازه‌گیری هریک از این مؤلفه‌ها، در تحلیل سریهای زمانی از روشهای مربوطه استفاده می‌شود. روشهای میانگین متحرک و غیره که قبلاً به آنها اشاره شد، برخی از این روشها می‌باشند. یکی دیگر از روشهای مورد استفاده برای حذف جزء روند یا فصلی، استفاده از روش تفاضلی است. مثلاً برای داده‌های غیر فصلی، تفاضل مرتبه اول طبق رابطه (۸-۷) برای رسیدن به ایستایی کافی است.

$$y_t = x_{t+۱} - x_t \qquad (۷-۸)$$

همچنین برای برآورد اثر فصلی برحسب اینکه اثر فصلی جمعی یا ضربی باشد، می‌توان با استفاده از روابط (۸-۸) تا (۸-۹) آن را برآورد کرد.

$$x_t - S_m(x_t) \qquad (۸-۸)$$

$$\frac{x_t}{S_m(x_t)} \qquad (۹-۸)$$

$$S_m(x_t) = \frac{\frac{۱}{۲}x_{t-۶} + x_{t-۵} + \ldots + x_{t+۵} + \frac{۱}{۲}x_{t+۶}}{۱۲} \qquad \text{شاخص داربین} \qquad (۱۰-۸)$$

همچنین یکی از راههای حذف کردن اثر فصلی، استفاده از شیوه تفاضلی است. [۲]

۸-۱-۴- جستجوی تشابه در تحلیل سریهای زمانی

مطابق آنچه که در ابتدای فصل اشاره شد، مباحث مدلسازی و پیش‌بینی سریهای زمانی در بسیاری از مراجع آماری مورد بررسی قرارگرفته است. لیکن آماردانها روشهای مناسبی را برای تشابه و شاخص‌گذاری سریهای زمانی بررسی نکرده‌اند. تعداد زیادی از این مسائل توسط جامعه علمی کامپیوتر حل شده‌اند. یکی از مسائل جالب در داده‌های سری زمانی، یافتن سریهای زمانی متفاوتی است که رفتار مشابه داشته باشند. مسئله را می‌توان این‌گونه نیز مطرح کرد که آیا دو سری زمانی X و Y داده شده مشابه هستند یا خیر؟ به‌عبارت دیگر، تابع $Sim(X,Y)$ را تعریف کرده و میزان تشابه دو سری یا به‌طور مشابه، تابع فاصله $Dis(X,Y)$ را محاسبه می‌کنیم. برای نمونه، هر سری زمانی مقدار و سیر تدریجی یک شیء را به‌صورت تابعی از زمان در مجموعه‌ای از داده‌های جمع‌آوری شده توصیف می‌کند (مثلاً قیمت سهام). هدف می‌تواند خوشه‌بندی اشیاء مختلف در گروه‌های مشابه (مانند گروهی از سهامها که تغییر قیمت یکسان داشته‌اند) یا دسته‌بندی اشیاء براساس مجموعه‌ای از ویژگیهای شناخته شده باشد. این مورد مشکل است، چون مدل تشابه باید اجازه تطابق غیردقیق را بدهد.

یکی از مسائل جالب مطرح شده در تشابه دنباله‌ها[۱]، تشابه زیر دنباله‌ها می‌باشد که در آن برای یک سری زمانی داده شده X و الگوی سری زمانی کوتاهتر Y، می‌خواهیم زیر دنباله‌ای از X را که مشابه الگوی Y عمل می‌کند پیدا کنیم. برای پاسخ به این پرسش، نظریه‌های متفاوتی از تشابه سریهای زمانی در پژوهش‌های داده‌کاوی مطرح شده است. در این بخش مدلهای مختلف اندازه‌گیری تشابه سریهای زمانی مطرح می‌شود که براساس شاخصهای کارآیی و دقت، می‌توان آنها را ارزیابی کرد. نمونه‌هایی از اندازه‌گیری تشابه را که بر اساس نرم اقلیدسی، تخمینهای خطی قطعه‌ای[۲]، تاباندن زمانی پویا[۳] (DTW) و بزرگترین زیردنباله‌های مشترک[۴] (LCSS) هستند را بررسی خواهیم کرد. [۵]

---

[۱]- Sequences

[۲]- Piecewise Linear Approximations: PLA

[۳]- Dynamic Time Warping: DTW

[۴]- Longest Common Subsequences Similarity: LCSS

۸-۱-۵- مقیاسهای اندازه‌گیری تشابه در سریهای زمانی

فاصله‌اقلیدسی و نرم L_p [1]

یکی از ساده‌ترین راههای اندازه‌گیری تشابه در سریهای زمانی اندازه‌گیری فاصله اقلیدسی است. دو دنباله زمانی با طول یکسان n را فرض کنید. ما هر دنباله را در فضای n بعدی اقلیدسی، به‌عنوان یک نقطه می‌بینیم. عدم شباهت یا فاصله بین دو دنباله X و Y را با $L_p(X,Y)$ تعریف می‌کنیم (وقتی $p=۲$ است، این فاصله، همان فاصله اقلیدسی معروف است). این اندازه‌گیری مزایای مختلفی دارد. فهم آن آسان، محاسبه آن ساده و برای حل مشکلات دیگر مثل شاخص‌گذاری وخوشه‌بندی سریهای زمانی قابل استفاده است. هرچند معایب زیادی نیز دارد که آن را برای کاربردهای متعددی نامناسب می‌کند. یکی از اصلی‌ترین معایب آن این است که اجازه نمی‌دهد که دنباله‌های زمانی خط مبنای متفاوتی داشته باشند. برای مثال سهام X با نوسان حدود ۱۰۰$ و Y با نوسان حدود ۳۰$ را در نظر بگیرید. حتی اگر شکل هر دو دنباله زمانی به هم خیلی شبیه باشد، ممکن است فاصله اقلیدسی بین آنها خیلی زیاد شود. همچنین نمی‌توان به این روش در مقیاسهای مختلف اندازه‌گیری نمود. برای مثال ممکن است سهام X در دامنه کوچکی نوسان کند (بین ۹۵$ تا ۱۰۵$) درحالی که سهام Y در دامنه بزرگتری نوسان کند (بین ۲۰$ تا ۴۰$).[۵]

تبدیلات نرمال [2]

با استفاده از نرمال‌کردن دنباله‌ها می‌توان معایب نرم L_p را در اندازه‌گیری تشابه برطرف کرد. در رابطه (۸-۱۰) اگر $\mu(X)$ میانگین و $\sigma(X)$ واریانس دنباله باشد، دنباله $X=\{x_1,\ldots,x_n\}$ را با دنباله نرمال X' جایگزین می‌کنیم:

$$x_i' = (x_i - \mu(X))/\sigma(X) \qquad (۸-۱۱)$$

همچنین دنباله Y را با دنباله نرمال Y' جایگزین می‌کنیم. در نهایت عدم تشابه بین X و Y را با $L_p(X',Y')$ تعریف می‌کنیم. این تعریف تشابه، معایب استفاده مستقیم از نرم L_p را که دنباله‌ها نرمال نیستند، حل می‌کند. برای مثال دو سهام X و Y که قبلاً در مورد آن صحبت شد،

---

[1]- Euclidean Distances and Lp Norms

[2]- Normalization Transformations

را در نظر بگیرید. بعد از نرمال شدن هر دو دارای یک خط مبنا شده (چون میانگین هر دو با نرمال کردن یکسان می‌شود) و دامنه یکسانی خواهند داشت (چون با نرمال کردن واریانس داده‌ها یکسان شده است).

فرآیند نرمال کردن هم معایب خاص خود را دارد. مثلاً به تأخیر و تقدم فاز در زمان خیلی حساس است. به‌عنوان نمونه دو دنباله X و Y را در نظر بگیرید که X شبیه به موج سینوسی است، در صورتی‌که Y شبیه به موج کسینوسی است. هر دو اصولاً دارای شکل یکسانی هستند به‌جز اینکه یک تأخیر فاز دارند. ولی فاصله بین این دو دنباله قابل ملاحظه است (در هردو حالت نرمال شده یا نرمال نشده). همچنین فاصله‌اقلیدسی، شتاب وکاهش شتاب در طول محور اصلی را نشان نمی‌دهد. برای مثال دو دنباله X و Y که شبیه موج سینوسی هستند را در نظر بگیرید که فقط پریود X دو برابر دوره Y است. حتی اگر این دو دنباله نرمال شوند، فاصله‌اقلیدسی نمی‌تواند شباهت بین این دو سیگنال را نشان دهد [۵].

تبدیلات عمومی[1]

تشخیص اهمیت نظریه حالت، در محاسبات تشابه در سال ۱۹۹۵ مطرح شد. یک چارچوب تشابه عمومی شامل زبان قواعد تبدیلات توصیف شده است. هر روش در زبان تبدیلات، یک دنباله ورودی را گرفته و با هزینه مربوط به آن روش، دنباله خروجی متناظرش را تولید می‌کند. تشابه بین دنباله X و Y، حداقل هزینه برای تغییر X به Y با استفاده از چنین روشهایی می‌باشد. برای مثال در شکل زیر دنباله‌هایی که به شکل منحنی‌های خطی-قطعه‌وار هستند، نشان داده شده است. به‌عنوان نمونه، یک قاعده تبدیل می‌تواند یکی کردن بخش‌های مجاور روی یک بخش باشد. هزینه این قاعده می‌تواند تابعی از طول و شیب بخش جدید و بخش ابتدایی باشد. قاعده دیگر می‌تواند یک بخش منفرد را با یک جفت بخش مجاور جایگزین کند.

شکل ۸-۹) دنباله‌هایی به شکل منحنی‌های خطی

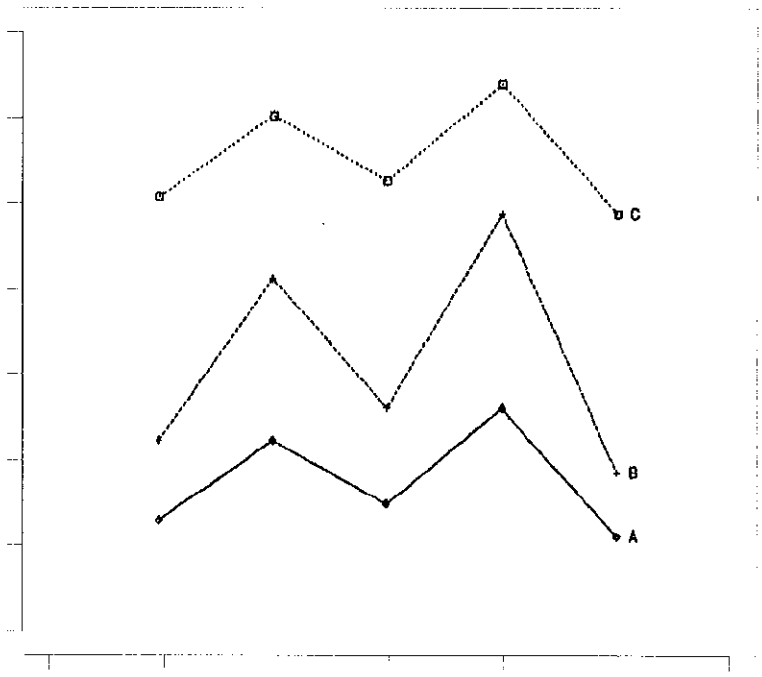
---

[1]- General Transformations

یکی از قواعد تبدیلات استفاده از میانگین متحرک برای هموارکردن سریهای زمانی است. برای مثال سری زمانی X که قیمت روزانه سهام در طول یک فاصله زمانی چند روزه است را در نظر می‌گیریم. برای هر ۳ روز یک میانگین متحرک حساب کرده و در دنباله X تغییر می‌دهیم و نتیجه دنباله X' می‌شود که $x' = (x_{i-1} + x_i + x_{i+1})/3$.

یکی دیگر از روشهای تبدیل، تبدیل مقیاس و انتقال است. در تبدیل مقیاس، هر جزء با مقیاس ثابتی افزایش می‌یابد. برای مثال هر x_i با cx_i جایگزین می‌شود که c مقداری ثابت است. هر تبدیل انتقال هر جزء را با یک عدد ثابت از موقعیت فعلی به سمت راست یا چپ منتقل می‌کند (هر x_i با x_i+c جایگزین می‌شود که c یک عدد صحیح ثابت است).

مثال: سه دنباله شکل زیر را در نظر بگیرید.



شکل ۸-۱۰) سه دنباله نمونه

$A = (۵.۱۰,٦,۱۲,٤)$
$B = (۱۰,۲۰,۱۲,۲٤,۸)$
$C = (۲۵,۳۰,۲٦,۳۲,۲٤)$

این سه سری زمانی متفاوتند، اما باهم رابطه نزدیکی دارند. سری A می‌تواند با دو برابر کردن جملات به B تبدیل شود و C با ۲۰ واحد انتقال می‌تواند به A تبدیل شود. به علاوه B می‌تواند با نصف کردن جملاتش و سپس ۲۰ واحد انتقال تبدیل به C شود. این بدین معنی است که این سریها با تبدیل مقیاس‌بندی و انتقال مناسب، در واقع یکی هستند. اگر سه دنباله بالا را به‌عنوان روند قیمت سه سهام در نظر بگیریم، با وجود اینکه قیمت سهام شرکت C بیشتر از شرکت A است، اما چون نوسان یکسانی دارند دقیقاً از روند قیمتی یکسانی پیروی می‌کنند. یا اگر چه قیمت سهام شرکت B همیشه دو برابر قیمت سهام شرکت A ولی نوسان آنها متناسب با قیمتشان است در نتیجه روند قیمتی آنها باید یکسان در نظرگرفته شود. سری A با سری B مشابه است، اگر A بتواند با یکی از تبدیلات گفته شده به B تبدیل شود [۷].

قواعد تبدیل، یک روش عمومی را برای تعریف تشابه که مناسب کاربردهای خاصی است، پیشنهاد می‌دهد. هرچند بعضی از معایب آنها، مشکلاتی نیز ایجاد می‌کنند. مثلاً محاسبات زیر دنباله‌ها (مانند شاخص‌گذاری) مشکل می‌شود، چون استخراج مشخصه‌ها از دنباله X، مخصوصاً اگر قواعد استفاده شده به هر دو دنباله X و Y بستگی داشته باشد، کار پیچیده‌ای است. همچنین فاصله‌اقلیدسی در فضای مشخصه ممکن است تقریب خوبی برای عدم تشابه دنباله‌های ابتدایی نباشد [۵].

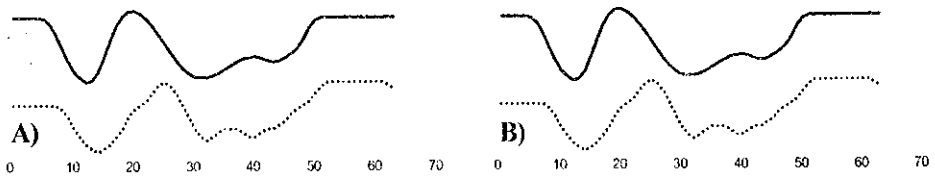
۸-۱-۶- تاباندن محور زمان به‌صورت پویا[1]

یکی از معمول‌ترین کارهایی که با داده‌های سریهای زمانی انجام می‌دهند، مقایسه یک دنباله با دنباله دیگر است. در بعضی از کاربردها یک مقیاس اندازه‌گیری ساده مانند اندازه‌گیری فاصله‌اقلیدسی، کافی است. اگرچه اغلب این حالت پیش می‌آید که دو دنباله با اجزای تقریباً یکسان در محور xها، در قسمتی نسبت به هم کشیده‌تر هستند. شکل زیر این موضوع را با یک مثال ساده نشان می‌دهد. برای فهمیدن شباهت چنین دنباله‌هایی، قبل از میانگین گرفتن از آنها،

---

[1]- Dynamic Time Warping

به‌عنوان یک قدم پیش‌پردازش، یک یا هر دو دنباله را روی محور زمان می‌پیچانیم. روش *DTW* برای این نوع تاباندن روی محور زمان، کارآمد می‌باشد. علاوه بر داده‌کاوی، *DTW* در تشخیص حرکات علم روباتیک، تحلیل سخنرانی و پزشکی کاربرد دارد.



شکل ۸-۱۱) دو دنباله که وضعیت دستخط یک شخص را هنگام نوشتن کلمه Pen در زبان علامت روی محورها نشان می‌دهد.

شکل (۸-۱۱) یک مثال از کاربرد *DTW* می‌باشد. نمودار *A* شامل دو دنباله است که وضعیت دستخط یک شخص را هنگام نوشتن کلمه *Pen* در زبان علامت روی محورها نشان می‌دهد. دنباله‌ها در دو روز مختلف ضبط شده‌اند. توجه کنید با وجود اینکه دنباله‌ها شکل عمومی یکسانی دارند، ولی در محور زمان بر هم منطبق نیستند. یک مقیاس فاصله که فرض می‌کند *i*امین نقطه روی یک دنباله منطبق بر *i*امین نقطه روی دنباله دیگر است، باعث یک عدم تشابه ناامید کننده می‌شود. نمودار *B* می‌تواند به صورت کارآمد یک تطابق بین دو دنباله ایجاد کند که محاسبه فاصله پیچیده‌تری را ایجاب می‌کند.

اگر چه استفاده از *DTW* در بسیاری از زمینه‌ها موفق بوده است، می‌تواند نتایج غیر قابل کنترل و نامطلوبی داشته باشد. مشاهدات حاکی از آن است که الگوریتم می‌تواند تغییرات در محور *y*ها را با تاباندن روی محور *x* ها توضیح دهد. این می‌تواند باعث تطابق‌های غیرشهودی هنگام تصویر یک نقطه منفرد از یک سری زمانی، روی یک زیربخش بزرگ از سری زمانی دیگر شود. ما چنین رفتار غیردلخواهی را مقادیر منفرد یا تکینها[1] می‌نامیم. بخش وسیعی از روشها برای مقابله با این رفتارها ارائه شده‌اند. دستاورد این روشها، چگونگی تاباندن مجاز را مشخص

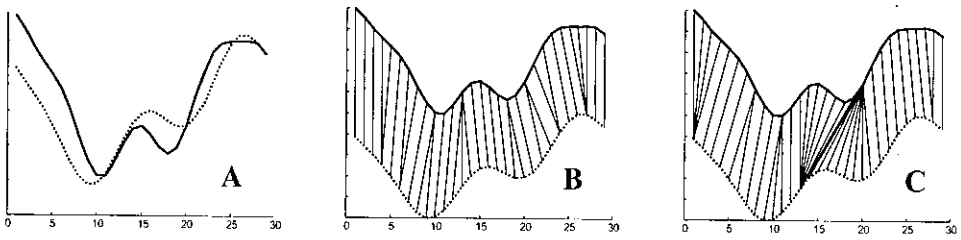
---

[1]- Singularities

می‌کنند. اگر چه استفاده از این روشها در بعضی مواقع از یافتن روش تاباندن صحیح جلوگیری می‌کند.

در موارد شبیه‌سازی شده، تاباندن وقتی صحیح تشخیص داده می‌شود که ما ابتدا یک سری زمانی را بتابانیم و سپس سعی کنیم سری اصلی را از روی سری تابانده شده به‌دست آوریم. در رویدادهای طبیعی، منظور ما از روش تاباندن صحیح آن است که مانند شکل (۸-۱۲- B) به‌طور شهودی تطابق یک به یک مشخصه‌ها واضح باشد.

یک مشکل دیگر با DTW آن است که الگوریتم ممکن است در پیدا کردن تطابق‌های طبیعی و واضح در دو دنباله، تنها به این دلیل که یک مشخصه در یک دنباله کمی بالاتر یا پایین‌تر از مشخصه مربوطه در دنباله دیگر است، اشتباه کند. برای مثال نقاط اوج یا حضیض، نقطه خمیدگی، قسمت مسطح و غیره شکل (۸-۱۲) این مسئله را نشان می‌دهد.



شکل ۸-۱۲- A) دو سیگنال ترکیبی (با میانگین وواریانس یکسان). B) تطابق نظیر به نظیر مشخصه‌ها. C) تطابق ایجاد شده به‌وسیله DTW

توجه کنید که DTW، دو نقطه اوج مرکزی را به‌علت اینکه آنها در محور Yها کمی با هم فاصله دارند، منطبق در نظر گرفته است [۸].

۸-۱-۷- الگوریتم کلاسیک DTW

فرض کنید دو سری زمانی Q و C را با طول‌های m و n داریم[۸] :

$$Q = q_1, q_2, \ldots, q_i, \ldots, q_n$$
$$C = c_1, c_2, \ldots, c_j, \ldots, c_m$$

برای تطابق دو دنباله با استفاده از DTW یک ماتریس $m*n$ می‌سازیم به‌طوری‌که عنصر (ith, jth) ماتریس، شامل فاصله $d(q_i, c_j)$ بین دو نقطه c_j و q_i می‌باشد (معمولاً از فاصله اقلیدسی

استفاده می‌کنیم)، بنابراین $d(q_i,c_j)=(q_i-c_j)^2$. هر عنصر *(i,j)* ماتریس به تطابق بین نقاط q_i و c_j مربوط است که در شکل (۸-۱۳) نشان داده شده است. یک مسیر تاباندن W، یک دنباله از عناصر پیوسته ماتریس است که یک نگاشت را بین Q و C مشخص می‌کند. عنصر k ام W به‌صورت $w_k = (i,j)$ تعریف می‌شود و در نتیجه خواهیم داشت.

$$W = w_1, w_2, \ldots, w_k, \ldots, w_k \qquad \max(m,n) \le K < m+n-1 \qquad (۱۲-۸)$$

مسیر تاباندن معمولاً محدودیتهایی دارد.

شرایط حدی: $w_k=(m,n)$ و $w_1=(1,1)$ به‌راحتی نشان می‌دهند که مسیر تاباندن از اولین نقطه روی قطر اصلی شروع و به نقطه مقابل آن در انتهای قطر اصلی خاتمه می‌یابد.

پیوستگی: اگر $w_k=(a,b)$ و $w_{k-1}=(a',b')$ باشند، لازم است که $b-b' \le 1$ و $a-a' \le 1$. این شروط گامهای مجاز را در مسیر تاباندن برای سلولهای مجاور از جمله سلولهای مجاور قطری مشخص می‌کند (هیچ جزئی نمی‌تواند در دنباله حذف شود).

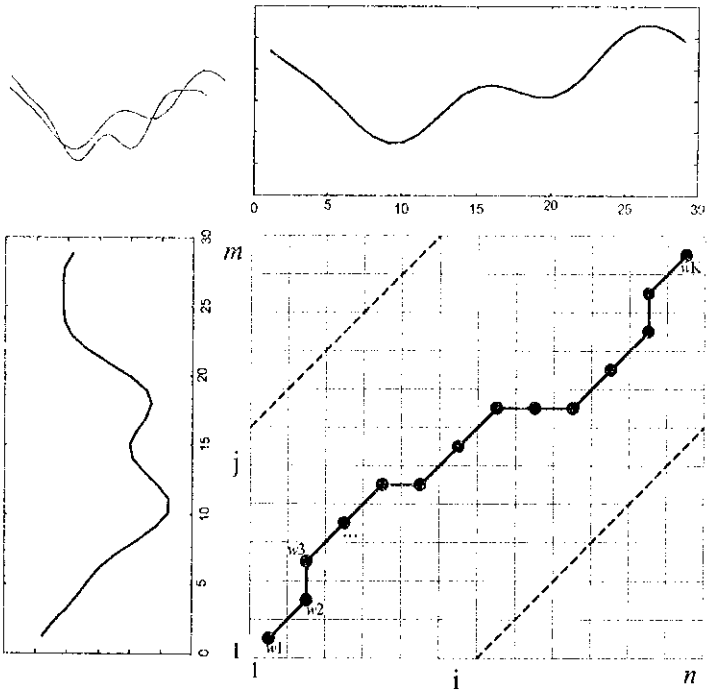
یکنواختی: اگر $w_k=(a,b)$ و $w_{k-1}=(a',b')$ لازم است که $a-a' \ge 0$ و $b-b' \ge 0$. این باعث می‌شود نقاط در W به‌صورت یکنواخت و بر حسب زمان قرار بگیرند.

باتوجه به محدودیتهای گفته شده تعداد زیادی مسیر تاباندن وجود دارد ولی تنها مسیرهایی که هزینه تاباندن را حداقل می‌کنند مورد نظر ما هستند.

$$DTW(Q,C)=\min\left\{\sqrt{\sum_{k=1}^{K} w_k} \,/\, K\right. \qquad (۱۳-۸)$$

برای نشان دادن اینکه مسیرهای مختلف می‌توانند طول‌های مختلف داشته باشند، K در رابطه (۸-۱۲) در مخرج ظاهر می‌شود. این مسیر را می‌توان به‌صورت کارآمد با استفاده از برنامه‌ریزی پویا به‌دست آورد. برای این کار $\gamma(i,j)$ را که فاصله تجمعی است به‌دست می‌آوریم. *d(i,j)* در فرمول (۸-۱۳) همان فاصله i و j ای است که در ماتریس به‌دست آوردیم و قسمت بعد حداقل مقدار فاصله سلولهای مجاور است.

$$\gamma(i,j)=d(q_i,c_j)+\min\{\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\} \qquad (۱۴-۸)$$



شکل ۸-۱۳) مثالی از یک مسیر تاباندن

محدودیت‌های الگوریتم کلاسیک *DTW*

مشکل مقادیر تکین از اوایل سال ۱۹۷۸ توسط ساکوئی و چیبا[1] مورد توجه قرارگرفت. روشهای مختلفی برای کم‌رنگ‌تر کردن این مشکل مطرح شده است که ما اجمالاً به بررسی آنها می‌پردازیم [۸].

۱) **پنجره بندی**[2]: عناصر مجاز ماتریس به آنهایی که در پنجره می‌افتند، محدود می‌شوند $|i-(n/(m/j))|<R$ که R یک عدد صحیح مثبت است و نمایانگر عرض پنجره است. این بدین معنی است که گوشه‌های ماتریس هرس می‌شوند. همان‌طورکه در شکل (۸-۱۳) با خط‌چین نشان داده شده است. دیگران پنجره‌بندی را با اشکال مختلف دیگری تجربه کردند این دستاورد تا جای ممکن مشکل تکینها را محدود می‌کند اما از رخداد آن جلوگیری نمی‌کند.

---

[1]- Sakoe & Chiba

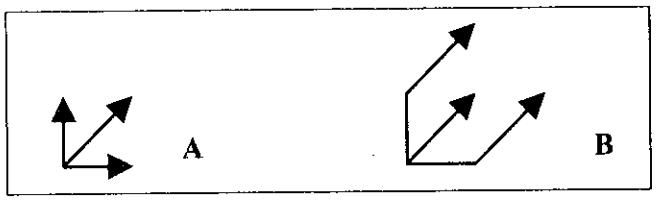
[2]- Windowing

۲)وزن‌دهی به شیب[1]: اگر در معادله (۸-۱۵) فاصله را به‌صورت زیر در نظر بگیریم:

$$\gamma(i,j) = d(i,j) + min\{\gamma(i-1,j-1), X\gamma(i-1,j), X\gamma(i,j-1)\} \qquad (۸-۱۵)$$

به‌طوری‌که X یک عدد حقیقی مثبت است و می‌توانیم شکل تاب دادن را با عوض‌کردن مقدار X محدود کنیم. هرچه مقدار X بزرگ‌تر شود مسیرتاب دادن به سمت قطری شدن تمایل پیدا می‌کند.

۳) محدودیت‌های شیب[2]: ما می‌توانیم معادله فاصله را به‌صورت یک نمودار الگوی گامهای قابل قبول به‌تصویر درآوریم. مثلاً در شکل (۸-۱٤-A) پیکانها نشان‌دهنده گامهای مجازی هستند که مسیر تاباندن در هر مرحله می‌تواند بردار را نشان دهد. می‌توانستیم معادله (۸-۱۵) را با معادله‌های زیر که به الگوی گام نشان داده شده در شکل (۸-۱٤-B) مربوط است جایگزین کنیم با استفاده از این معادله، مسیر تاباندن یک قدم قطری، گام برمی‌دارد که این قدم می‌تواند بعد از طی یک قدم موازی با یکی از محورها صورت گیرد.



شکل ۸-۱٤) یک نمایش تصویری از رو نوع الگوی مسیر متفاوت

$$\gamma(i,j) = d(i,j) + \min(\gamma(i-1,j), \gamma(i-1,j), \gamma(i,j-1) \qquad \text{الگوی } (A)$$

$$\gamma(i,j) = d(i,j) + \min[\gamma(i-1,j-1), \gamma(i-1,j-2), \gamma(i-2,j-1) \qquad \text{الگوی } (B)$$

همه موارد گفته شده می‌توانند در محدود کردن مشکل تکینها کمک کنند ولی ریسک از دست دادن روش تاب‌دادن صحیح، هنوز وجود دارد. مشکل دیگر آن است که چگونگی انتخاب پارامترهای موجود در الگوها هنوز برای ما واضح نیست. برای مثال نمی‌دانیم عدد R را در پنجره‌بندی و عدد صحیح X را در الگوی وزن‌دهی شیب چگونه به‌دست آوریم.

---

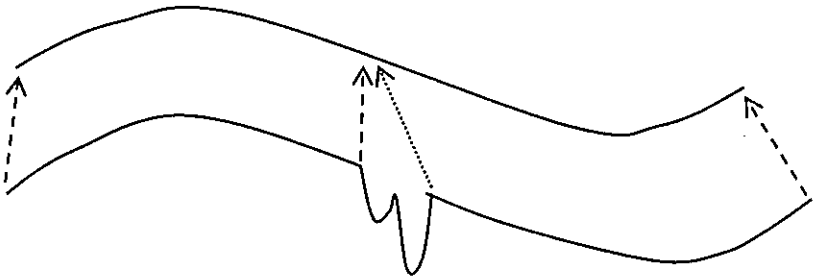
[1]- Slope Weighting

[2]- Slope Constraints

۸-۱-۸- شباهت بزرگترین زیردنباله مشترک(LCSS)

شباهت بزرگترین زیردنباله مشترک برای اندازه‌گیری اختلاف سریهای زمانی، در مواردی مانند تشخیص صدا و تطابق الگوی متن به‌کار می‌رود. ایده اصلی، تطبیق دو دنباله بر یکدیگر است، حتی اگر چند جزء آنها منطبق نباشد. روش *LCSS* دو مزیت دارد: (الف) بعضی از اجزا می‌توانند منطبق نباشد (مثل نقاط پرت)، ولی در فاصله اقلیدسی و *DTW* همه اجزاء حتی نقاط پرت باید منطبق باشد. (ب) همان‌طورکه در ادامه دیده خواهد شد اندازه‌گیری *LCSS*، کارآیی محاسبات تقریبی را بیشتر می‌کند. حال به بررسی یک مثال شهودی‌تر می‌پردازیم. دو دنباله X و Y را در نظر بگیرید:

$$Y = \{۲,۵,٤,۷,۳,۱۰,۸\} \quad , \quad X = \{۳,۲,۵,۷,٤,۸,۱۰,۷\}$$



شکل ۸-۱۵) اندازه‌گیری LCSS

شکل (۸-۱۵) ایده اصلی اندازه‌گیری *LCSS* را مشخص می‌کند. فرض کنید قسمتهای مشخصی از هر دنباله در فرآیند تطابق حذف شده باشد (مثل نقاط پرت یا داده‌های مغشوش). مقدار *LCSS* بین X و Y برابر $\{۲,۵,۷,۱۰\}$ است.

دو دنباله X و Y، به ترتیب با طول m و n را در نظر می‌گیریم. مشابه همان کاری که در *DTW* انجام شد، یک رابطه بازگشتی برای طول *LCSS* دنباله X و Y به‌دست می‌آید. مقدار $L(i,j)$ نشان دهنده *LCSS* زیر دنباله $\{x_۱,...,x_i\}$ و $\{y_۱,...,y_j\}$ می‌باشد. $L(i,j)$ می‌تواند یک رابطه بازگشتی به‌صورت زیر باشد.

$$If\ x_i = y_i\ then\ L(i,j) = ۱ + L(i-۱, j-۱), else\ \ l(i,j) = max\{D(i-۱,j), D(i,j-۱)\}$$

ما عدم تشابه بین X و Y را با $LCSS(X,Y)=(m+n-2l)/(m+n)$ تعریف می‌کنیم، که lطول $LCSS$ است. این کمیت، حداقل مقدارنرمال شده تعداد عناصری است که باید از X حذف و یا به X اضافه شوند تا X به Y تبدیل شود. مشابه DTW، اندازه‌گیری $LCSS$ به‌وسیله برنامه‌ریزی پویا درزمان $O(mn)$، می‌تواند محاسبه شود. اگر پنجره تطابق با طول w مشخص شده باشد که $|i-j|$ می‌تواند حداکثر w باشد، می‌توان پیچیدگی آن را به مقدار $O(wn)$ کاهش داد. الگوریتم یافتن طولانی‌ترین مسیر مشترک بین دو دنباله به‌صورت زیر است:

```
LCSS - LENGTH(X, Y)
 m = length [X], n = length [Y]
 for i = ۱ to m do c[i,۰] ← ۰
  for j = ۱ to n do c[۰, j] ← ۰
  for i = ۱ to m
    do for j = ۱ to n
      do if x_i = y_j
        then c[i, j] ← c[i − ۱, j − ۱] + ۱  b[i, j] ="↖"
        else if c[i − ۱, j] ≥ c[i, j − ۱]
          then c[i, j] ← c[i − ۱, j]  b[i, j]="↑"
          else c[i, j] ← c[i, j − ۱]  b[i, j]="←"
```

در $LCSS$ نیاز به اینکه داده‌های متناظر در زیر دنباله‌های مشترک دقیقاً منطبق باشند، کار را کمی سخت می‌کند. این مشکل با استفاده از تلرانسهای ($\varepsilon>0$) هنگام مقایسه عناصرقابل حل است. بنابراین، مطابق گفته بالا، دو جزء a وb (به ترتیب از دنباله‌های X و Y) مطابقند اگر:

$$a(1-\varepsilon)<b<a(1+\varepsilon)$$

به جای تلرانس نسبی، ممکن است از تلرانس مطلق استفاده شود، یعنی b باید بین $a-\varepsilon$ و $a+\varepsilon$ باشد[۵]. برای آشنایی بیشتر با این الگوریتم، مثال زیر توضیح داده می‌شود.

مثال: فرض کنید دنباله‌های X وY به‌صورت زیر داده شده باشند. برای به‌دست آوردن $LCSS$ این دو دنباله، قدمهای زیر را دنبال می‌کنیم.

$X = ABCB$
$Y = BDCAB$

همه عناصر رشته‌ها را در ستونها و ردیفها می‌چینیم

	Y_j	B	D	C	A	B
X_i						
A						
B						
C						
B						

همه عناصر ستون و ردیف اول را صفر می‌کنیم

	Y_j	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰					
B	۰					
C	۰					
B	۰					

طبق قدمهای الگوریتم همه عناصر ستونها و ردیفها را با هم مقایسه کرده و عناصر متناظر در جدول را مقداردهی می‌کنیم اول را صفر می‌کنیم

j	۰	۱	۲	۳	٤	٥
	Y_j	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰					
B	۰					
C	۰					
B	۰		۰			

	Y_j	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰		
B	۰					
C	۰					
B	۰					

	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱	
B	۰					
C	۰					
B	۰					

	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱ →	۱
B	۰					
C	۰					
B	۰					

	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱	۱
B	۰ →	۱ →	۱ →	۱	↓ ۱	
C	۰					
B	۰					

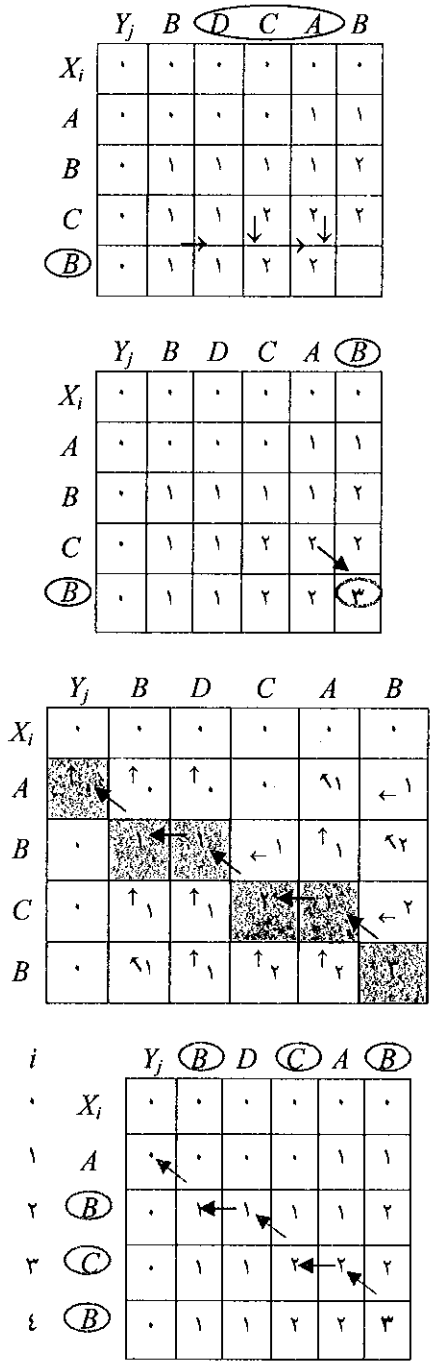
	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱	۱
B	۰	۱	۱	۱	۱	۲
C	۰					
B	۰					

	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱	۱
B	۰	۱	۱	۱	۱	۲
C	۰	۱	۱			
B	۰					

	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱	۱
B	۰	۱	۱	۱	۱	۲
C	۰	۱	۱	۲		
B	۰					

	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱	۱
B	۰	۱	۱	۱	۱	۲
C	۰	۱	۱	۲	۲	۲
B	۰					

	Y_i	B	D	C	A	B
X_i	۰	۰	۰	۰	۰	۰
A	۰	۰	۰	۰	۱	۱
B	۰	۱	۱	۱	۱	۲
C	۰	۱	۱	۲	۲	۲
B	۰	۱				



شکل ۸-۱۶) قدم‌های الگوریتم

LCS (reversed order): ***B C B***

LCS (straight order): ***B C B***

خوشه‌بندی جریان کلیکها[1] با کمک بزرگترین دنباله‌های مشترک وزنی

گروه‌بندی بازدیدکنندگان براساس تعامل آنها با یک وب سایت، یک مشکل کلیدی در کاوشهای کاربردی وب است. جریان کلیکهای تولید شده به‌وسیله کاربرهای مختلف اغلب الگوهای مشخصی را دنبال می‌کنند. برای خوشه‌بندی کاربران وب، براساس جریان کلیک در یک وب سایت و زمان صرف شده روی هر صفحه، از یک الگوریتم LCSS استفاده شده است [۱۱].

با افزایش سریع کاربردهای تجارت الکترونیک، آگاهی از رفتار کاربر بر اساس تعاملاتش با یک وب سایت برای صاحبان وب سایت اهمیت بیشتری پیدا کرده است. تشخیص رفتار هر کاربر در بازدید از یک وب سایت می‌تواند مدیران سایت را برای تهیه محتوای سفارش‌شده برای دیگرکاربران قادر سازد. این موضوع کاربردهای زیادی در کسب وکار دارد. جریان کلیکهای یک کاربر، دنباله‌ای از صفحه‌های بازدید شده توسط او در یک وب سایت خاص در یک جلسه[2] می‌باشد. هدف، خوشه‌بندی کاربران بر اساس جریان کلیکها در یک وب سایت خاص و یافتن گروه‌هایی از کاربران است که با علایق و انگیزه‌های مشابه از یک سایت بازدید می‌کنند. در نتیجه همبستگی قوی بین جریان کلیکهای کاربران وجود دارد که نشان دهنده تشابه علایق کاربران است.

۸-۱-۹- روشهای شاخص‌گذاری برای جستجوی تشابه در سریهای زمانی

مسئله دیگری که در بحثهای کاربردی سریهای زمانی مطرح است مسئله شاخص‌گذاری/بازیابی[3] است [۵]. مجموعه سریهای زمانی $S=\{Y_1,...,Y_N\}$ را در نظر گرفته وسری مورد نظر X را داریم. سریهای زمانی موجود در S را که بیشترین تشابه با سری X دارند، پیدا می‌کنیم. برای مثال، روزهایی از سال که یک سهام مشخص تغییرات مشابهی مانند امروز داشته باشد را جستجو می‌کنیم.

---

[1]- Clickstream

[2]- Session

[3]- Indexing/Retrieval

مسئله دیگر، شاخص‌گذاری زیر دنباله است. مجموعه دنباله‌های S و دنباله یا الگوی مورد نظر X داده شده است. دنباله‌هایی را در S می‌یابیم که شامل زیر دنباله‌های مشابه X باشد. برای حل کارآی این نوع مسائل، باید از روشهای مناسب شاخص‌گذاری استفاده کنیم.

مسئله تشابه، مرتبط با مسئله شاخص‌گذاری می‌باشد. معمولاً تعیین شاخص برای روشهای اندازه‌گیری ساده شباهت، آسان و احتمالاً کم‌دقت است. ولی اندازه‌گیری شباهتهای پیچیده، تعیین شاخص را دشوار و جالب می‌کند.

یک سری زمانی با طول n می‌تواند به‌عنوان یک مشاهده در فضای n بعدی درنظر گرفته شود. شاخص‌گذاری مستقیم در این فضا به‌علت ابعاد خیلی زیاد آن ناکارآ می‌باشد. راه حل، استفاده از روشهای کاهش بُعد است که در آن سری زمانی X با n مشاهده درنظرگرفته شده، چند مشخصه کلیدی آن استخراج و به نقطه $f(X)$ در فضای مشخصه با ابعاد کمتر k نگاشت می‌شود (امید داریم که $k<<n$ باشد). این نگاشت باید طوری انجام شود که تشابه یا فاصله بین X و Y تقریباً با فاصله اقلیدسی دو نقطه $f(X)$ و $f(Y)$ برابر باشد. می‌توان از روشهای شناخته‌شده دسترسی فاصله‌ای برای شاخص‌گذاری فضای مشخصه‌های دارای ابعاد کمتر استفاده کرد، مانند *R-trees*، *kd-trees* یا *VP-trees*.

در بسیاری از موارد سرعت دستیابی به سریهای زمانی مشابه با سری زمانی مورد جستجو، حائز اهمیت است. این مسئله در بسیاری از کاربردها دیده می‌شود. مثلاً یافتن سهامهایی که شبیه یک سهام خاص رفتار می‌کنند، پیدا کردن محصولاتی که چرخه تقاضای یکسانی دارند و پیدا کردن ژنهایی که الگوی مبین آنها شبیه ژن خاصی است. این کاربردها نیازمند مکانیزم بازیابی برای دسته‌بندی سریهای زمانی با استفاده از روش دسته‌بندی نزدیک‌ترین همسایگی جهت بهبود زمان اجرا در الگوریتمهای خوشه‌بندی یا برای تحلیل اکتساب در داده‌های سریهای زمانی هستند.

شاخص‌گذاری در سریهای زمانی، توجه بسیاری از محققان را در سالهای اخیر جلب کرده است. با توجه به گسترش روزافزون اندازه بانکهای اطلاعاتی، شاخص‌گذاری می‌تواند در داده‌کاوی بسیار مؤثر باشد. اگر کاربر بخواهد در بانکهای اطلاعاتی گسترده شروع به اکتشاف کند، باید داده‌ها به شکلی سازمان‌دهی شوند که او بتواند به شکلی مؤثر و کارا داده‌های مورد

نظر خود را بازیابی کند. به‌طور عمومی می‌توان مسئله بازیابی سریهای زمانی را به‌صورت زیر تعریف کرد. بانک اطلاعاتی D شامل مجموعه‌ای از سریهای زمانی داده شده است. یک روش پیش‌پردازشی تعریف می‌کنیم که هدف آن پیدا کردن کارآی سری X عضو D نزدیک به سری داده شده Q است، (سری Q لزوماً در بانک اطلاعاتی وجود ندارد). برای حل این مسئله باید موارد زیر را در نظر گرفت:

- یک تابع فاصله که با درک کاربر از آنچه شباهت نامیده می‌شود، مطابقت دارد.
- یک رویه مؤثر شاخص‌گذاری که سرعت جستجوی کاربر را بالا می‌برد.

درزیربخش قبلی روشهای مختلف برای تعریف شباهت (یا فاصله) بین دوسری زمانی بررسی شد. آسان‌ترین روش، تعریف فاصله بین دو سری با نگاشت هریک برروی یک بردار وسپس استفاده از نرم L_p برای محاسبه بود. فاصله نرم L_p بین دوبردار n بعدی $\bar{x}$ و $\bar{y}$ به‌صورت زیر تعریف می‌شود:

$$L_p(\bar{x},\bar{y}) = \left(\sum_{i=1}^{n}(x_i - y_j)^p\right)^{\frac{1}{p}} \qquad (۱۶-۸)$$

برای $p=2$ این فرمول همان فاصله‌اقلیدسی مشهور و برای $p=1$ فاصله مانهاتان است. اگر چه محاسبه چنین مقیاس فاصله‌ای آسان است، اما به کوچک‌ترین تغییرات در محور زمان بسیار حساس بوده و برای داده‌های مغشوش به‌خوبی عمل نمی‌کند. همان‌طور که گفته شد، انعطاف پذیرترین روشها برای تعریف تشابه در سریهای زمانی، روشهای DTW و $LCSS$ هستند.

صورت مسئله برای کاربران مختلف می‌تواند متفاوت باشد. مثلاً کاربران می‌توانند به دنبال تطابق کل دنباله و یا تنها به دنبال تطابق یک زیردنباله باشند. مقیاس‌های اندازه‌گیری تشابه یا فاصله، کاربردهای مختلفی دارند. همچنین کاربر ممکن است علاقه‌مند باشد که k تا از شبیه‌ترین سریهای زمانی که با فاصله ε ازسری مورد جستجو قرار دارند را بیابد. شاخص‌گذاری سریهای زمانی در دو حالت مورد بررسی قرار می‌گیرد که در حالت اول تابع فاصله متریک و در حالت بعدی تابع فاصله غیر متریک است. در ادامه به‌طور خلاصه به‌بررسی هر یک از این حالات پرداخته می‌شود.

شاخص‌گذاری سریهای زمانی با تابع فاصله متریک

شاخص‌گذاری علاوه بر اینکه موارد مشابه را درسازمان داده‌ها گردآوری می‌کند، امکان هرس داده‌های غیر مرتبط را نیز فراهم می‌نماید. هرس کردن کاملاً به متریک بودن تابع فاصله بستگی دارد.

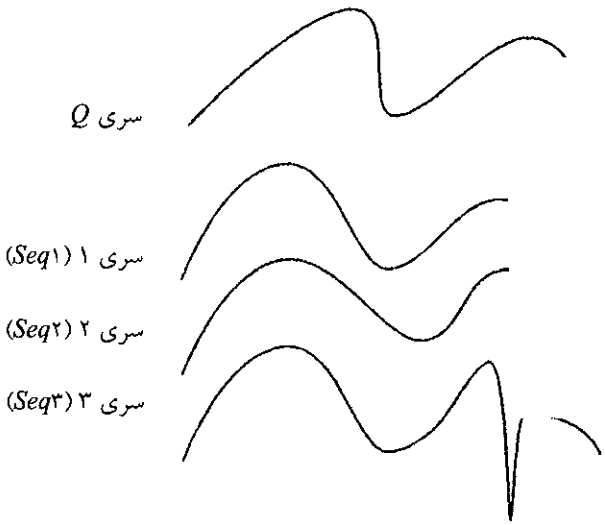
یک تابع فاصله $d(X,Y)$ بین دو شیء X و Y متریک است اگر دارای شرایط زیر باشد:

مثبت بودن $d(X,Y) \geq ۰$ و $d(X,Y) = ۰$ اگر $X=Y$

تقارن $d(X,Y) = d(Y, X)$

نامساوی مثلثی $d(X,Y) + d(Y,Z) \geq d(X,Z)$

فاصله‌اقلیدسی یک تابع فاصله متریک است. قدرت هرس کردن چنین تابعی در مثال زیر نشان داده شده است.



شکل ۸-۱۷) قدرت هرس نامساوی مثلثی

مثال: فرض کنید مجموعه دنباله‌های $S = \{Seq۱, Seq۲, Seq۳\}$ در شکل (۸-۱۷) داده شده است. همچنین فرض کنید فاصله سه دنباله را با روش فواصل جفتی محاسبه و در جدول (۸-۱) مرتب کرده‌ایم.

جدول ۸-۱) فواصل جفتی

	سری ۱	سری ۲	سری ۳
سری ۱	۰	۲۰	۱۱۰
سری ۲	۲۰	۰	۹۰
سری ۳	۱۱۰	۹۰	۰

برای پیدا کردن شبیه‌ترین دنباله به دنباله Q از بین سه دنباله فوق بهترین روش، تصویر دنباله‌هاست. در این روش فاصله همه آنها با Q محاسبه شده و یکی از آنها که کوتاهترین فاصله را با Q دارد، انتخاب می‌شود. اگر تابع فاصله، نامساوی مثلثی را برآورده کند و داشته باشیم $D(Q,Seq۱)=۲۰$ و $D(Q,Seq۲)=۱۳۰$ آنگاه به این علت که:

$$D(Q,Seq۳) \geq D(Q,Seq۲) - D(Seq۲,Seq۳) \rightarrow D(Q,Seq۳) \geq ۱۳۰ - ۹۰ = ٤۰$$

ما می‌توانیم به‌راحتی Seq ۳ را حذف کنیم زیرا راه حل بهتری را پیشنهاد نمی‌کند. برای بهبود بیشتر باید از یک شاخص چند بعدی استفاده کنیم.

شاخصهای چند بعدی و پیدا کردن نزدیک‌ترین همسایگی‌ها به شکلی مؤثر در فضایی با ابعاد بالا، توجه بسیاری از متخصصین را در علوم کامپیوتری و تحقیقات بانکهای اطلاعاتی به خود جلب کرده است. یک روش ساده شاخص‌گذاری، درنظرگرفتن سری زمانی با طول n به‌عنوان یک نقطه n بعدی است. ما می‌توانیم هر سری زمانی را به‌عنوان یک نقطه در ساختار n بعدی *R-tree* ذخیره کنیم. برای پیدا کردن نزدیک‌ترین همسایگی، با تبدیل هر سری زمانی به یک نقطه n بعدی و استفاده از ساختار شاخص‌گذاری مانند استفاده از *R-tree* به جستجوی نزدیک‌ترین همسایگی می‌پردازیم.

متأسفانه این ایده در عمل به‌خوبی کار نمی‌کند زیرا طول بلند سریهای زمانی معمولاً نقاطی با ابعاد بسیار بالا می‌سازد. هنگامی‌که ابعاد زیاد شود، کارآیی ساختارهای شاخص‌گذاری مختلف به تدریج کاهش می‌یابد.

شاخص گذاری تشابه سریهای زمانی بازگشتی با تابع فاصله غیر متریک

توابع فاصله که نسبت به داده‌های بسیارمغشوش، مقاوم هستند، معمولاً نامساوی مثلثی را نقض می‌کنند. چنین توابعی، همه بخشها در سری زمانی را یکسان در نظر نمی‌گیرند. اگر چه

این رفتار مفید است، زیرا مدل صحیح‌تری از ادراک بشری را نشان می‌دهد. زمانی‌که مردم هر نوع داده‌ای (تصویر، سری زمانی و. . .) را مقایسه می‌کنند، بیشتر روی قسمتهایی که شبیه هستند، تمرکز کرده و توجه کمتری به قسمتهایی که شبیه نیستند، می‌کنند. فواصل غیرمتریک امروزه در بسیاری از دامنه‌ها به‌کار گرفته می‌شوند، مانند تطابق رشته (*DNA*)، فیلتر کردن مشارکتی (هنگامی‌که مشتری با الگوی ازپیش ذخیره شده مشتریان منطبق شود) و بازیابی تصاویر مشابه از بانکهای اطلاعاتی. علاوه بر آن تحقیقات علم روانشناسی مطرح می‌کند که قضاوت‌های مشابه بشری نیز غیرمتریک هستند. به‌علاوه برای توابع فاصله غیرمتریک، حالت‌های زیر مطرح می‌شود:

- سریهای زمانی در نرخ‌های نمونه‌گیری متفاوت یا سرعت‌های مختلف جمع‌آوری می‌شود. سریهای زمانی به‌دست آمده نتایج نمونه‌گیری در فواصل زمانی ثابت را تضمین نمی‌کنند. گیرنده‌های جمع‌آوری‌کننده داده، ممکن است برای یک دوره زمانی خاص، یکسان عمل کنند و به نرخ‌های نمونه گیری متناقض منتهی شوند. علاوه برآن وقتی دوسری زمانی دقیقاً در یک جهت حرکت می‌کنند، ولی یکی با سرعتی دو برابر نسبت به دیگری در حرکت است، نتیجه به احتمال زیاد، فاصله‌اقلیدسی بسیار بزرگی است.
- سریهای زمانی شامل نقاط پرت هستند. نقاط پرت به دوصورت تعریف می‌شوند. یکی رخداد غیر عادی در برگیرنده جمع‌آوری‌کننده داده و دیگری واکنش به خطای بشری می‌باشد.
- سریهای زمانی طول‌های مختلفی دارند. فاصله‌اقلیدسی، مربوط به سریهای زمانی با طول یکسان است. هنگامی‌که طول‌ها متفاوت هستند، ما باید تصمیم بگیریم که آیا می‌خواهیم سری طولانی‌تر را کوتاه کنیم یاسری کوتاه‌تر را باجملات صفر لایه‌گذاری کنیم.

مثالهایی از چنین مترهای فاصله، توابع فاصله *LCSS* و *DTW* می‌باشند. مشاهده اینکه هیچ‌کدام از آن دو متریک نیستند ساده است.

مثال: سه دنباله زیر را داریم:

$$Seq۱ = (۰,۱,۱,۰,۰,۱,۰,۰)$$
$$Seq۲ = (۰,۰,۰,۱,۱,۰,۰,۰)$$
$$Seq۳ = (۰,۱,۰,۰,۱,۰)$$

$$DTW(Q,C) = min\left\{\sqrt{\sum_{k=۱}^{K} w_k} \Big/ K\right\}$$

$$LCSS(X,Y) = (m+n-۲l)/(m+n)$$

	DTW	LCSS
(seq۱ , seq۲) Dis	۱/۰	۰/۲۵۰
(seq۱ , seq۳) Dis	۰	۰/۱۴۳

		۰	۱	۱	۰	۰	۱	۰	۰
	۰	۰	۰	۰	۰	۰	۰	۰	۰
۰	۰	۰	۱	۱	۰	۰	۱	۰	۰
۰	۰	۰	۱	۲	۰	۰	۱	۰	۰
۰	۰	۰	۱	۲	۰	۰	۱	۰	۰
۱	۰	۱	۰	۰	۱	۱	۰	۱	۱
۱	۰	۱	۰	۰	۱	۲	۰	۱	۲
۰	۰	۰	۱	۱	۰	۰	۱	۰	۰
۰	۰	۰	۱	۲	۰	۰	۱	۰	۰
۰	۰	۰	۱	۲	۰	۰	۱	۰	۰

$Dis(seq۲, seq۳)$ ۰/۱۱۱ ۰/۲۸۶

$DTW(Seq۱, Seq۲) = ۱/۱۰$

		۰	۱	۱	۰	۰	۱	۰	۰
	۰	۰	۰	۰	۰	۰	۰	۰	۰
۰	۰	۰	۱	۱	۰	۰	۱	۰	۰
۱	۰	۱	۰	۰	۱	۱	۰	۱	۱
۰	۰	۰	۱	۱	۰	۰	۱	۰	۰
۰	۰	۰	۱	۲	۰	۰	۱	۰	۰
۱	۰	۱	۰	۰	۱	۱	۰	۱	۱
۰	۰	۰	۱	۱	۰	۰	۱	۰	۰

$DTW(Seq۱, Seq۳) = ۰$

		۰	۰	۰	۱	۱	۰	۰	۰
	۰	۰	۰	۰	۰	۰	۰	۰	۰
۰	۰	۱	۱	۱	۱	۱	۱	۱	۱
۱	۰	۱	۱	۱	۲	۲	۲	۲	۲
۱	۰	۱	۱	۱	۲	۳	۳	۳	۳
۰	۰	۱	۲	۳	۳	۳	٤	٤	٤
۰	۰	۱	۲	۳	۳	۳	٤	۵	۵
۱	۰	۱	۲	۳	٤	٤	٤	۵	۵
۰	۰	۱	۲	۳	٤	٤	۵	۵	٦
۰	۰	۱	۲	۳	٤	٤	۵	٦	٦

$$DTW(Seq۳, Seq۲) = ۱/۹$$

		۰	۰	۰	۱	۱	۱	۰	۰
	۰	۰	۰	۰	۰	۰	۰	۰	۰
۰	۰	۰	۰	۰	۱	۱	۰	۰	۰
۱	۰	۱	۱	۱	۰	۰	۱	۱	۱
۰	۰	۰	۰	۰	۱	۱	۰	۰	۰
۰	۰	۰	۰	۰	۱	۲	۰	۰	۰
۱	۰	۱	۱	۱	۰	۰	۱	۱	۱
۰	۰	۰	۰	۰	۱	۱	۰	۰	۰

$$LCSS(Seq۱, Seq۲) = (۸ + ۸ - ۲ \times ٦)/(۸ + ۸) = ۱/٤$$

		۰	۱	۰	۰	۱	۰
۰	۰	۰	۰	۰	۰	۰	۰
۰	۰	۱	۱	۱	۱	۱	۱
۰	۰	۱	۱	۲	۲	۲	۲
۱	۰	۱	۱	۲	۳	۳	۳
۱	۰	۱	۲	۲	۳	٤	٤
۰	۰	۱	۲	۳	۳	٤	٤
۰	۰	۱	۲	۳	٤	٤	۵
۰	۰	۱	۲	۳	٤	٤	۵
	۰	۱	۲	۳	٤	٤	۵

$$LCSS(Seq,۳Seq۲) = (۸ + ٦ - ۲ \times ۵)/(۸ + ٦) = ٤/۱٤$$

		۰	۱	۰	۰	۱	۰
	۰	۰	۰	۰	۰	۰	۰
۰	۰	۱	۱	۱	۱	۱	۱
۱	۰	۱	۲	۲	۲	۲	۲
۱	۰	۱	۲	۲	۲	۳	۳
۰	۰	۱	۲	۳	۳	۳	٤
۰	۰	۱	۲	۳	٤	٤	٤
۱	۰	۱	۲	۳	٤	۵	۵
۰	۰	۱	۲	۳	٤	۵	٦
۰	۰	۱	۲	۳	٤	۵	٦

$$LCSS(Seq۱, Seq۳) = (۸+٦-۲\times٦)/(۸+٦) = ۱/۷$$

شکل ۸-۱۸) قدم‌های الگوریتم

$$DTW(Seq۳, Seq۲) = ۱/۹$$
$$LCSS(Seq۱, Seq۲) = (۸+۸-۲\times٦)/(۸+۸) = ۱/٤$$
$$LCSS(Seq۳, Seq۲) = (۸+٦-۲\times۵)/(۸+٦) = ٤/۱٤$$
$$LCSS(Seq۱, Seq۳) = (۸+٦-۲\times٦)/(۸+٦) = ۱/۷$$

برای شاخص گذاری مناسب سریهای زمانی براساس مفهوم تشابه، نیاز به استفاده از روشهای کاهش بعد می‌باشد. در ادامه به بررسی ضرورت استفاده از روشهای کاهش بعد و تبدیل داده‌ها پرداخته می‌شود.

روشهای تبدیل و کاهش داده

به سبب اندازه بسیار زیاد و ابعاد بالای داده‌هایی به شکل سری‌های زمانی، روشهای کاهش داده به‌عنوان اولین قدم در تحلیل سریهای زمانی به‌کار می‌رود. روشهای کاهش داده نه تنها منجر به اشغال فضای کمتر می‌شود، بلکه سرعت پردازش داده‌ها را نیز افزایش می‌دهد. همان‌طورکه در فصل دوم اشاره شد، مهم‌ترین استراتژی برای کاهش داده، انتخاب زیرمجموعه‌ای از ویژگیها می‌باشد که در آن ویژگیهای نامرتبط و اضافی این داده‌ها در نظر گرفته نمی‌شود. علاوه بر این روشهای کاهش بعد و کاهش حجم داده‌ها (نظیر نمونه گیری، خوشه‌بندی و. . .) نیز از دیگر ویژگیهای مؤثر و کارآمد برای کاهش مقادیر عظیم داده‌ها می‌باشد. به‌دلیل اینکه سریهای زمانی به‌عنوان یک نوع از داده‌های با حجم بالا مد نظر قرار می‌گیرند و هر نقطه بر حسب زمان

به‌عنوان یک بعد در نظر گرفته می‌شود، کاهش بعد یکی از مهم‌ترین مباحث مرتبط با تحلیل سریهای زمانی است. یکی از دلایل توجه روزافزون به کاهش بعد، در تحلیل و داده‌کاوی سریهای زمانی، کاهش پیچیدگی محاسبات در اثر کاهش حجم و بعد داده‌ها می‌باشد [٦].

ایده کلیدی شاخص‌گذاری مؤثر سریهای زمانی، کاهش بُعد فضاست. برای کاهش ابعاد فضا، نمونه n بعدی را که نشان‌دهنده سری زمانی است، در یک فضای k بعدی ($k<<n$) تصویر می‌کنیم، به‌طوری‌که فاصله‌ها تاجای ممکن بدون تغییر باقی بمانند. سپس می‌توانیم از یک روش شاخص‌گذاری روی فضای جدید که ابعاد کمتری دارد، استفاده کنیم. چارچوب عمومی‌روش *GEMINI*، برای شاخص‌گذاری سریهای زمانی، با روشهای کاهش بعد از گامهای زیر تبعیت می‌کند.

- مجموعه سریهای زمانی را بر روی یک فضای کاهش بعد یافته، نگاشت می‌کنیم.
- از یک روش شاخص‌گذاری برای شاخص‌گذاری فضای جدید، استفاده می‌کنیم.
- برای سری زمانی مورد جستجویQ، دنباله Q را بر روی فضای جدید نگاشت کرده و نزدیک‌ترین همسایگی‌ها به Q را در فضای جدید، با استفاده از شاخص‌گذاری پیدا می‌کنیم. سپس فواصل واقعی را محاسبه کرده و نزدیک‌ترین را انتخاب می‌کنیم.

روشهای کاهش بعد متعددی در تحلیل سریهای زمانی مورد استفاده قرار می‌گیرند که از آن جمله می‌توان به تبدیل فوریه گسسته، تبدیل موجک گسسته، تجزیه مقدار منفرد بر مبنای تحلیل مؤلفه‌های اصلی و تصویرکردن تصادفی اشاره کرد. این فنون در فصل دوم، قسمت کاهش بعد توضیح داده شده‌اند.

تبدیلات *SVD* دارای مزیت کاهش بعد بهینه تصاویر خطی می‌باشد. یعنی بهترین حفظ را از میانگین مربع خطا بین تصاویر اصلی و تصاویر تقریبی انجام می‌دهد. اگر چه محاسبه آن در مقایسه با روشهای دیگر دشوار است، مخصوصا اگر سریهای زمانی خیلی طولانی باشند. علاوه بر این، این روش برای شاخص‌گذاری زیر دنباله‌ها کاربرد ندارد.

تبدیلات گسسته فوریه، طیف فرکانس یک سیگنال یک بعدی را توصیف می‌کند. روش *DFT* به‌عنوان یک روش کاهش بعد برای سریهای زمانی ارائه شده است.

تجزیه موجک[1] که سریهای زمانی را به شکل مجموع توابع اولیه نشان می‌دهد، مشابه روش *DFT* می‌باشد. ساختار روش *WD* با *DFT* فرق دارد، زیرا تأثیر ضرایب مختلف در آن بیشتر در زمان متمرکز شده‌اند تا در فرکانس. فواید *WD* آن است که سریهای زمانی تبدیل شده در همان دامنه (دامنه موقت) باقی می‌ماند و الگوریتم کارآیی با پیچیدگی *O(n)* برای محاسبه تبدیلات وجود دارد. معایب آن، مشابه روش *DFT* است.

تصویر کردن تصادفی یک روش کاهش بعد عمومی است که درسال ۱۹۹۸ ارائه ودرسال ۲۰۰۱ برای حوزه سریهای زمانی به کار گرفته شد.

استفاده از مقیاس‌بندی چند بعدی برای شاخص‌گذاری سریهای زمانی، دشوار است. روش *MDS*، نگاشت یک مجموعه از سریهای زمانی به نقاط k بعدی است (که k یک مقدار کوچک است). برای پاسخ‌گویی به جستجوی تشابه، باید قادر باشیم سری زمانی قابل جستجو را روی فضای k بعدی نگاشت کنیم. به علت نوع روشی که الگوریتم *MDS* برای یافتن جستجوهای مورد نظر استفاده می‌کند، باید فواصل همه سریهای زمانی داخل مجموعه آن را بیابیم و این عملیاتی خطی است.

نگاشت سریع، تخمینی بوده و روشی بسیار شبیه به روش مقیاس‌دهی چندبعدی است. برای بررسی مبسوط روشهای کاهش بعد به فصل پیش‌پردازش، بخش کاهش بعد رجوع کنید.

تخمین قطعه‌ای خط[2]

به جای استفاده از *DFT* برای تقریب یک سری زمانی می‌توانیم از تقریب چندجمله‌ای استفاده کنیم. اگر چه با استفاده از یک روش خطی، یک روش عمومی برای تعریف مقیاس اندازه‌گیری غیراقلیدسی به دست می‌آید، ولی هیچ روش شاخص‌گذاری عمومی قابل قبولی شناخته نشده است. تحقیقات اخیر نشان می‌دهند که شاخص‌گذاری این سریهای زمانی درصورتی‌که سریهای زمانی با تابع ثابت قطعه‌ای تقریب زده شود، امکان‌پذیر است. ایده اصلی این روش، کاهش بعد از طریق تقسیم سریهای زمانی به k قطعه هم‌اندازه می‌باشد. مقدار میانگین داده‌هایی که در هر قطعه می‌افتند محاسبه شده و برداری از این مقادیر، نمایش داده‌های کاهش

---

[1]- WD

[2]- Line Segment Approximation

بعد یافته می‌باشد. این نوع نمایش داده‌ها، تخمینهای خطی- قطعه‌وار نامیده می‌شود. این نوع ارائه، نسبت به روش DFT مزایایی دارد. مثلاً تبدیلات می‌تواند در زمانهای خطی اجرا شود. مهم‌تر از آن، این روش گستره‌ای از مقیاس‌های اندازه‌گیری، مانند نرم L_p گسسته، DTW و جستجوی اقلیدسی وزندار را پشتیبانی می‌کنند. این روش مانند WD است، هنگامی‌که K، ضریب ۲ داشته و میانگین برای تعریف زیردنباله‌ها استفاده شود.

منابع

۱) فاطمی قمی م. ت. (۱۳۷۵) «پیش‌بینی وتجزیه وتحلیل سریهای زمانی"، نشردانش امروز .

۲) نیرومند ح. ، بزرگ‌نیا ا. (۱۳۷۲) "مقدمه‌ای برتحلیل سریهای زمانی"، نشردانشگاه فردوسی مشهد .

۳) نیرومند، حسینعلی، ۱۳۷۱، تجزیه وتحلیل سریهای زمانی، نشردانشگاه فردوسی مشهد

٤) ابریشمی، حمید، ۱۳۷۳، اقتصادسنجی کاربردی

5) Ye N. (2003) "The hand book of data mining"

6) Han. J, Kamber. M,(2006) "data mining concepts and techniques"

7) Chu K. W. ,Wong M. H. ,(1998) "fast time series searching with scaling and shifting"

8) Keogh. E, Pazzani M. (2000) "Scaling up dynamic time warping for datamining application",proceeding of the sixth acm sigkdd conference on knowledge discovery and data mining

9) Keogh E. , Pazzani. M,(2001) "derivative dynamic time warping"

10) Keogh E. et al (2004) "Exact indexing of dynamic time warping"

11) Banerjee A. , Ghosh J. (2000) "clickstream clustering using weighted longest common subsequences"

12) Faloutsos C. , Lin K. (1995) "Fast map"

13) Smith L. I. (2002) "A tutorial on principal components analysis"

فصل نهم

تحلیل شبکه‌های اجتماعی

در دهه‌های اخیر، نظریه شبکه‌های اجتماعی (که در آن رابطه میان موجودیت‌ها [1] به صورت پیوندهای درون یک گراف نشان داده می‌شوند) توجهات روز افزونی را به خود جلب کرده است. بدین ترتیب تحلیل شبکه‌های اجتماعی از دید داده‌کاوی، تحلیل پیوندها یا پیوندکاوی نیز نامیده می‌شود. در این بخش، ما ابتدا مفهوم شبکه‌های اجتماعی را مطرح نموده و به مطالعه ویژگی‌های شبکه‌های اجتماعی می‌پردازیم. سپس نگاهی به وظایف و چالشهای موجود در پیوندکاوی خواهیم داشت و در نهایت چند نمونه از شبکه‌های اجتماعی را مورد کاوش قرار می‌دهیم.

۹-۱-تعریف شبکه اجتماعی

از دید داده‌کاوی، شبکه اجتماعی مجموعه داده‌های ناهمگن[2] (نامتجانس) و چندرابطه‌ای[3] است که توسط یک گراف نمایش داده می‌شود. نوعاً گرافها بسیار بزرگ و متشکل از گره‌هایی معادل اشیاء و یالهایی[4] معادل پیوندها (که معرف رابطه یا تعامل بین اشیا می‌باشند) هستند. هم

---

[1]- Entity

[2]- Heterogeneous

[3]- Multirelational

[4]- Edges

گره‌ها و هم پیوندها ویژگیهایی دارند. اشیا ممکن است برچسب دسته داشته باشند. پیوندها می‌توانند یک‌طرفه باشند و لزومی ندارد که دوحالته باشند.

لازم نیست شبکه‌های اجتماعی زمینه‌ای اجتماعی داشته باشند. مثالهای واقعی بسیاری از شبکه‌های اجتماعی تکنولوژیکی، تجاری، اقتصادی و زیستی وجود دارد: شبکه‌های توزیع نیروی الکتریسیته، گرافهای تماسهای تلفنی، انتشار ویروسهای کامپیوتری، شبکه گسترده جهانی، شبکه‌های استناد و هم‌نویسندگی دانشمندان. مثال دیگر عبارت است از شبکه‌های مشتریان در مواقعی که توصیه محصول بر اساس اولویت سایر مشتریان صورت می‌گیرد. در بیولوژی، نمونه‌ها طیف وسیعی را از شبکه‌های شیوع بیماری، شبکه‌های متابولیک و سلولی، شبکه غذایی تا شبکه عصبی کرم (تنها موجودی که شبکه عصبی‌اش کاملاً نگاشت شده است) دربرمی‌گیرند. تبادل پیامهای پست الکترونیک بین شرکتها، گروه‌های خبری، اتاقهای گپ، شبکه دوستی، هم‌پوشانی هیئت مدیره‌های شرکتهای بزرگ آمریکایی و غیره نمونه‌هایی از حوزه جامعه‌شناسی هستند.

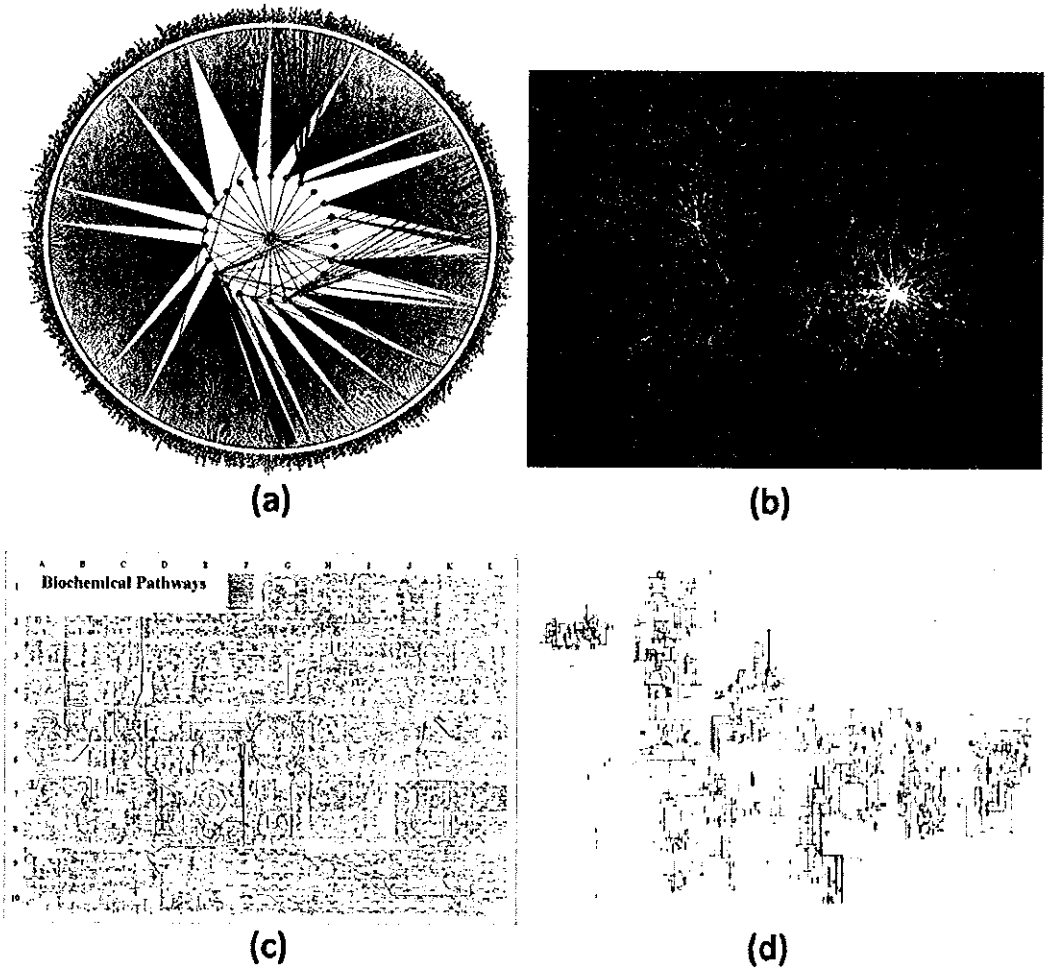
«شبکه‌های (اجتماعی) دنیای کوچک[1]» اخیراً نظرات بسیاری را به خود معطوف داشته است. آنها مفهوم «دنیای کوچک» را منعکس می‌نمایند که در اصل بر شبکه‌های بین افراد تمرکز دارد. همان عبارتی‌که نشان دهنده تعجب زیاد اولیه میان دو غریبه است، وقتی در می‌یابند که به‌صورت غیر مستقیم از طریق نفر سومی که هر دو با وی سابقه آشنایی دارند، با هم رابطه دارند: «چه دنیای کوچکی!». در ۱۹٦۷ استنلی میلگرام[2] جامعه‌شناس هاروارد و همکارانش آزمایشی را ترتیب دادند که در آن از مردم کانزاس و نبراسکا درخواست شده بود تا نامه‌هایی را به دست افرادی غریبه در بوستون برسانند، بدین صورت‌که این نامه‌ها را به دوستانی ارسال کنند که گمان می‌کنند ممکن است آن افراد غریبه را بشناسند. نیمی از نامه‌ها به‌طور موفقیت‌آمیزی از طریق کمتر از ۵ واسطه به مقصد رسیدند. مطالعات دیگری که میلگرام و سایرین در شهرهای دیگر صورت دادند، نشان داد که ظاهراً به‌طور عمومی «شش سطح جدایی[3]» بین هر دو نفر در جهان وجود دارد. نمونه‌هایی از شبکه‌های دنیای کوچک در شکل (۹-۱) نشان داده شده است.

۱- Small World

۲- Stanley Milgram

۳- Six Degrees of Separation

شبکه‌های دنیای کوچک با ویژگی دارا بودن درجه بالایی از خوشه‌پذیری محلی برای نسبت کوچکی از گره‌ها شناخته می‌شوند (به‌عنوان مثال این گره‌ها با دیگری به هم وصل می‌شوند)، که در عین حال بیش از چند سطح از سایر گره‌ها باقیمانده جدا نیستند. این باور وجود دارد که بسیاری از شبکه‌های زیستی، اجتماعی، و فیزیکی ساخته دست بشر این ویژگی‌های دنیای کوچک را به معرض نمایش می‌گذارند. این ویژگیها بعداً توضیح داده شده و مدلسازی می‌شوند.



شکل ۹-۱) مثالهای شبکه اجتماعی در دنیای حقیقی: a) هم‌نویسندگی علمی، b) صفحات مرتبط در بخشی از اینترنت، c) مسیر بیوشیمی، d) شبکه قدرت الکتریسیته نیویورک

«به‌طور کلی چرا این همه علاقه به شبکه‌های دنیای کوچک و شبکه‌های اجتماعی وجود دارد؟ چه فایده‌ای در مشخص نمودن ویژگی‌های شبکه‌ها و کاوش آنها به منظور بیشتر آموختن از ساختارشان وجود دارد؟» علت آن است که ساختار همواره بر عملکرد مؤثر است. به‌عنوان مثال توپولوژی شبکه‌های اجتماعی بر شیوع بیماریهای مسری در یک جمعیت ساختار یافته تأثیر دارد. توپولوژی شبکه نیرو بر ثبات و انسجام انتقال نیرو مؤثر است.

به‌طور نمونه، یک نارسایی مرتبط در تاریخ ١٤ آگوست ۲۰۰۳ در کلیولند واقع در اوهایو در سیستم شبکه ایجاد شده. این نارسایی منجر به از کار افتادن کارخانجات نیروی هسته‌ای در ایالت نیویورک و اوهایو گردید و باعث خاموشی گسترده در بخشهای زیادی از شمال شرقی ایالات متحده و جنوب شرقی کانادا شد که حدود ۵۰ میلیون نفر را در بر می‌گرفت.

توجه به شبکه‌ها بخشی از مطالعات وسیع‌تری است که در توصیف کامل و دقیق «سیستمهای پیچیده[1]» انجام می‌شود. قبلا شبکه‌هایی که برای مطالعات تجربی در دسترس بودند، محدود و کوچک بودند و اطلاعات ناچیزی از گره‌های آنها وجود داشت. باید از اینترنت ممنون باشیم که در حال حاضر مقادیر عظیمی از داده‌های مرتبط با شبکه‌های اجتماعی بسیار بزرگ را در دسترس ما قرار داده است. این شبکه‌ها نوعاً دهها هزار تا میلیونها گره را دربرمی‌گیرد و اغلب اطلاعات وسیعی در سطح هر گره موجود می‌باشد.

دسترسی به کامپیوترهای قدرتمند نیز بررسی ساختار شبکه‌ها را ممکن ساخته است. مطالعه شبکه‌های اجتماعی می‌تواند به ما در دسترسی بهتر و آسانتر به سایر مردم دنیا کمک کند. به‌علاوه، مطالعه دنیای کوچک، با توجه به جدایی نسبتا کم بین گره‌هایشان، می‌تواند به ما در طراحی شبکه‌هایی که انتقال کارای اطلاعات یا سایر منابع را تسهیل می‌کنند، یاری رساند، بدون آنکه مجبور باشیم از شبکه‌ای با تعداد زیادی رابطه زائد استفاده کنیم. به‌عنوان مثال می‌تواند به ما در طراحی موتورهای جستجو هوشمندتری در وب کمک کند، به‌طوریکه در پاسخ به یک پرس‌وجو، وب سایتهای مرتبطی که کمترین میزان جدایی از وب سایت اولیه را دارند، بیابند.

---

[1]- Complex Systems

۹-۲- ویژگیهای شبکه‌های اجتماعی

همان‌طورکه در قسمت قبل توضیح داده شد، دانستن ویژگیهای شبکه‌های دنیای کوچک در موقعیتهای بسیاری مفید است. می‌توان مدلهای مولد گراف را که دارای همین ویژگیها باشند، ایجاد نمود. تا قادر به پاسخگویی به سؤالات «اگر – آنگاه» و پیش‌بینی چگونگی یک شبکه در آینده باشند. به‌عنوان مثال در مورد اینترنت، می‌پرسیم «اگر تعداد گره‌ها در اینترنت دو برابر شود، آنگاه اینترنت چگونه به نظر می‌رسد؟» و «تعداد یالها چندتا خواهد بود؟». چنانچه فرضیه‌ای با ویژگیهای عموماً پذیرفته شده، تناقض داشته باشد، پرچمی در برابر مشکوک بودن فرضیه برمی‌افرازد.

این موارد به کشف ناهنجاریها در گرافهای موجود کمک می‌نماید که ممکن است تقلب، هرزنگاری[1] یا حمله‌های *Ddos* را مشخص کند. به‌علاوه مدلهای مولد گراف می‌تواند برای شبیه سازی مواردی که گرافهای واقعی بیش از اندازه بزرگ هستند و بدین لحاظ جمع‌آوریشان غیرممکن است (مثل شبکه بسیار بزرگی از رابطه دوستی) کمک کند. در این قسمت، ویژگیهای اساسی شبکه‌های اجتماعی به همراه مدلی برای تولید گراف بررسی می‌شود.چه ویژگیهایی شبکه‌های اجتماعی را مشخص می‌کند؟ اکثر مطالعات، امتیاز گره‌ها[2] را بررسی کرده‌اند که همان تعداد یالهای منتهی به هر گره است و یا فاصله بین هر دو گره را که با محاسبه طول کوتاهترین مسیر تعیین می‌شود. سایر محاسبات فاصله میان دو گره شامل فاصله متوسط[3] و قطر مؤثر[4] (به‌عنوان مثال، حداقل فاصله d به‌طوری‌که لااقل برای ۹۰٪ گره‌ها در دسترس، طول مسیر حداکثر d باشد) می‌باشد.

شبکه‌های اجتماعی به ندرت ایستا هستند. نمایش گراف آنها با افزایش یا حذف گره‌ها و یالها در طول زمان، تکامل می‌یابد. به‌طورکلی، شبکه‌های اجتماعی پدیده‌های زیر را نشان می‌دهند:

---

[1]- Spam

[2]- Nodes Degrees

[3]- Average Distance

[4]- Effective Diameter

قانون تراکم توانی[1]: در گذشته، باور بر این بود که با تکامل یک شبکه، امتیاز گره‌ها به صورت خطی نسبت به تعداد گره‌ها افزایش می‌یابد. این باور با نام فرض میانگین امتیاز ثابت[2] شناخته می‌شد. با این وجود، آزمایشات گسترده‌ای نشان داده‌اند که برعکس، با افزایش متوسط امتیاز در طول زمان، شبکه‌ها به‌طور فزاینده‌ای چگال می‌شوند (بدین ترتیب، تعداد یالها به نسبت تعداد گره‌ها به‌صورت فوق خطی[3] افزایش می‌یابند). این متراکم شدن از قانون تراکم توانی (یا قانون رشد توانی[4]) پیروی می‌کند و عبارتست از:

$$e(t) \propto n(t)^a \qquad (۱-۹)$$

که در آن $e(t)$ و $n(t)$ به ترتیب نمایانگر تعداد یالها و گره‌های گراف در زمان t هستند و توان a عموما بین ۱و۲ قرار می‌گیرد. توجه داشته باشید که چنانچه $a=۱$ باشد، معادل میانگین امتیاز ثابت در طول زمان است، درحالی‌که $a=۲$ بیانگر گرافی به شدت متراکم است که در آن هر گره توسط یالهایش به نسبت ثابتی از کل گره‌ها، متصل است.

جمع شدن قطر[5]: به‌طور تجربی اثبات شده که قطر مؤثر با رشد شبکه کاهش می‌یابد. این با باور اولیه‌ای که معتقد بود قطر به آهستگی به‌صورت تابعی از اندازه شبکه افزایش می‌یابد، درتناقض است.

برای درک شهودی این مطلب یک شبکه استناد را درنظر بگیرید که در آن گره‌ها مقالات هستند و استناد (ارجاع) از یک مقاله به مقاله‌ای دیگر با یک یال جهت‌دار نمایش داده می‌شود. یالهای خروجی از گره v (نمایانگر مقالاتی که v به آنها ارجاع داده است) در لحظه پیوستن به گراف «منجمد[6]» هستند. متعاقبا به نظر می‌رسد فاصلهٔ در حال کاهش بین دو گره، مقالات بعدی باشد که به‌عنوان «پل[7]» عمل کرده و مقالات اولیه را مورد ارجاع قرار داده‌اند.

۱- Densification Power Law

۲- Constant Average Degree Assumption

۳- Superlinearly

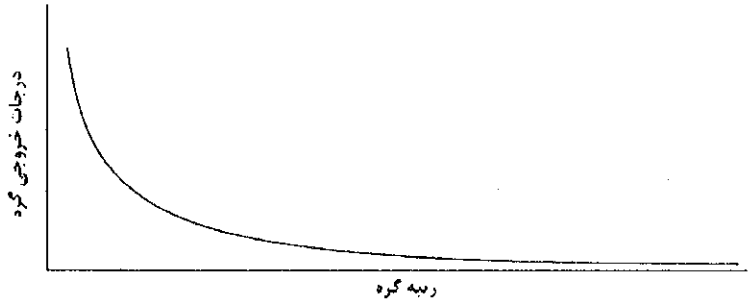
٤- Growth Power Law

٥- Shrinking Diameter

٦- Frozen

۷- Bridge

توزیع‌های دم‌پهن[1] درجات ورودی و خروجی: با توجه به «قانون توان» تعداد امتیازات خروجی یک گره به پیروی از توزیعی دم‌پهن گرایش دارد: $1/n^a$ که در آن n رتبه گره در ترتیب کاهش امتیاز خروجی است و نوعاً $0<a<2$ است (شکل (۹-۲)). هرچه مقدار a کوچک‌تر باشد، دنباله پهن‌تر خواهد بود. این حالت بیانگر الحاق ترجیحی[2] است که در آن هر گره جدید با تعداد ثابتی یالهای خروجی به شبکهٔ موجود وصل می‌شود و از قاعده «ثروتمند، ثروتمندتر می‌شود[3]» پیروی می‌کند. امتیازات ورودی هم از توزیع دم‌پهن پیروی می‌کند که در ضمن نسبت به توزیع امتیازت خروجی چولگی بیشتری دارد.



شکل ۹ – ۲) تعداد درجات خروجی (محور y) یک گره تمایل به پیروی از توزیعی دم‌پهن دارد. رتبه گره (محور x) ترتیب نزولی درجات خروجی گره است.

برای تولید گراف مدل آتش جنگل[4] پیشنهاد شد که این خصوصیات تکامل گراف در طول زمان را دارا است. این مدل بر پایه این نظریه استوار است که گره‌های جدیدی از طریق سوزاندن[5] یالهای موجود به طریق مسری به شبکه متصل شده و از دو پارامتر بهره می‌گیرد: احتمال سوزاندن پیشرو (p) و نسبت سوزاندن پسرو (r) که در ادامه توضیح داده می‌شوند. فرض کنید گره جدید v در زمان t اضافه شود، این گره در طی گامهای زیر به G_t، گرافی که تاکنون ایجاد شده، متصل می‌شود:

---

[1]- Heavy-Tailed

[2]- Preferential Attachment Model

[3]- Rich-Get-Richer

[4]- Forest fire Model

[5]- Burning

گام اول: به‌طور تصادفی یک گره سفیر[1]، w انتخاب می‌کند و یک یال به w متصل می‌کند.

گام دوم: x یال متصل به w را انتخاب می‌کند. x مقداری تصادفی است که دارای توزیع برنولی با میانگین $(1-p)^{-1}$ است. هم یالهای ورودی به w و هم یالهای خروجی از w در نظر گرفته می‌شوند ولی یالهای ورودی را با احتمال r بار کمتر از یالهای خروجی برمی‌گزینند. گره‌هایی را که در سر دیگر یالهای انتخاب شده قرار دارند، $w_1, w_2, ..., w_x$ بنامید.

گام سوم: گره جدید (w) یالهای خروجی را به $w_1, w_2, ..., w_x$ متصل نمود، حال گام ۲ را به‌طور بازگشتی برای هر کدام از گره‌های $w_1, w_2, ..., w_x$ انجام دهید. به‌منظور جلوگیری از افتادن در حلقه، هیچ گره‌ای برای بار دوم نباید انتخاب شود. این فرآیند تا زمانی که خاموش شود ادامه می‌یابد.

برای درک شهودی از مدل، به مثال خودمان از شبکهٔ استناد، باز می‌گردیم. نویسنده مقاله جدید، v، ابتدا به w مراجعه می‌کند. سپس یک زیر مجموعه از منابع w را دنبال می‌کند (که ممکن است پیشرو یا پسرو باشند) و به مقالات $w_1, w_2, ..., w_x$ دست می‌یابد. با ارجاع دادن این مقالات، تجمیع مراجع به‌صورت بازگشتی ادامه می‌یابد.

بسیاری از مدلهای تکامل شبکه بر گرافهای ایستا مبتنی است که ویژگیهای شبکه را بر اساس یک یا تعداد معدودی تصویر آنی از آن با کمی تأکید بر یافتن روندها در طول زمان تعیین می‌کنند. مدل آتش جنگل درحالی‌که به تکامل شبکه در طول زمان توجه دارد، ماهیت بسیاری از مدلهای قبلی را درهم می‌آمیزد. مثلاً خاصیت امتیازهای خروجی دم‌پهن که به طبیعت بازگشتی تشکیل یالها تعلق دارد را نیز در نظر می‌گیرد؛ بدین ترتیب، گره‌های جدید شانس خوبی در سوزاندن بسیاری یالها و لذا ایجاد امتیازات خروجی بزرگ دارند. خاصیت امتیازات ورودی دم‌پهن نیز حفظ می‌شود، زیرا آتش جنگل از قاعده «ثروتمند، ثروتمندتر می‌شود» پیروی می‌کند. گره‌هایی که خیلی زیاد به هم متصل هستند، بدون در نظر گرفتن اینکه گره جدید از کدام گره سفیر آغاز می‌شود، به آسانی به گره‌ای جدید متصل می‌شوند. گونه‌ای از مدل کپی‌کننده[2] نیز مورد نظر قرار گرفته است بدین صورت که یک گره جدید بسیاری از همسایگان گره سفیر خود

---

[1]- Ambassador

[2]- Copying Model

را کپی می‌کند. قانون تراکم توانی نیز تأیید می‌شود: یک گره جدید یالهای بسیاری در نزدیکی اجتماع گره سفیر خود خواهد داشت. مطالعات تجربی خاصیت جمع شدن قطر را تأیید می‌کنند. گره‌هایی که امتیازات خروجی دم‌پهن دارند، ممکن است به‌عنوان «پلهایی» به‌کار روند که قسمتهایی از شبکه را که پیش‌تر ناهمگن بودند، متصل کنند و قطر شبکه را کاهش دهند.

۹-۳- پیوندکاوی[1]: وظایف و چالشها

چگونه می‌توان شبکه‌های اجتماعی را مورد کاوش قرار داد؟ روشهای سنتی یادگیری ماشینی و داده‌کاوی که به‌عنوان ورودی، نمونه‌های تصادفی اشیاء همگنی را از یک رابطهٔ واحد اخذ می‌کنند، ممکن است برای این حالت مناسب نباشند. داده‌های تشکیل‌دهنده شبکه‌های اجتماعی بیشتر ناهمگن، چندرابطه‌ای و نیمه‌ساخت‌یافته هستند. در نتیجه، حوزه جدیدی از پژوهش با نام پیوندکاوی پدید آمده است. پیوندکاوی محل تلاقی پژوهش در شبکه‌های اجتماعی، تحلیل پیوندها، وب‌کاوی و ابرمتنها[2]، گراف‌کاوی، یادگیری رابطه‌ای[3] و برنامه‌ریزی منطقی قیاسی[4] است. پیوندکاوی متضمن مدلهای توصیفی و پیشبینانه است. با درنظر گرفتن پیوندها (رابطه میان اشیاء) اطلاعات بیشتری برای فرآیند کاوش حاصل می‌شود. این امر وظایف جدید بسیاری را به همراه می‌آورد، لیست این وظایف با مثالهایی از حوزه‌های مختلف در زیر فهرست شده‌اند.

دسته‌بندی اشیاء مبتنی بر پیوندها: در روشهای سنتی دسته‌بندی، اشیاء بر اساس ویژگیهایی که آنها را توصیف می‌کردند، دسته‌بندی می‌شدند. در دسته‌بندی مبتنی بر پیوندها، دستهٔ یک شیء نه تنها بر اساس ویژگیهایش، بلکه بر اساس پیوندهایش و ویژگیهای اشیایی که با آنها پیوند دارد، پیش‌بینی می‌شود.

دسته‌بندی صفحات وب مثال بسیار مناسبی از دسته‌بندی مبتنی بر پیوندهاست. در دسته‌بندی صفحات وب، دسته یک صفحه هم بر اساس رخداد واژه‌ها (کلماتی که بر روی آن صفحه واقع

---

[1]- Link Mining

[2]- Hypertext

[3]- Relational Learning

[4]- Inductive Logic Programming

شده‌اند) و هم بر اساس متن لنگر[1] (کلمات ابَرپیوندی[2] یعنی کلماتی که شما هنگام کلیک بر روی یک پیوند، بر روی آنها کلیک می‌کنید) انجام می‌شود و هر دوی آنها به‌عنوان ویژگیهای صفحه به‌کار می‌روند. به‌علاوه، دسته‌بندی بر پایه پیوندهای بین صفحات و سایر ویژگیهای صفحات و پیوندها نیز می‌باشد. در حوزه کتاب‌شناسی[3]، اشیاء ما مقالات، نویسندگان، مؤسسات، مجلات و کنفرانسها هستند. در این حالت وظیفه دسته‌بندی، پیش‌بینی مبحث یک مقاله است که بر اساس رخداد واژه‌ها، مورد ارجاع بودن (سایر مقالاتی که به این مقاله ارجاع داده‌اند)، و استناد کردن (سایر مقالاتی که در این مقاله به آنها استناد شده است) انجام می‌شود. استنادها به‌عنوان پیوند عمل می‌کنند. در مبحث شناخت بیماریهای همه‌گیر، وظیفه دسته‌بندی، پیش‌بینی نوع بیماری یک فرد است که بر اساس ویژگیهای (علائم بیماری) فرد بیمار و ویژگیهای سایر افرادی که فرد بیمار با آنها تماس داشته است (از این افراد با عنوان تماسهای بیمار یاد می‌شود) صورت می‌گیرد.

پیش‌بینی نوع شیء: این پیش‌بینی، نوع یک شیء را بر اساس ویژگیهای خودش و پیوندهایش و ویژگیهای اشیایی که به آن متصلند تعیین می‌کند. در حوزه کتاب‌شناسی ممکن است بخواهیم محل یک مطلب منتشر شده را مشخص کنیم مثلاً کنفرانس یا مجله یا کارگاه. در حوزه ارتباطات، وظیفه مشابهی وجود دارد که پیش‌بینی آن است که آیا یک رابطه از طریق پست الکترونیک است، یا تلفن یا پست؟

پیش‌بینی نوع پیوند: این پیش‌بینی، بر اساس مشخصه‌های اشیاء درگیر در یک پیوند، نوع یا هدف آن پیوند را معلوم می‌سازد. به‌عنوان مثال با در دست داشتن داده‌ها شیوع بیماریهای همه‌گیر، ممکن است بخواهیم پیش‌بینی کنیم آیا دو نفر که یکدیگر را می‌شناسند از اعضای یک خانواده‌اند، یا همکارند، یا صرفاً با یکدیگر آشنا هستند. در نمونه‌ای دیگر ممکن است بخواهیم بدانیم آیا رابطه بین دو مؤلف به‌صورت مشاور-مشورت کننده است؟ با دستیابی به داده‌های صفحات وب می‌توانیم پیش‌بینی کنیم آیا یک پیوند روی یک صفحه، پیوند تبلیغاتی است یا پیوندی مربوط به پیمایش وب؟

---

[1]- Anchor Text

[2]- Hyperlink

[3]- Bibliography

پیش‌بینی وجود پیوند: در پیش‌بینی نوع پیوند ما می‌دانستیم پیوندی بین دو شیء موجود است و می‌خواستیم نوع این پیوند را معلوم کنیم اما در این حالت می‌خواهیم پیش‌بینی کنیم آیا اصلاً بین دو شیء پیوندی وجود دارد؟ مثال: پیش‌بینی اینکه آیا پیوندی بین دو صفحه وب وجود خواهد داشت؟ آیا مقاله‌ای به مقاله دیگری استناد خواهد کرد؟ و یا در علم شناخت بیماریهای همه‌گیر می‌توانیم پیش‌بینی کنیم یک بیمار با چه کسی در تماس بوده است.

تخمین عدد اصلی[1] پیوند: دو شکل تخمین عدد اصلی پیوند وجود دارد. اول، می‌توان تعداد پیوندهای متصل به یک شیء را پیش‌بینی نمود. به‌طور مثال تعیین مأخذیت[2] یک صفحه وب می‌تواند بر اساس تعداد پیوندهایی که به آن منتهی شده‌اند (پیوندهای ورودی) مشخص شود. به‌طور مشابه تعیین تعداد پیوندهای خروجی[3] می‌تواند مبین این باشد که آیا آن صفحه وب به‌عنوان قطب[4] عمل می‌کند (منظور از قطب، یک یا مجموعه‌ای از صفحات وب است که به تعدادی صفحات مأخذ در همان مبحث استناد می‌کنند)، در حوزه کتاب‌شناسی، تعداد استنادات به یک مقاله می‌تواند نشانگر تأثیر آن مقاله باشد (هرچه استنادات به مقاله بیشتر باشد، انتظار می‌رود تأثیرگذاری بیشتری داشته باشد)، در علم بیماریهای همه‌گیر، پیش‌بینی تعداد پیوندهای بین یک بیمار و تماسهایش، نشانگر انتقالهای بالقوه بیماری است.

حالت مشکل‌تر تخمین عدد اصلی پیوند، پیش‌بینی تعداد اشیایی است که در خلال یک مسیر از یک شیء قرار دارند. این مسئله در تخمین تعداد اشیائی که در پاسخ به یک پرس‌وجو باز می‌گردد، حائز اهمیت است. در حوزه صفحات وب، می‌توان تعداد صفحاتی را که از طریق پویش یک سایت بازیابی می‌شود، پیش‌بینی نمود (منظور از پویش، جستجویی خودکار و روشمند در وب است که اساساً به‌منظور تهیه یک کپی از تمام صفحات مشاهده شده است تا برای پردازشهای بعدی یک موتور جستجو استفاده شود). با در نظر گرفتن مسئله استناددهی، می‌توان از تخمین کاردینالیتی پیوندی برای پیش‌بینی تعداد استنادات یک نویسنده خاص در یک مجله خاص بهره برد.

---

[1]- Cardinality

[2]- Authoritativeness

[3]- Out-Links

[4]- Hub

مصالحه[1] اشیاء: در مصالحه اشیاء، وظیفه پیش‌بینی این است که آیا در واقع دو شیء بر اساس ویژگیها و پیوندهایشان یکسان هستند؟ این وظیفه در استخراج اطلاعات[2]، حذف دوباره‌کاری[3]، یکی‌سازی اشیاء و انطباق[4] استنادات امری معمول است و با عنوان عدم‌قطعیت موجودیتها[5] یا ارتباط رکورد[6] نیز شناخته می‌شود. به‌عنوان نمونه: پیش‌بینی اینکه آیا دو وب سایت آینه یکدیگرند، آیا دو استناد واقعاً به یک مقاله استناد می‌کنند، و آیا دو درد آشکار حاصل از بیماری، در واقع یکی هستند؟

کشف گروه‌ها: کشف گروه نوعی خوشه‌بندی است و به پیش‌بینی این امر می‌پردازد که چه هنگام مجموعه‌ای از اشیاء به یک گروه یا خوشه تعلق دارند و این کار را بر اساس ویژگیهای آن اشیاء و ساختار پیوندهایشان انجام می‌دهد. یک حوزه کاربرد آن تعیین اجتماعات وبی[7] است یعنی مجموعه‌ای از صفحات وب که بر یک عنوان یا زمینه[8] خاص متمرکز هستند. کاربرد مشابه، تعیین اجتماعات پژوهشی در حوزه کتاب‌شناسی می‌باشد.

کشف زیر گرافها[9]: تعیین زیر گرافها، زیر گرافهای ویژه‌ای را در درون شبکه می‌یابد و نوعی از جستجوی گرافی است. مثالی از بیولوژی کشف زیر گرافهای متناظر با ساختار پروتئینهاست. در شیمی نیز می‌توان زیر گرافهایی را جستجو کرد که زیر ساختارهای شیمیایی را نمایش می‌دهند.

فراداده‌کاوی: فراداده‌ها، داده‌هایی در مورد داده‌ها هستند. فراداده‌ها، داده‌هایی نیمه‌ساخت‌یافته درباره داده‌های ساخت‌نیافته فراهم می‌کنند که طیف وسیعی از داده‌های وب و متن گرفته تا پایگاههای داده‌های چند رسانه‌ای را در برمی‌گیرد. این وظیفه برای یکپارچه‌سازی داده‌ها در

---

[1]- Reconciliation
[2]- Information Extraction
[3]- Duplication
[4]- Matching
[5]- Identity Uncertainty
[6]- Record Linkage
[7]- Web Communities
[8]- Theme
[9]- Subgraph

بسیاری حوزه‌ها مفید است. فراداده‌کاوی را می‌توان برای نگاشت شماتیک[۱] (مثلاً ویژگی شماره مشتری از یک پایگاه داده به شماره مشتری از پایگاه داده‌ای دیگر نگاشت می‌شود چون هر دو آنها به یک موجودیت اشاره دارد)؛ کشف شماتیک[۲] (از داده‌های نیمه‌ساخت‌یافته طرحهایی ایجاد می‌کند)؛ و تشکیل مجدد شماتیک[۳] (بر اساس فراداده‌های داده‌کاوی شده، طرح را اصلاح می‌کند) به‌کار گرفت. نمونه‌ها شامل این موارد است: انطباق دو منبع کتاب‌شناختی، کشف طرح از داده‌های نیمه‌ساخت‌یافته یا غیرساخت‌یافته روی وب، و نگاشت بین دو هستان‌شناسی[۴] دارویی.

به‌طور خلاصه، استفاده از اطلاعات پیوند بین اشیاء، وظیفه‌ای اضافی برای پیوندکاوی در مقایسه با رویکردهای سنتی کاوش به همراه می‌آورد. به‌کارگیری این وظایف در هر حال چالشهای بسیاری را دربرخواهد داشت. در اینجا چند نمونه از این چالشها را بررسی می‌کنیم.

وابستگیهای منطقی در برابر وابستگیهای آماری: دو نوع وابستگی در ساختارهای گراف قرار دارند: ساختارهای پیوند (مبین رابطه منطقی بین اشیاء) و وابستگیهای احتمالی[۵] (مبین رابطه آماری مثل همبستگی بین ویژگیهای اشیاء درحالتی‌که نوعاً چنین اشیایی منطقاً به هم مرتبطند). کار هم‌زمان با این وابستگیها نیز خود چالشی برای کاوش داده‌های چندرابطه‌ای است که در آن داده‌هایی که مورد کاوش قرار می‌گیرند، در جداول چند لایه‌ای قرار دارند. علاوه بر جستجوهای استاندارد بر روی وابستگیهای احتمالی بین ویژگیها، می‌بایست بر روی رابطه‌های منطقی مختلف ممکن بین اشیاء نیز جستجو صورت گیرد. این امر فضای جستجوی وسیعی را می‌طلبد که یافتن یک مدل ریاضی موجه را مشکل می‌سازد. روشهای توسعه‌یافته در برنامه‌ریزی منطق قیاسی که بر روی ارتباطات منطقی جستجو می‌کند، اینجا می‌تواند مفید واقع شود.

ساخت مشخصه‌ها[۶]: در دسته‌بندی مبتنی بر پیوندها، ما هم به ویژگیهای یک شیء توجه می‌کنیم و هم به ویژگیهای اشیای متصل به آن. به‌علاوه، ممکن است پیوندها هم ویژگیهایی

۱- Schema Mapping

۲- Schema Discovery

۳- Reformulation Schema

۴- Ontologies

۵- Probabilistic Dependencies

۶- Feature Construction

داشته باشند. هدف از ساخت مشخصه‌ها، ایجاد یک صفت منفرد مبین این ویژگیهاست. این امر می‌تواند شامل انتخاب مشخصه‌ها[1] و تجمیع مشخصه‌ها[2] باشد. در انتخاب مشخصه‌ها، فقط مشخصه‌های مهم و متمایزکننده درنظر گرفته می‌شوند. تجمع مشخصه‌ها، چند مجموعه از مقادیر بر روی مجموعه اشیاء مرتبط را می‌گیرد و خلاصه‌ای از آن را باز می‌گرداند. این خلاصه می‌تواند به‌عنوان مثال مُد مجموعه (مقداری که بیشترین تعداد رخداد را دارد)؛ میانگین مجموعه (اگر مقادیر عددی باشد)؛ یا میانه (اگر مقادیر به ترتیب مرتب شده باشند) باشد. برخی اوقات این روش مناسب نیست.

نمونه‌ها[3] در برابر دسته‌ها: این چالش مربوط است به اینکه آیا این مدل صریحاً به تک‌تک نمونه‌ها اشاره دارد یا به دسته‌هایی (طبقه‌های عمومی) از نمونه‌ها. یک مزیت مدل پیشین در این است که می‌تواند برای اتصال تک تک نمونه‌های خاص با احتمال بالا به‌کار رود. یک مزیت مدل آخر این است که می‌تواند برای عمومیت بخشیدن به موقعیتهای جدید با نمونه‌های منفرد مختلف استفاده شود.

دسته‌بندی جمعی و یکی‌سازی جمعی[4]: آموزش دادن[5] یک مدل برای دسته‌بندی را در نظر بگیرید که بر اساس مجموعه‌ای از اشیاء که برچسب یا نام دسته‌شان مشخص است، انجام شود. روشهای دسته‌بندی سنتی تنها به ویژگیهای شیء توجه می‌نمود. فرض کنید پس از آموزش مدل، مجموعه جدیدی از اشیایی که برچسب دسته‌شان معلوم نیست در اختیار داریم. به‌کارگیری مدل برای تعیین برچسب دسته اشیاء جدید، با توجه به همبستگی‌های احتمالی بین اشیاء پیچیده است (برچسب دسته اشیاء به هم مرتبط ممکن است همبسته باشد). بنابراین دسته‌بندی می‌بایست گام تکرار شونده دیگری را هم در بربگیرد که برچسب دسته هر شیء را بر اساس برچسب دسته اشیاء مرتبط با آن تغییر دهد (یا تثبیت کند). در این معنا، دسته‌بندی بیشتر جمعی انجام می‌شود تا مستقلاً.

---

[1]- Feature Selection

[2]- Feature Aggregation

[3]- Instances

[4]- Collective Consolidation

[5]- Training

استفاده مؤثر از داده‌های برچسب خورده و برچسب نخورده: یک استراتژی جدید در یادگیری این است که مخلوطی از داده‌های برچسب خورده و برچسب نخورده را شرکت دهید. داده‌های برچسب نخورده می‌تواند به استنتاج توزیع ویژگیهای اشیاء کمک کند. پیوند بین داده‌های برچسب خورده (داده‌های آموزشی) و برچسب نخورده (داده‌های آزمون) پی به وابستگیهایی می‌برد که می‌تواند به استدلالهای دقیق‌تر کمک کند.

پیش‌بینی پیوندها: یک چالش موجود در پیش‌بینی پیوندها این است که احتمال پیشین یک پیوند خاص بین اشیاء بسیار کم است. رویکردهای مختلف در باب پیش‌بینی پیوندها بر اساس تعدادی معیار برای تحلیل مجاورت گره‌ها در یک شبکه، مطرح شده‌اند. مدلهای احتمالی هم مطرح شده‌اند. برای مجموعه‌های بزرگ داده ممکن است مدل کردن پیوندها در سطحی بالاتر مؤثر باشد

فرض دنیای باز در مقابل دنیای بسته[1]: سنتی‌ترین رویکردها فرض می‌کنند که ما تمام موجودیتهای بالقوه را در این حوزه می‌شناسیم. این فرض دنیای بسته در کاربردهای دنیای واقع، غیرواقعی است. کار در این زمینه، معرفی یک زبان برای تعیین توزیعهای احتمال ساختارهای رابطه‌ای را دربرمی‌گیرد که متضمن مجموعه‌ای متغیر از اشیاء است.

اجتماع‌کاوی بر روی شبکه‌های چندرابطه‌ای: کار بر روی تحلیل شبکه‌های اجتماعی نوعاً کشف گروه‌هایی از اشیاء را که در ویژگیهای مشابهی سهیم هستند، دربرمی‌گیرد. به این‌کار اجتماع[2]کاوی گفته می‌شود. پیوند[3] صفحات وب یک مثال است که در آن اجتماع کشف شده می‌تواند مجموعه‌ای از صفحات وب مربوط به یک مبحث خاص باشد. اغلب الگوریتمهای اجتماع‌کاوی فرض می‌کنند که تنها یک شبکه اجتماعی وجود دارد که نشانگر رابطه نسبتاً همگنی است. در واقعیت، چندین شبکه اجتماعی ناهمگن وجود دارد که بیانگر روابط گوناگونی هستند. چالش جدید کاوش اجتماعات پنهان در چنین شبکه‌های اجتماعی ناهمگن است که به

---

[1]- Closed Versus Open World Assumption

[2]- Community

[3]- Linkage

آن اجتماع‌کاوی بر روی شبکه‌های اجتماعی چندرابطه‌ای گفته می‌شود. این چالشها ادامه دارد تا انگیزه‌ای برای تحقیقات بیشتر در پیوند کاوی باشد.

۹-۴- کاوش شبکه‌های اجتماعی

در این قسمت ما چند مثال در حوزه‌های مختلف کاوش بر روی شبکه‌های اجتماعی را بررسی می‌کنیم. این نمونه‌ها شامل پیش‌بینی پیوندها، کاوش شبکه‌های مشتریان برای بازاریابی ویروسی[1]، کاوش گروه‌های خبری با استفاده از شبکه‌ها و اجتماع‌کاوی از شبکه‌های چندرابطه‌ای است. سایر نمونه‌ها شامل کشف زیرگرافهای مشخصه در گراف‌کاوی و کاوش ساختارهای پیوند در وب‌کاوی است. کاربردهای دیگری مانند خوشه‌بندی و دسته‌بندی مبتنی بر پیوند نیز وجود دارند.

پیش‌بینی پیوندها: چه یالهایی به شبکه افزوده خواهد شد؟

شبکه‌های اجتماعی پویا هستند. پیوندهای جدیدی ظاهر می‌شده که نشان دهنده تعاملی جدید بین اشیاء هستند. در مسئله پیش‌بینی پیوندها، یک تصویر آنی[2] از شبکه اجتماعی در زمان t در اختیار ما قرار داده می‌شود و پیش‌بینی اینکه در بازه زمانی t تا $t+1$ چه یالهایی به این شبکه افزوده خواهد شد، از ما خواسته می‌شود. در این حالت ما به دنبال این هستیم که با استفاده از صفات حقیقی خود مدل، از توسعه‌ای که می‌تواند تکامل یک شبکه اجتماعی را مدل کند، پرده برداریم. به‌عنوان مثال یک شبکه اجتماعی هم‌نویسندگی یا تألیف مشترک، بین دانشمندان را درنظر بگیرند. به‌طور شهودی ممکن است پیش‌بینی کنیم که دو دانشمند که در شبکه نزدیک یکدیگر قرار دارند، احتمال دارد در آینده با هم همکاری داشته باشند. بر این اساس پیش‌بینی پیوندها را می‌توان به عنوان بخشی از مطالعات مدلهای تکامل شبکه‌های اجتماعی در نظرگرفت.

رویکردهایی که در باب پیش‌بینی پیوندها مطرح شده‌اند، مبتنی بر معیارهای مختلفی از تحلیل مجاورت گره‌های یک شبکه می‌باشند. بسیاری از معیارها از روشهای تحلیل شبکه‌های

---

[1]- Viral Marketing

[2]- Snapshot

اجتماعی و نظریه گراف سرچشمه می‌گیرند. روش کلی بدین صورت است: در تمام روشها به هر جفت گره X و Y، بر اساس گراف ورودی G و اندازه مجاورت موجود، یک وزن اتصال[۱] $Scroe(X,Y)$ تخصیص می‌دهند. سپس یک ترتیب نزولی از $Scroe(X,Y)$ تهیه می‌شود که پیوندهای جدید پیش‌بینی شده را به ترتیب نزولی اطمینان به ما می‌دهد. این پیش‌بینیها را می‌توان بر اساس مشاهدات واقعی از مجموعه داده‌های تجربی ارزیابی نمود. ساده‌ترین رویکرد، جفتهای (X,Y) را براساس طول کوتاهترین مسیرشان در G مرتب می‌کند. این رویکرد نظریه دنیاهای کوچک را مجسم می‌کند که در آن تک تک اجزا از طریق کوتاهترین زنجیره به یکدیگر متصل هستند. از آنجا که هدف مشترک همه روشها، مرتب کردن تمام جفتها به ترتیب کاهش امتیاز می‌باشد، در اینجا $Scroe(X,Y)$ به‌صورت منفی طول کوتاهترین مسیر تعریف می‌شود. بسیاری از معیارها از اطلاعات همسایگی استفاده می‌کنند. ساده‌ترین نوع چنین معیارهایی همسایگان مشترک[۲] است (هر قدر تعداد همسایگان مشترک X,Y بیشتر باشد، احتمال اینکه در آینده بین X,Y پیوندی ایجاد شود بیشتر است). از لحاظ شهودی، اگر نویسنده X,Y هرگز مقاله مشترکی تالیف نکرده باشند ولی همکاران مشترک بسیاری داشته باشند، با احتمال بیشتری آنها در آینده با یکدیگر همکاری خواهند داشت. سایر معیارها بر اساس مجموع تمام مسیرهای[۳] بین دو گره استوار است. به‌عنوان مثال معیار کتز[۴] تمام مسیرهای وزن دهی شده بین X و Y را محاسبه می‌کند به‌طوری‌که به مسیرهای کوتاهتر وزن بیشتری اختصاص می‌دهد. تمام این معیارها را می‌توان با ترکیبی از رویکردهای سطح بالاتر مانند خوشه‌بندی به‌کار برد. به‌طور نمونه، روش پیش‌بینی پیوند را می‌توان برای نسخه اصلاح شده یک گراف به‌کار برد که در آن یالهای قلابی حذف شده‌اند.

در آزمایشات انجام شده برروی مجموعه داده‌ها استناد یا نقل قول، هیچ‌کدام از روشها بر دیگر روشها مقدم نیست. بسیاری روشها به‌طور عمده یک پیشگویی تصادفی[۵] را به‌دست

۱- Connection Weight

۲- Common Neighbors

۳- Ensemble of All Paths

۴- Katz

۵- Random Predictor

می‌آورند که معتقد است توپولوژی شبکه‌ها می‌تواند اطلاعات مفیدی برای پیش‌بینی پیوندها فراهم آورد. معیار کتز و نسخه‌ها گوناگون مبتنی بر خوشه‌بندی آن، همواره خوب عمل کرده‌اند هرچند دقت پیش‌بینی همچنان بسیار پایین است. کارهای آینده در زمینه پیش‌بینی پیوندها ممکن است هم بر یافتن راههای بهتر استفاده از اطلاعات توپولوژی شبکه متمرکز شود و هم کارآیی محاسبات فاصله گره‌ها را مثلاً از طریق تخمین زدن بهبود بخشد.

کاوش شبکه‌های مشتریان به منظور بازاریابی ویروسی

بازاریابی ویروسی کاربردی از کاوش شبکه‌های اجتماعی است که مشخص می‌کند چگونه افراد می‌توانند بر رفتار خرید سایرین تأثیر بگذارند. به‌طور سنتی شرکتها، بازاریابی مستقیم[۱] (که در آن تصمیم فروش از طریق یک فرد خاص و صرفاً بر اساس خصوصیات آن فرد گرفته می‌شد) یا بازاریابی انبوه[۲] (که در آن افراد بر اساس اینکه به کدام بخش[۳] از جمعیت متعلق باشند، صورت می‌گیرد) را به‌کار برده‌اند. در هر حال این رویکردها تأثیری را که مشتریان می‌توانند بر تصمیم خرید سایرین داشته باشند نادیده می‌گیرند. به‌عنوان مثال فردی را در نظر بگیرید که تصمیم می‌گیرد یک فیلم خاص را ببیند و گروهی از دوستان را نیز برای تماشای آن فیلم تشویق می‌کند. هدف بازاریابی ویروسی بهینه‌کردن تأثیر مثبت گفتار شفاهی[٤] در میان مشتریان است. ممکن است بخواهیم هزینه بیشتری برای جذب یک نفر که تماسهای اجتماعی زیادی دارد، اختصاص دهیم. بنابراین با در نظر گرفتن این تعاملات بین مشتریان، بازاریابی ویروسی می‌تواند سود بیشتری نسبت به بازاریابی سنتی که چنین تعاملاتی را نادیده می‌گرفت، کسب نماید. رشد اینترنت در دهه‌های اخیر موجب در دسترس قرار گرفتن شبکه‌های اجتماعی بسیاری شده است که می‌توان با هدف بازاریابی ویروسی به کاوش آنها پرداخت، مثل لیستهای پست الکترونیک، گروه‌های خبری[٥]، اجتماعات برخط[٦] گپهای *IRC*[۷]، پیام آنی[۱]، سیستمهای فیلتر

۱- Direct Marketing
۲- Mass Marketing
۳- Segment
٤- Word of Mouth
٥- Use Net Groups
٦- On Line Forums
۷- Instant Relay Chat

مشارکتی و سایتهای شراکت دانش[۲]. سایتهای شرکت دانش (مانند *Epinions* در *www.epinions. com*) به‌کاربران خود اجازه می‌دهند (نوعاً به‌صورت مجانی) محصولات را به سایرین توصیه کنند یا برای محصولات برآورد قیمت نمایند تا به سایرین کمک کرده باشند. کاربران می‌توانند مفید بودن یا قابلیت اعتماد[۳] یک مقاله مروری[۴] را ارزش‌گذاری کنند، و حتی درصورت امکان سایر منتقدان را نیز ارزیابی نمایند. در این صورت یک شبکه روابط با عنوان شبکه اعتماد[۵] بر اساس اعتماد بین کاربران شکل می‌گیرد که بیانگر یک شبکه اجتماعی است که می‌تواند مورد کاوش قرار بگیرد.

ارزش شبکه‌ای[۶] یک مشتری، افزایش مورد انتظار در فروش به سایرین است که از جذب آن مشتری حاصل می‌شود. در مثال فوق، اگر مشتری مورد نظر ما سایرین را متقاعد کند که یک فیلم خاص را ببینند، شرکت فیلم‌سازی به صرف هزینه بیشتر برای ترغیب وی به مشاهده فیلم مشتاق می‌شود. در عوض اگر این مشتری نوعاً هنگام تصمیم‌گیری درباره اینکه چه فیلمی را تماشا کند، به سایرین گوش می‌کند، هزینه بازاریابی صرف شده برای وی را می‌توان اتلاف منابع محسوب نمود. بازاریابی ویروسی ارزش شبکه‌ای یک مشتری را درنظر می‌گیرد. در حالت ایده‌آل، ترجیح می‌دهیم شبکه یک مشتری را مورد کاوش قرار دهیم (به‌عنوان مثال شبکه دوستان و اقوام وی) تا نه تنها بر اساس ویژگیهای مشتری، بلکه بر اساس تأثیر همسایگان مشتری در شبکه، پیش‌بینی کنیم، چقدر احتمال دارد وی یک کالای خاص را بخرد. اگر مجموعه مشخصی از مشتریان را مورد بازاریابی قرار دهیم، از طریق بازاریابی ویروسی می‌توانیم میزان سود مورد انتظار از کل شبکه را پس از آنکه تأثیر آن مشتریان در شبکه تکثیر شد، در تحقیقاتمان دنبال کنیم. این کار می‌تواند به ما در یافتن مجموعه بهینه مشتریانی که مورد بازاریابی قرار می‌گیرند، یاری رساند. ارزش شبکه‌ای مشتریان (که در بازاریابی مستقیم سنتی نادیده گرفته می‌شوند) می‌تواند به طرح بازاریابی بهبود یافته‌ای بیانجامد.

۱- Instant Messaging

۲- Knowledge- Sharing Sites

۳- Trust Worthiness

۴- Review

۵- Web of Trust

۶- Network Values

مجموعه n مشتری بالقوه را در نظر بگیرید، X_i را یک متغیر دوحالته فرض کنید که اگر مشتری i محصول بازاریابی شده را بخرد مقدار یک می‌گیرد و در غیر این صورت مقدارش صفر می‌شود. همسایگان X_i مشتریانی هستند که مستقیماً بر X_i تأثیر می‌گذارند. M_i اقدام بازاریابی[1] است که در مورد مشتری iام صورت می‌گیرد. M_i می‌تواند دوحالته (اگر کوپنی[2] برای این مشتری ارسال شود، یک و در غیر این صورت صفر) یا طبقه‌ای (نشانگر اقدامات متفاوت ممکن اتخاذ شده در قبال مشتری) باشد. در حالتی دیگر M_i می‌تواند مقادیری پیوسته (مثلاً نشان دهنده میزان تخفیفی که برای مشتری قائل شده‌ایم) باشد. یافتن طرح بازاریابی که سود را بیشینه کند، مطلوب ماست. یک مدل احتمالی مطرح شده است که M_i را به‌عنوان مقداری پیوسته، بهینه می‌کند. در این مدل به جای اینکه تنها تصمیمی دوحالته اتخاذ شود که آیا این مشتری را مورد بازاریابی قرار بدهد یا خیر، مقدار هزینه صرف شده برای بازاریابی هر مشتری، بهینه می‌شود.

این مدل عواملی را که بر ارزش شبکه مؤثرند، در نظرمی‌گیرد. اول این مشتری می‌بایست اتصالات زیادی در شبکه داشته باشد و به‌علاوه محصول در نظر وی خوب قلمداد شود. درصورتی‌که یک مشتری دارای اتصالات زیاد، نگرش منفی نسبت به محصول مورد نظر داشته باشد، ارزش شبکه‌ای او می‌تواند منفی باشد که در این حالت، بازاریابی (جذب) وی توصیه نمی‌شود. دوم، این مشتری می‌باید ترجیحاً بر دیگران تأثیرگذار باشد تا تأثیرپذیر. سوم طبیعت بازگشتی[3] اثر گذاری تبلیغ شفاهی را می‌بایست مد نظر قرار داد. یک مشتری ممکن است بر آشنایان خود تأثیر بگذارد. آنها هم به نوبه خود ممکن است محصول را بپسندند و بر سایر افراد تأثیر بگذارند وضع به همین ترتیب ادامه می‌باید تا جایی که کل شبکه پوشش داده می‌شود. در ضمن این مدل ملاحظات مهم دیگری را نیز در نظر می‌گیرد: ممکن است تولید کننده بپذیرد که مقداری پول را برای مشتریانی که به اندازه کافی بر دیگران تأثیر مثبت دارند از دست بدهد. به‌عنوان مثال، فروش مجانی به یک مشتری برگزیده مناسب می‌تواند چندین برابر در فروش به سایر مشتریان جبران شود. این رویکرد چرخشی بزرگ نسبت به بازاریابی سنتی مستقیم است که در آن درصورتی‌که سود حاصل از فروش به یک مشتری به تنهایی از هزینه توصیه محصول به

---

[1]- Marketing Action

[2]- Coupon

[3]- Recursive

وی تجاوز کند، تحفیفی به آن مشتری تعلق می‌گیرد. این حقیقت را که، تنها دانش جزئی نسبت به شبکه داریم و جمع‌آوری چنین دانشی نیز می‌تواند هزینه‌های مربوط به خود را داشته باشد، در این مدل مورد توجه قرار می‌گیرد.

یافتن مجموعه مشتریان بهینه به‌صورت یک مسئله بهینه‌سازی که به خوبی تعریف شده، فرموله شده است: یافتن مجموعه مشتریانی که سود خالص را بیشینه می‌کنند. این مسئله از نوع دشوار *NP* شناخته شده است. با به‌کارگیری رویه جستجوی تپه‌نوردی ساده می‌توان با تقریب ۶۳٪ جواب بهینه را به‌دست آورد. مادامی‌که افزایش مشتریان سود سرانه را بهبود بخشد، مشتریان جدیدی به مجموعه مشتریان افزوده خواهند شد. این روش با وجود دانش ناقص از شبکه باثبات[1] شناخته شده است.

روشهای بازاریابی ویروسی را می‌توان برای حوزه‌های دیگر نیز به‌کار برد. کاهش انتشار ویروس *HIV* مبارزه با استعمال دخانیات در نوجوانان و ابتکار سیاسی اجتماعی محلی[2] مثالهایی از این نوع هستند. به‌کارگیری روشهای بازاریابی ویروسی برای حوزه وب و عکس آن، زمینه‌های جالبی برای تحقیقات آتی هستند.

۹-۴-۱- کاوش گروه‌های خبری با کمک شبکه‌ها

تحلیل شبکه‌های اجتماعی مبتنی بر وب با وب‌کاوی رابطه نزدیکی دارد. در وب‌کاوی دو الگوریتم رتبه‌بندی متداول به نام صفحه‌رتبه و *HITS* به‌کار می‌رود. این الگوریتمها بر این پایه استوارند که پیوندی از صفحه وبی A به B معمولاً نشانه تأیید B توسط A است.

وضعیت در گروه‌های خبری روی مباحث موضوعی متفاوت است. یک پست نوعی در گروه خبری شامل یک یا چند خط نقل قول[3] از پستی دیگر و به دنبال آن نظر نویسنده پست فعلی است. این پاسخ‌های نقل قول‌دار پیوندهای نقل قولی را شکل داده و شبکه‌ای را ایجاد می‌کنند که در آن رئوس، بیانگر افراد و پیوندها رابطه پاسخ هستند. پدیده جالب این است که مردم بیشتر به پیامی که با آن مخالفند پاسخ می‌دهند تا پیامی که موافقند. این رفتار که در بسیاری از گروه‌های

---

[1]- Robust

[2]- Grass- Root Political Initiative

[3]- Quote

خبری وجود دارد کاملاً با گراف پیوند صفحات وب که در آن پیوند نشانه توافق یا علاقه مشترک می‌باشد، متضاد است. بر پایه این رفتار می‌توان با تحلیل ساختار گراف پاسخ‌ها به‌طور مؤثر، نویسندگان داخل گروه خبری را به دسته‌های مخالف دسته‌بندی و افراز کرد. این فرایند دسته‌بندی گروه خبری، با نظریه گراف قابل انجام است. اگر فرد i از پست قبلی فرد j نقل قول کند، پیوند نقل قول بین i و j ساخته شده و از روی این پیوندها شبکه یا گراف نقل قول ایجاد می‌شود. حال دوبخشی کردن رئوس به دو مجموعه را در نظر می‌گیریم:

مجموعه F بیانگر موافقین یک مطلب و مجموعه A بیانگر مخالفین آن مطلب است. اگر یالهای گراف گروه خبری بیانگر مخالفت باشند آن‌گاه انتخاب بهینه، حداکثر کردن تعداد یالها بین این دو مجموعه است. از آنجا که مسئله حداکثر برش (حداکثر کردن تعداد یالهای برش خورده برای دو بخش کردن گراف) به‌طور نظری مسئله‌ای از نوع NP است، نیاز به راه حل عملی دیگری داریم. به‌خصوص می‌توان از دو حقیقت دیگر در وضعیت فعلی استفاده کرد:

نمونه ما بیشتر از آن‌که گراف عامی باشد، گرافی دوبخشی است که برخی گره‌های مغشوش به آن اضافه شده‌اند.

هیچ‌کدام از دو بخش چندان از دیگری کوچکتر نیستند.

در چنین وضعیتهایی می‌توان مسئله را به مسئله برش تقریباً متوازن حداقل وزن تبدیل کرد، که به نوبه خود با روشهای طیفی ساده تقریب زده می‌شود. برای بهبود دقت دسته‌بندی می‌توان ابتدا به‌طور دستی تعداد کمی از پست‌کنندگان فعال را دسته‌بندی کرد و رئوس متناظر در گراف را علامت زد. سپس از این اطلاعات برای دستیابی به افرازی بهتر استفاده کرد به این نحو که در حین اجرای الگوریتم افراز، قرار داشتن موارد دستی در دو سو حفظ شود.

بر پایه این ایده‌ها، الگوریتم مؤثری ارائه شده است. آزمایشاتی با چند مجموعه داده گروه خبری در مباحث اجتماعی بحث‌انگیز مانند سقط جنین، کنترل اسلحه و مهاجرت نشان می‌دهد که پیوندها حاوی اطلاعاتی کم‌اغتشاش‌تر از متون هستند. روشهای مبتنی بر تحلیل زبانی و آماری متن، صحت کمتری از تحلیل پیوند در این نوع گروه‌های خبری داشتند، زیرا الفاظ مورد استفادهٔ طرفهای مقابل تا حد زیادی مشابه بوده و بسیاری از پستها حاوی متنی بسیار مختصر هستند که امکان تحلیل زبانی قابل اعتمادی نمی‌دهد.

۹-۴-۲- اجتماع کاوی شبکه‌های چندرابطه‌ای

با رشد وب، اجتماع کاوی توجهات روز افزونی را به خود جلب نموده است. قسمت اعظم چنین کارهایی بر کاوش اجتماعات ضمنی صفحات وب، اجتماعات ضمنی ادبیات علمی برگرفته از وب و اجتماعات ضمنی استنادات متمرکز شده است. در اصل یک اجتماع را می‌توان به‌صورت گروهی از اشیاء که در چندین خصوصیات مشترک سهیم هستند، تعریف نمود. اجتماع کاوی را می‌توان به‌صورت تعیین زیر گرافها در نظر گرفت. به‌عنوان مثال، در صفحات وب مربوط، دو صفحه وب (اشیاء) در صورتی به هم مرتبط هستند که ابرپیوندی بین آنها وجود داشته باشد. گرافی از روابط صفحات وب را می‌توان به منظور تعیین اجتماع یا تعیین مجموعه‌ای از صفحات وب در مورد یک موضوع خاص، مورد کاوش قرار داد.

اغلب روشهای گراف‌کاوی و اجتماع کاوی مبتنی بر گرافهای همگن است. یعنی آنها فرض می‌کنند تنها یک نوع رابطه بین اشیاء وجود دارد. در حالی‌که در شبکه‌های اجتماعی واقعی، همواره انواع مختلف روابط بین اشیاء برقرار است. به هر نوع رابطه می‌توان در قالب یک شبکه روابط[1] نگریست. (که به آن شبکه‌های اجتماعی همگن نیز گفته می‌شود). در این‌صورت چند نوع رابطه یک شبکه اجتماعی چندرابطه‌ای[2] را تشکیل می‌دهند (که به آن شبکه‌های اجتماعی غیر همگن نیز گفته می‌شود). هر نوع رابطه ممکن است نقش معینی در یک وظیفه خاص ایفا نمایند. در اینجا گرافهایی با روابط مختلف می‌توانند برای ما اجتماعات متفاوتی را ایجاد نمایند.

به‌منظور یافتن اجتماعی با خصوصیات معلوم ابتدا نیاز است تعیین شود، کدام رابطه نقشی مهم در چنین اجتماعی ایفا می‌کند. چنین رابطه‌ای ممکن است به‌طور صریح وجود نداشته باشد، یعنی ما نیاز داشته باشیم قبل از آنکه آن اجتماع را در چنان شبکه رابطه‌ای بیابیم ابتدا چنین رابطه پنهانی را کشف کنیم. کاربران مختلف ممکن است به روابط متفاوتی در درون شبکه علاقه‌مند باشند. بدین لحاظ اگر ما شبکه‌ها را تنها با در نظر گرفتن تنها یک نوع رابطه مورد کاوش قرار دهیم، ممکن است اطلاعات ارزشمند بسیاری از اجتماع پنهان را از دست بدهیم. چنین کاوشی نمی‌تواند نیازهای متنوع اطلاعاتی کاربران مختلف را برآورده کند. این امر ما را به مسئله کاوش

---

[1]- Relation Network

[2]- Multi Relational Social Network

اجتماعات چندرابطه‌ای می‌رساند که کاوش اجتماعات پنهان در شبکه‌های اجتماعی ناهمگن را دربرمی‌گیرد.

مثالی ساده را بررسی می‌کنیم. در یک اجتماع انسانی، ممکن است چندین نوع رابطه وجود داشته باشند: برخی از مردم در یک محل کار می‌کنند؛ بعضی علایق مشترکی دارند؛ برخی به یک درمانگاه می‌روند، والی آخر. از لحاظ ریاضی این اجتماع را می‌توان با یک گراف بزرگ نمایش داد که در آن گره‌ها نماد افرادهستند و یالها قوت رابطه بین آنها را مشخص می‌کنند. از آنجا که انواع مختلف رابطه وجود دارد، یالهای این گراف می‌بایست غیر همگن باشند. در برخی موارد، می‌توان این اجتماع را با به‌کارگیری چندین گراف همگن نیز مدلسازی نمود. هر گراف یک نوع رابطه را منعکس می‌کند. فرض کنید یک بیماری مسری شیوع پیدا کند، و دولت تلاش می‌کند افرادی را که بیشترین احتمال بیمار شدن را دارند، شناسایی نماید. به وضوح پیداست که روابط موجود بین مردم نمی‌توانند نقش یکسانی داشته باشند. منطقی است که فرض کنیم در چنین شرایطی رابطه «کارکردن در یک محل» یا «زندگی کردن با یکدیگر» می‌بایست نقشی حیاتی ایفا نمایند. سؤالی که پیش می‌آید این است: «چگونه می‌توانیم رابطه‌ای که بیشترین نقش را در شیوع بیماری دارد، انتخاب کنیم؟ آیا رابطه پنهانی (بر اساس روابط صریح و آشکار) وجود دارد که به بهترین نحو مسیر شیوع بیماری را فاش نماید؟»

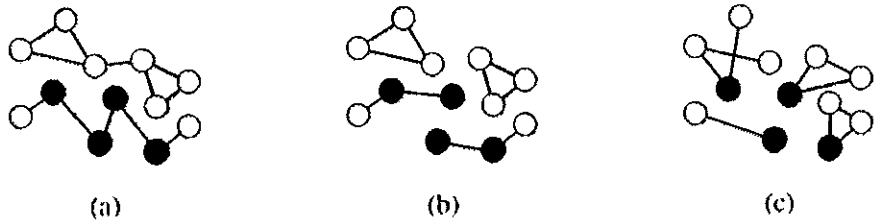
سؤال از لحاظ ریاضی می‌تواند به‌صورت انتخاب و استخراج رابطه[1] در تحلیل شبکه‌های اجتماعی چندرابطه‌ای، مدلسازی شود. مسئله استخراج رابطه را می‌توان به صورت ساده آن‌گونه که در ادامه خواهد آمد، بیان نمود: در یک شبکه اجتماعی چندرابطه‌ای، بر اساس چند نمونه برچسب گذاری شده است (به‌عنوان مثال به‌صورت پرس‌وجوهایی که کاربر فراهم می‌کند). چگونه می‌توان اهمیت روابط مختلف را ارزیابی نمود؟ درضمن چگونه می‌توان به ترکیبی از روابط موجود دست یافت که بیشترین تطابق را با رابطه نمونه‌های برچسب خورده داشته باشد؟

به‌عنوان مثال شبکه نشان داده شده در شکل (۹- ۳) را ببینید که سه نوع رابطه مختلف دارد؛ به ترتیب با (a)، (b) و (c) مشخص شده‌اند. فرض کنید کاربری نیاز دارد که ٤ شیء رنگی به یک اجتماع تعلق داشته باشند و این امر را با یک پرس‌وجو مشخص می‌کند. پیداست که اهمیت

---

[1]- Relation Selection & Extraction

نسبی هر کدام از سه نوع رابطه با توجه به نیاز اطلاعاتی کاربر فرق می‌کند. در بین این سه نوع رابطه، (a) بیشترین تطابق را با نیاز کاربر دارد و لذا مهم‌ترین است، در حالی که (b) در رتبه دوم قرار می‌گیرد. رابطه (c) را می‌توان با توجه به نیاز اطلاعاتی کاربر، یک اغتشاش در نظر گرفت. تحلیل سنتی شبکه‌های اجتماعی بین این روابط تمایزی قائل نمی‌شود و با روابط مختلف یکسان برخورد می‌کند. آنها به سادگی برای توصیف ساختار بین اشیا با یکدیگر ترکیب می‌شوند. متاسفانه، در این مثال، رابطه (c) تأثیری منفی بر این هدف می‌گذارد. با این وجود، اگر ما این روابط را بر اساس اهمیتشان ترکیب کنیم، رابطه (c) به آسانی حذف می‌شود و روابط (a) و (b) را برای کشف ساختار اجتماع باقی می‌گذارد که با نیاز کاربر مطابق است.



شکل ۹-۳) شبکه‌هایی با روابط مختلف

بعضی مواقع یک کاربر ممکن است پرس‌وجوی پیچیده‌تری را ارائه دهد. مثلاً ممکن است مشخص کند که دو شیء رنگی پایینی می‌بایست به دو اجتماع متفاوت متعلق باشند. در این صورت، اهمیت سه نوع رابطه شکل (۹-۳) تغییر می‌کند. رابطه (b) مهمترین رابطه می‌شود، درحالی که رابطه (a) بی فایده قلمداد می‌شود (حتی با توجه به پرس‌وجو اثری منفی دارد). درنتیجه، در شبکه‌های اجتماعی چندرابطه‌ای، اجتماع‌کاوی می‌بایست بر اساس پرس‌وجوی کاربر (یا اطلاعاتی که لازم دارد) صورت بگیرد. پرس‌وجوی یک کاربر می‌تواند بسیار منعطف باشد. روشهای اولیه تنها بر یک شبکه رابطه متمرکز بودند و بر اساس پرس‌وجوی کابر انجام نمی شدند و لذا نمی‌توانستند پاسخگوی چنین موقعیتهای پیچیده‌ای باشند.

الگوریتمی برای استخراج و انتخاب رابطه مطرح شده که مسئله را به‌صورت یک مسئله بهینه‌سازی مدل می‌کند. این مسئله از لحاظ ریاضی می‌تواند بدین صورت تعریف شود: مجموعه‌ای از اشیاء و مجموعه‌ای از روابط را در اختیار داریم که به‌صورت مجموعه‌ای از

گرافهای $G_i(V,E_i)$ ، $i=1,...,n$ در آن n تعداد روابط است، V مجموعه گره‌هاست (اشیاء) و E_i مجموعه یالهای مربوط به iامین رابطه است. وزن یالها را می‌توان به صورت طبیعی بر اساس قوت رابطه بین دو شیء تعیین نمود. این الگوریتم هر رابطه را با یک گراف و یک ماتریس اوزان مشخص می‌کند.

M_i را نماد ماتریس اوزان مربوط به Gi در نظر بگیرید. هر مؤلفه در ماتریس قوت رابطه بین یک جفت شیء مربوط را مشخص می‌کند. فرض کنید یک رابطه پنهانی با گراف $G^\wedge(V,E^\wedge)$ مشخص شود و $M^\wedge$ هم ماتریس اوزان مرتبط با $G^\wedge$ است.

یک کاربر نیاز اطلاعاتی خود را به‌صورت یک پرس‌وجو مشخص می‌کند که در آن مجموعه‌ای از اشیاء برچسب‌گذاری‌شده $X=[x_1,...,x_m]$ را اعلان می‌نماید.

$Y=[y_1,...,y_m]$ نیز به گونه‌ای است که در آن y_j برچسب x_j است (چنین اشیاء برچسب گذاری شده‌ای اطلاعات جزئی از روابط پنهان $G^\wedge$ را نشان می‌دهند). هدف این الگوریتم یافتن ترکیب خطی این ماتریسهای اوزان است، به نحوی‌که به بهترین صورت $G^\wedge$ (ماتریس اوزان مربوط به اشیاء دارای برچسب) را تخمین بزنند. ترکیب حاصله با احتمال بیشتری نیازمندی اطلاعاتی کاربر را برآورده ساخته و لذا موجب عملکرد بهتری در اجتماع کاوی خواهد شد.

این الگوریتم روی داده‌های کتاب‌شناسی آزمایش شده است. به‌طور طبیعی روابط چندگانه‌ای بین نویسندگان وجود دارد. نویسندگان می‌توانند مقالات خود را در هزاران کنفرانس مختلف به چاپ برسانند، و هر کنفرانس را می‌توان به‌صورت یک رابطه در نظر گرفت که یک شبکه اجتماعی چندرابطه‌ای را ایجاد خواهد کرد. با در دست داشتن چند مثال به‌دست آمده از کاربران (مانند گروه‌های نویسندگان)، این الگوریتم می‌تواند یک رابطه جدید را با استفاده از مثالها استخراج کند و تمام گروه‌های مرتبط دیگر را بیابد. رابطه استخراجی می‌تواند به‌صورت گروهی از نویسندگان باشد که علایق مشترکی را دنبال می‌کنند.

منابع

1) *Han. J, Kamber. M. (2006) "Chapter 9: Graph Mining, Social Network Analysis, and Multirelational Data mining", Data mining concepts and techniques, 2nd edition, , Morgan Kaufmann Publishers*

فصل دهم

کاربرد داده‌کاوی در مدیریت ارتباط با مشتری

توجه: مطالب این فصل به‌طور مستقل از فصول قبل قابل مطالعه بوده و برای دانشجویان رشته‌های تجارت الکترونیک و مدیریت بازرگانی مناسب می‌باشد.

در طی چند سال گذشته تعامل شرکتها با مشتریانشان به‌طور قابل توجهی تغییر کرده است به‌طوری‌که تداوم کسب وکار با مشتری تضمین بلند مدت ندارد. به همین دلیل برای موفقیت یک سازمان لازم است سازمان‌ها مشتریانشان را به‌درستی درک کرده، نیازها و خواسته‌های آنها را پیش‌بینی کنند و با مجهز شدن به این اطلاعات، سلامت کاری خود را بهبود بخشند. بسیاری از سازمانها داده‌های بسیار زیادی را درباره مشتریان، تامین کنندگان و شرکای تجاریشان جمع‌آوری و ذخیره می‌کنند ولی ناتوانی این سازمانها برای کشف دانش پنهان باارزش در این داده‌ها سبب می‌شود که این داده‌ها به دانش تبدیل نشوند و این کار عملاً بیهوده باشد. صاحبان کسب‌وکار میل به استخراج اطلاعاتی ناشناخته، معتبر و قابل درک از بانک‌های اطلاعاتی عظیم خود و استفاده از این اطلاعات برای کسب سود بیشتر دارند. برای برخی، داده‌کاوی از نظر فنی جالب است ولی برای بیشتر مردم، این علم وسیله‌ای برای رسیدن به نتایج جالب می‌باشد. داده‌کاوی به تنهایی مفید نیست، بلکه زمانی که به صورت کاربردی در یک مورد خاص استفاده می‌شود، معنا پیدا می‌کند. برای محقق شدن این اهداف سازمانها باید مراحل زیر را طی نمایند:

- جمع‌آوری و یکپارچه‌سازی داده‌های داخلی و خارجی (خرید) درکل سازمان به شکلی قابل درک.
- کاوش داده‌های یکپارچه برای تولید دانش.
- سازماندهی و ارائه اطلاعات و دانش به شیوه‌ای که فرآیندهای تصمیم‌گیری پیچیده را تسریع نماید.

برای محقق شدن همه این اهداف، سازمانها نیاز به یکپارچه‌سازی مؤلفه‌های مختلف برنامه‌های کاربردی خود دارند. یکی از حوزه‌هایی که به سرعت درحال رشد است فنآوری تصمیم‌گیری علمی[۱] (معمولاً به آنها موتورهای تصمیم‌گیری گفته می‌شود) و داده‌کاوی در مدیریت ارتباط با مشتری است.

برای حفظ رقابت، سازمانها نیاز به تدوین استراتژیهای تمرکز بر مشتری[۲]، مشتری محوری[۳]، مشتری مداری[۴] دارند. همه این موارد خواسته‌های سازمانها را در راستای ارتباط با مشتریان تعریف می‌کند. مدیریت ارتباط با مشتری[۵] راه حلی است که این تلاشها را برای سازمانها و همچنین مشتریان محقق می‌سازد.

فرض اولیه این است که همه مشتریان با یکدیگر برابر نیستند. هدف اصلی *CRM*، بهینه‌سازی نسبت وقایع سودآور به وقایع زیانبار برای گروه معلومی از مشتریان است. برخی وقایع خاص مانند فروش محصول، ایجاد درآمد و بعضاً تولید سود می‌کنند درحالی‌که برخی دیگر مانند تماسهای تلفنی و دیگر موارد موارد مشابه به منظور پشتیبانی این‌گونه نیستند. در این‌گونه موارد هدف داده‌کاوی افزایش درآمد و کاهش هزینه می‌باشد. چیزی که باید سازمانها بدانند این است که مشتریان به چه شیوه‌ای و چگونه تمایل به تعامل دارند تا بتوانند وفاداری مشتری را کسب نموده و به طبع سودآوری خود را نیز بهبود بخشند. مدیریت ارتباط با مشتری به سازمانها این اجازه را می‌دهد که مشتریان خود را بهتر بشناسند و تفاوت بین آنها را بهتر درک

۱- Decision Science

۲- Customer Focused

۳- Customer Driven

۴- Customer Centric

۵- Customer Relationship Management (CRM)

نمایند. در نتیجه در تخصیص منابع به مشتریان مطلوب‌تر، کارآمدی بیشتری داشته باشند. از طریق تلاشهای *CRM*، سازمانها می‌توانند هماهنگی بهتری در ارتباط با مشتری ایجاد نمایند، بنابراین سازمان می‌تواند مدیریت مؤثرتری برروی منابع بازاریابی و ارتباط بامعناتری با مشتریانش داشته باشد. ارتباط کارا با مشتری نیاز به درک الزامات این رابطه دارد. توانایی ارائه خدمات شخصی شده، ایجاد ارزش و پذیرش دوطرفه، تعهد به ارتباط متقابل و موارد مشابه همه در ایجاد ارتباط قوی تأثیر به‌سزایی دارند

۱۰-۱- معماری مدیریت ارتباط با مشتری

ازدید معماری، کل چارچوب کاری *CRM* به سه مؤلفه اصلی تقسیم می‌شود:

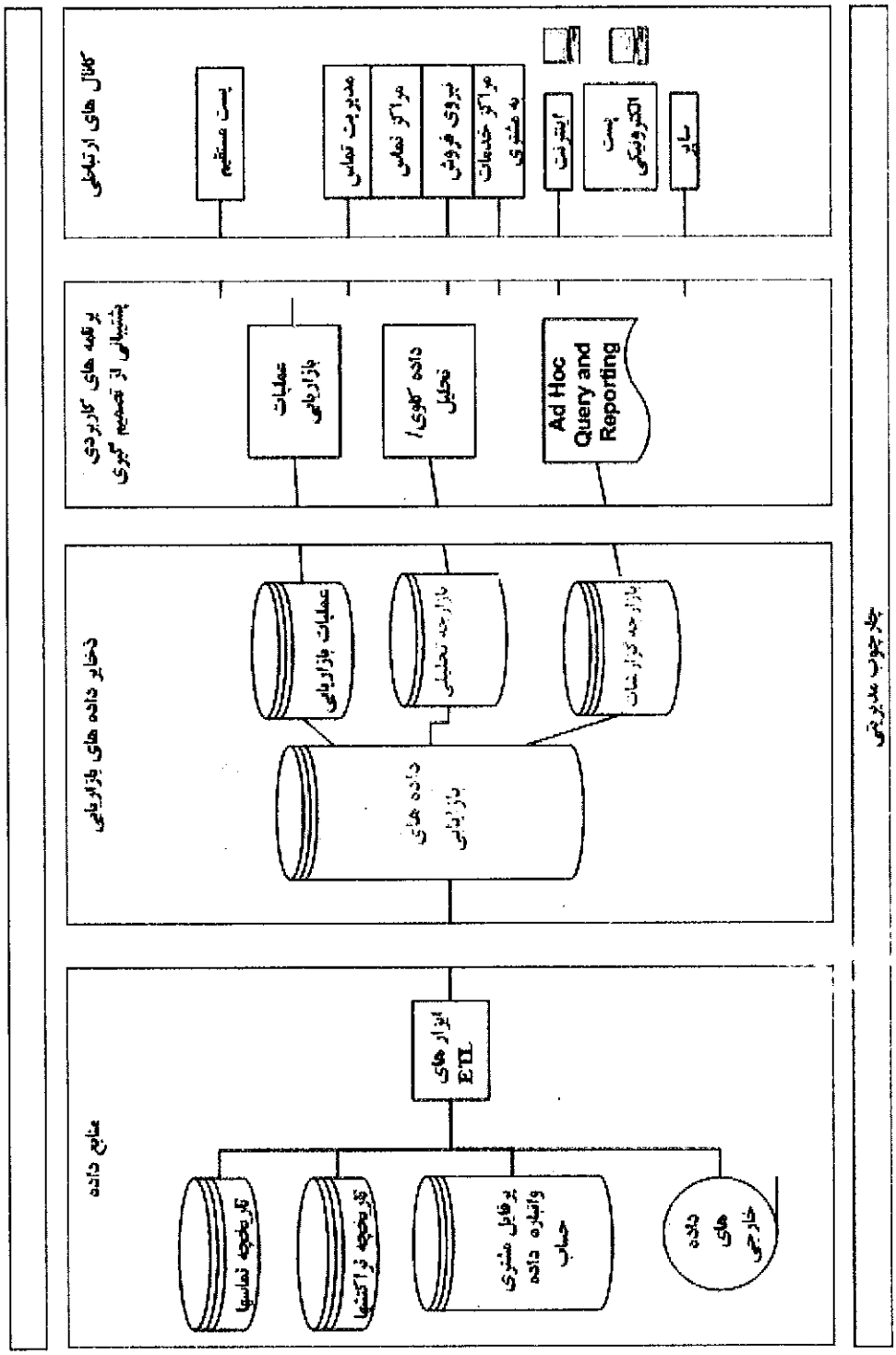
- *CRM* عملیاتی[۱]: عبارتست ازخودکارسازی فرآیندهای کسب‌وکاری افقی، شامل نقاط تماس مشتری، کانالها و یکپارچه کردن موضوعات پشت صحنه و با موارد قابل مشاهده در روی صحنه.
- *CRM* تحلیلی[۲]: تحلیل داده‌های ایجاد شده توسط *CRM* عملیاتی می‌باشد.
- *CRM* مشارکتی[۳]: برنامه‌های کاربردی خدمات مشارکتی شامل موارد زیر می‌باشد: پست الکترونیکی، انتشارات شخصی شده، ارتباطات الکترونیکی، رسانه‌های مشابهی که برای تسهیل تعامل بین مشتری و سازمان طراحی شده است.

همان‌طور که در (۱۰-۱) مشاهده می‌کنید معماری CRM شامل نقاط تماس با مشتری و کانالهای ارسال می‌باشد که اطلاعات را تولید و مصرف می‌نمایند. این اطلاعات نیاز به یکپارچگی و تحلیل به‌منظور تعیین تصویر کامل و دقیق از مشتریان، عملکرد، نیازها، شکایات و مشخصه‌های آنها دارد [۳].

۱- Operational CRM

۲- Analytical CRM

۳- Collaborative CRM



شکل ۱۰-۱) معماری CRM

۱۰-۱-۱- یافتن مشتریان احتمالی

هر یک از اهداف کسب و کار به فنون داده‌کاوی مناسب با آن مسئله مربوط می‌شوند. مباحث کسب و کار مورد بررسی متناسب با به ترتیب پیچیدگی رابطه مشتری می‌باشد. این بحث در ارتباط با مشتریان احتمالی[1] که کمتر شناخته شده‌اند شروع شده و تا فرصتهای متنوع بیان شده توسط روابط مستمر مشتری ادامه می‌یابد. موضوعات مورد نظر می‌تواند شامل محصولات، کانالهای ارتباطی متعدد و تعاملات فردی فزاینده باشد[٤].

به‌نظر می‌رسد جستجوی مشتریان احتمالی، شروع مناسبی برای بحث در مورد کاربردهای داده‌کاوی در این مبحث باشد. در بازاریابی، یک مشتری احتمالی کسی است که اگر به روش درستی با او برخورد شود، انتظار می‌رود به یک مشتری خوب تبدیل شود.

برای بیشتر کسب و کارها، تعداد کمی از شش میلیارد مردم زمین، مشتری احتمالی هستند. می‌توان بر اساس محل، سن، توانایی مالی و نیاز به محصول یا خدمت، اکثر مردم را کنار گذاشت. به‌عنوان مثال شرکتی که صندلی تاب حیاط می‌فروشد، قاعدتاً کاتالوگ خود را به خانوارهای دارای فرزند در مناطقی که حیاط خلوت دارند، ارسال می‌کند. یک مجله، گروهی را هدف‌گیری می‌کند که آن را مطالعه کرده و به آگهی‌هایش علاقه‌مند باشند.

داده‌کاوی می‌تواند نقش مهمی در یافتن مشتریان احتمالی داشته باشد. مهم‌ترین نقشهای آن عبارتند از:

- تشخیص مشتریان احتمالی خوب
- انتخاب کانال ارتباطی مناسب به منظور دسترسی به مشتریان احتمالی
- انتخاب پیام مناسب برای گروه‌های مختلف مشتریان احتمالی

داده‌کاوی تا کنون بیشتر در تشخیص مشتریان احتمالی خوب استفاده شده است.

تشخیص مشتریان احتمالی خوب

بنابر ساده‌ترین تعریف، « مشتری احتمالی خوب » کسی است که ممکن است علاقه به مشتری شدن نشان دهد. مشتریان احتمالی خوب نه فقط علاقه‌مند هستند تا مشتری شوند، بلکه

---

[1]- Prospects

استطاعت مالی مشتری شدن را نیز داشته و به‌عنوان مشتری سودآور خواهند بود. این مشتریان به احتمال زیاد صورتحساب‌هایشان را به‌موقع پرداخته، احتمال کمی دارد که سر شرکت کلاه بگذارند، اگر با آنها درست رفتار شود مشتریان وفاداری[1] خواهند بود و دیگران را نیز همراه خواهند کرد. جدا از سادگی و پیچیدگی تعریف مشتری احتمالی خوب، اولین وظیفه هدف‌گیری آنها است.

اگر قرار است پیام از طریق تبلیغ یا کانالهای مستقیم‌تر مانند پست، تلفن یا پست الکترونیک انتقال یابد، هدف‌گیری[2] بسیار مهم است. پیامهای تابلوهای تبلیغاتی تا حدودی هدف‌گیری شده‌اند. مثلاً تابلوهای تبلیغاتی خطوط هوایی و شرکتهای اجاره ماشین معمولاً در کنار بزرگراههای منتهی به فرودگاهها دیده می‌شوند، جایی که احتمالاً در بین رانندگان این راهها، استفاده‌کنندگان این خدمات وجود خواهند داشت.

برای به‌کاربردن داده‌کاوی در این مسئله، باید اول مشتری احتمالیِ خوب را تعریف کرده و سپس قواعد را یافت که اجازه هدف‌گیری مشتریان دارای خصوصیات مورد نظر را می‌دهد. برای اکثر شرکتها، اولین قدم به منظور استفاده از داده‌کاوی برای تشخیص مشتریان احتمالی خوب، ساختن یک مدل پاسخ[3] است.

انتخاب کانال ارتباطی

یافتن مشتریان احتمالی نیاز به ارتباط دارد. شرکتها به عمد از راههای مختلفی با مشتریان احتمالی تماس می‌گیرند. یکی از راه‌ها، روابط عمومی است که هدف آن تشویق رسانه‌ها برای پوشش داستانی درباره شرکت و نیز انتشار پیامهای مثبت افواهی می‌باشد. این کار برای برخی شرکتها بسیار مؤثر است ولی پیامهای بازاریابی مستقیم با روابط عمومی متفاوتند.

در اینجا تبلیغات و بازاریابی مستقیم مورد نظر است. تبلیغات می‌توانند روی جلد مجله، پنجره‌های باز شده مزاحم در برخی سایتهای تجاری، زیرنویسهای تلویزیونی در حین وقایع ورزشی مهم یا تبلیغات محصول در فیلمها باشد. این نوع تبلیغات، گروه‌های مردم را بر اساس

---

[1]- Loyal

[2]- Targeting

[3]- Response Model

صفات مشترک هدف‌گیری می‌کند ولی پیام را برای هر فرد مشخص نمی‌کند. در بخشهای آتی، از طریق تطابق پروفایل[1] مشتریان احتمالی با پروفایل یک ناحیه جغرافیایی، به‌منظور انتخاب محل مناسب تبلیغ بحث می‌شود.

پروفایل مشتری

اگر مشخصات عمومی مشتری مانند سن، جنسیت و آدرس با مشخصات مشتریان دیگر مقایسه شود، شرکت را قادر به شناسایی نوع افرادی می‌کند که محصولاتش را می‌خرند. این مشخصات به شرکت کمک می‌کند که محصولات دیگری برای همان گروه تولید کند یا استراتژیهای متفاوتی برای فروختن همان محصولات به بازارهای هدف دیگر توسعه دهد.

بازاریابی مستقیم، اجازه می‌دهد پیام برای هر فرد، شخصی شود. این کار می‌تواند از راه تماس تلفنی (مثل *SMS*)، پست الکترونیکی، کارت پستال و یا کاتالوگ رنگی گلاسه باشد. داده‌کاوی می‌تواند به تعیین کانالهای مؤثر برای هر گروه از مشتریان احتمالی کمک کند.

انتخاب پیامهای مناسب

حتی برای فروختن محصول یا خدمتی یکسان، پیامهایی متفاوت برای افراد مختلف مناسب است. مثلاً، ممکن است یک روزنامه برای یکی به دلیل پوشش اخبار ورزشی و برای دیگری به خاطر پوشش اخبار سیاسی یا هنری جذاب باشد. وقتی محصول دارای تنوع بوده یا چند محصول پیشنهاد می‌شود، انتخاب پیام مناسب مهم‌تر هم می‌شود.

حتی در یک محصول نیز پیام مهم است. یک مثال کلاسیک، تعامل میان معیار هزینه و معیار راحتی است. برخی مردم به قیمت خیلی حساس بوده و تمایل دارند از تعاونی خرید کرده، نصفه شب تلفن کنند، بین مسیر هواپیما عوض کنند و سفرهایشان شامل شب آخر هفته باشد. دیگران حاضرند برای خدمات راحت‌تر، مبالغ اضافه بپردازند. پیامی بر مبنای قیمت کمتر، نه

---

[1]- Profile

فقط در انگیزش راحت‌طلبان شکست می‌خورد بلکه ممکن است خطر راندن آنان به سمت محصولات کم‌سودتر را داشته باشد، درحالی‌که آنان تمایل دارند پول بیشتری خرج کنند.

مدلهای پاسخ که شامل یک فعالیت تبلیغی[1] هستند، می‌توانند با هم ترکیب شده تا بهترین پیشنهاد را به مشتری بدهند. برای دسته‌بندی مشتریان به بخشهای دارای تشابه فکری در پاسخگویی به پیشنهادات، می‌توان از فیلترکردن مشارکتی استفاده کرد.

۱۰-۱-۲- داده‌کاوی برای انتخاب محل مناسب تبلیغ

یک راه هدف‌گیری مشتریان احتمالی، نگاه کردن به مشتریان فعلی است. برای مثال یک نشریه کشوری از طریق پرسشنامه، مشخصات زیر را برای خوانندگانش به‌دست آورده است:

- ۵۸٪ تحصیلات عالی داشتند.
- ٤٦٪ مشاغل تخصصی یا مدیریتی داشتند.
- ۲۱٪ درآمد خانواری بیش از ۸ میلیون تومان در سال داشتند.
- ۷٪ درآمد خانواری بیش از ۱۲ میلیون تومان در سال داشتند.

درک این پروفایل از دو طریق می‌تواند به نشریه کمک کند. اول اینکه با هدف‌گیری مشتریان احتمالی مطابق با این پروفایل، می‌توان نرخ پاسخ به فعالیتهای ترویجی[2] خود را افزایش داد. دوم، می‌توان فضای تبلیغاتی نشریه را به شرکتهای علاقه‌مند به این نوع خوانندگانِ تحصیل‌کرده و پردرآمد فروخت. از آنجا که موضوع این بخش، هدف‌گیری مشتریان احتمالی است، بیایید ببینیم چگونه این نشریه از این پروفایل برای متمرکز کردن فعالیتهای مشتری‌یابی خود استفاده کرده است. ایده اصلی ساده است. وقتی نشریه می‌خواهد در روزنامه آگهی بدهد، باید به دنبال روزنامه‌هایی باشد که شنوندگانشان مطابق این پروفایل باشند. باید در جاهایی کارتهای معرفی خود را روی پیشخوان مغازه‌ها بگذارد که منطبق بر این پروفایل باشند. وقتی می‌خواهد بازاریابی با پیامک (*SMS*) انجام دهد باید با مردمی مطابق این پروفایل تماس بگیرد. چالش داده‌کاوی، ارائه تعریف مناسبی از معنای انطباق با این پروفایل است.

---

[1]- Campaign

[2]- Promotional

چه کسی با این پروفایل مطابقت دارد؟

یک راه تعیین تطابق مشتری با پروفایل مورد نظر، اندازه‌گیری شباهت یا فاصله آن دو است. بسیاری از فنون داده‌کاوی از ایده اندازه‌گیری شباهت به‌عنوان فاصله استفاده می‌کنند. برای مثال، مدل استدلال بر مبنای حافظه[1]، روشی برای دسته‌بندی مشاهدات بر پایه دسته‌های مشاهده شده‌ای می‌باشد که در همان همسایگی وجود دارند. کشف خوشه خودکار، فن داده‌کاوی دیگری است که بر اساس امکان محاسبه فاصله بین دو مشاهده برای یافتن گروه مشاهدات شبیه به‌هم کار می‌کند. در این مثال، به دنبال تعریف معیار فاصله‌ای برای درجه تطابق مشتری احتمالی با پروفایل موجود هستیم. داده‌ها شامل نتایج پرسشنامه بوده و نشان‌دهنده مشترکین در یک مقطع زمانی می‌باشند. در اینجا با این پرسشها روبرو هستیم: چه نوع معیارهایی مناسب این داده‌ها هستند؟ پروفایل‌ها چه نیازهایی را برآورده می‌کنند؟

دو فرد شرکت‌کننده در تحقیق را در نظر بگیرید. مریم، تحصیل‌کرده بوده، ۹ میلیون تومان در سال درآمد داشته و متخصص است. داوود دیپلمه بوده و ٤ میلیون تومان در سال درآمد دارد. کدام یک بیشتر با پروفایل خوانندگان تطابق دارند؟ جواب، بسته به نحوه مقایسه متفاوت است. جدول (۱۰-۱) راه مناسبی را برای محاسبه امتیاز از روی پروفایل و معیار فاصله نشان می‌دهد.

جدول ۱۰-۱) محاسبه امتیازهای تطابق برای افراد با مقایسه معیارهای دموگرافیک

	خوانندگان نشریه	امتیاز بله	امتیاز خیر	مریم	داوود	امتیاز مریم	امتیاز داوود
تحصیل‌کرده	٪۵۸	۰,۵۸	۰,٤۲	بله	خیر	۰,۵۸	۰,٤۲
متخصص یا مدیر	٪٤٦	۰,٤٦	۰,۵٤	بله	خیر	۰,٤٦	۰,۵٤
درآمد بیشتر از ۸	٪۲۱	۰,۲۱	۰,۷۹	بله	خیر	۰,۲۱	۰,۷۹
درآمد بیشتر از ۱۲	٪۷	۰,۰۷	۰,۹۳	خیر	خیر	۰,۹۳	۰,۹۳
جمع						۲,۱۸	۲,٦۸

[1]- Memory Based Reasoning

این جدول امتیازی بر پایه نسبت توافق مخاطب با هر خصوصیت را محاسبه می‌کند. برای مثال چون ۵۸٪ خوانندگان تحصیل‌کرده‌اند، مریم امتیاز ۰٫۵۸ برای این خصوصیت می‌گیرد. داوود که دیپلمه است امتیاز ۰٫٤۲ می‌گیرد، زیرا ٤۲٪ خوانندگان نیز تحصیلات عالی ندارند. این امتیاز برای هر خصوصیت محاسبه شده و امتیازها جمع می‌شوند. امتیاز مریم ۲٫۱۸ و امتیاز داوود ۲٫٦۸ می‌شود. امتیاز بالاتر داوود نشان می‌دهد او از مریم به خوانندگان فعلی شبیه‌تر است. مشکل این روش این است که با اینکه طبق این امتیازدهی داوود مناسب‌تر از مریم به نظر می‌رسد ولی در واقع مریم به مخاطبین هدف نشریه یعنی افراد تحصیل‌کرده و پردرآمد شبیه‌تر است.

موفقیت این هدف‌گیری از مقایسه پروفایل خوانندگان با خصوصیات جمعیت‌شناسی[1] معلوم می‌شود. این موضوع لزوم برخوردی پخته‌تر با اندازه‌گیری تطابق فرد با مخاطبین نشریه از طریق در نظر گرفتن خصوصیات جمعیت عمومی علاوه بر خصوصیات خوانندگان را ایجاب می‌کند. این روش، درجه تفاوت یک مشتری احتمالی از جمعیت عمومی و تفاوت خواننده از جمعیت عمومی را محاسبه کرده و سپس شباهت این دو تفاوت را اندازه می‌گیرد.

دموگرافی

مطالعه آماری جمعیت (جمعیت شناسی) شامل خصوصیاتی مانند توزیع جغرافیایی، محیط فیزیکی، بیماری، ترکیب جنسیتی و سنی، و نرخ تولد و مرگ.

خواننده نشریه در مقایسه با جمعیت عمومی تحصیل‌کرده‌تر، متخصص‌تر و پردرآمدتر است. درجدول (۱۰-۱) ستونهای شاخص از تقسیم درصد خوانندگانی که ویژگی خاصی دارند بر درصد جمعیتی که دارای این ویژگی است به‌دست آمده‌اند. می‌توان خصوصیات خواننده را با جمعیت عمومی از طریق این شاخصها مقایسه کرد. دیده می‌شود که خواننده نشریه تقریباً سه برابر جمعیت عمومی تحصیل‌کرده است. به‌طور مشابه، خواننده تقریباً نصف جمعیت عمومی تحصیل‌نکرده است. با استفاده از شاخصها به‌عنوان امتیاز هر خصوصیت، مریم امتیاز

---

[1]- Demographic

۸/۴۲(=۰/۹۵+۲/۲۱+۲/۴۰+۲/۸۶) می‌گیرد در حالی که امتیاز داوود فقط ۳/۰۲(=۰/۹۵+۰/۸۷+۰/۶۷+۰/۵۳) می‌شود. امتیازهای بر پایه شاخصها با توجه به جمعیت مخاطب هدف، متناسب‌تر می‌باشند. این امتیازها با معناتر هستند زیرا شامل اطلاعات اضافه تفاوت مخاطبین هدف با جمعیت عمومی می‌باشند.

شاخصها به جای مقادیر خام

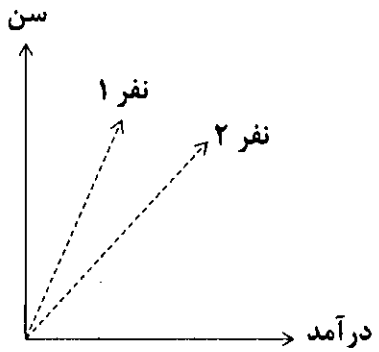
در مقایسه پروفایل مشتری، در نظر گرفتن پروفایل جمعیت عمومی مهم است. برای همین، استفاده از شاخصها اغلب از مقادیر خام بهتر است.

جدول ۱۰-۲) محاسبه امتیازها با در نظر گرفتن نسبتها در جمعیت

	بله			خیر		
	خوانندگان نشریه	جمعیت عمومی	شاخص	خوانندگان نشریه	جمعیت عمومی	شاخص
تحصیل‌کرده	٪۵۸	٪۲۰,۳	۲,۸۶	٪۴۲	٪۷۹,۹	۰,۵۳
متخصص یا مدیر	٪۴۶	٪۱۹,۲	۲,۴۰	٪۵۴	٪۸۰,۸	۰,۶۷
درآمدبیشتر از ۸	٪۲۱	٪۹,۵	۲,۲۱	٪۷۹	٪۹۰,۵	۰,۸۷
درآمدبیشتر از ۱۲	٪۷	٪۲,۴	۲,۹۲	٪۹۳	٪۹۷,۶	۰,۹۵

مفهوم تشابه بر پایه زاویه و تفاوت بر پایه فاصله

می‌توان مفهوم تشابه را بر پایه تفاوت دو زاویه توضیح داد. هر ویژگی اندازه‌گیری شده یک بُعد جدا در نظر گرفته می‌شود. با در نظر گرفتن مقدار متوسط هر ویژگی به‌عنوان مبدأ، پروفایل خوانندگان فعلی، برداری است که نشان می‌دهد خواننده نوعی، چقدر و در چه جهتی از جمعیت عمومی دور است. داده‌های یک مشتری احتمالی نیز یک بردار است. اگر زاویه بین این دو بردار کوچک باشد، آن گاه مشتری احتمالی نیز در همان جهت پروفایل خوانندگان با جمعیت عمومی تفاوت دارد.



شکل ۱۰-۲) شباهت (زاویه) و تفاوت (فاصله)

در شکل ۱۰-۲)، دو نفر (۱ و ۲) از نظر سن و درآمد مقایسه شده‌اند. می‌توان شباهت بین این دو را از طریق زاویه (کسینوس) بین دو بردار سنجید. ولی اگر یکی از آنها دقیقاً دو برابر دیگری سن و درآمد داشته باشد چه طور؟ در این صورت بردار یکی روی دیگری قرار می‌گیرد و شباهت ۱۰۰٪ می‌شود! پس معیار شباهت همیشه مناسب نیست و در بسیاری از موارد بهتر است از تفاوت دو نقطه ۱ و ۲ بر حسب معیار فاصله استفاده شود.

نرمال کردن

به نظر شما آیا اندازه‌گرفتن درآمد بر حسب تومان یا ریال باید تأثیری در شباهت یا تفاوت دو نفر بگذارد؟ برای جلوگیری از این مشکلات معمولاً ابتدا داده‌های هر مشخصه مانند درآمد طوری نرمال می‌شوند که در یک مقیاس قرارگیرد. برای مثال درآمد کلیه افراد را تقسیم بر حداکثر درآمد ممکن می‌کنیم تا کلیه درآمدها بین صفر تا یک قرارگیرند.

مشخصه‌های غیرعددی

چگونه مشخصه‌ای مانند جنسیت را در شباهت در نظر بگیریم؟ کافی است بنا بر قرارداد بگوییم اگر دو نفر هم‌جنس باشند، تفاوت آنها صفر و اگر دارای جنسیت مخالف باشند تفاوت آنها یک است. در این صورت با توجه به اینکه مشخصه‌های عددی را نیز نرمال کرده‌ایم (در فاصله صفر-یک) می‌توان همه مشخصه‌ها را در یک مقیاس در تفاوت یا شباهت در نظر گرفت.

اندازه‌گیری مطابقت برای گروه‌های مشتریان

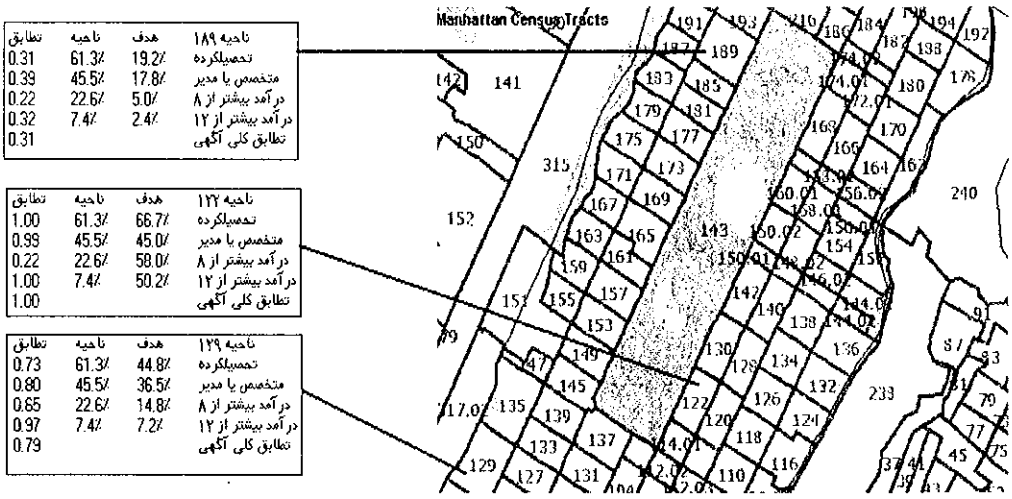
محاسبه امتیازهای بر پایه شاخص، قابل توسعه به گروه‌های بزرگ‌تر مردم است. اهمیت این موضوع در آن می‌باشد که ممکن است ویژگی خاص مورد استفاده برای هر مشتری یا مشتری احتمالی در دسترس نباشد. خوشبختانه و نه بر حسب تصادف، خصوصیات مطرح شده قبلی همگی خصوصیات جمعیت‌شناسی موجود در سرشماری آماری بوده و قابل اندازه‌گیری بر حسب نواحی جغرافیایی مانند منطقه سرشماری می‌باشند.

داده‌های هر منطقه سرشماری

یکی از فروض بازاریابی بر پایه ضرب‌المثل قدیمی "کبوتر با کبوتر، باز با باز، کند هم‌جنس با هم‌جنس پرواز. " می‌باشد. یعنی مردم دارای علایق و سلیقه‌های مشابه در نواحی مشابه زندگی می‌کنند (داوطلبانه یا به دلیل الگوهای تاریخی تمایز). با توجه به این فرض، معامله با مردمی که قبلاً در میانشان و یا در نواحی مشابه مشتری داشته‌اید، ایده خوبی است. آمار سرشماری، هم برای یافتن مکانهای تمرکز مشتری و هم برای تعیین پروفایل نواحی مشابه، باارزش است.

فرایند مورد نظر، نرخ‌گذاری هر منطقه سرشماری با توجه به تطابق آن با نشریه می‌باشد. ایده اصلی، تخمین نسبت هر منطقه سرشماری مطابق با پروفایل خوانندگان نشریه است. برای مثال اگر یک منطقه سرشماری دارای جمعیت بزرگسالی با ۵۸٪ تحصیل‌کرده باشد باز هم امتیاز مریم، ۱ یعنی تطابق کامل است. اگر فقط ۵٫۸٪ تحصیل‌کرده باشند، آنگاه امتیاز تطابق این خصوصیت ۰٫۱ است. امتیاز تطابق کل، متوسط امتیازهای هر خصوصیت است.

شکل (۱۰-۳) مثالی از سه منطقه سرشماری را نشان می‌دهد. هر منطقه دارای نسبت متفاوتی از چهار خصوصیت مورد نظر است. این داده‌ها را می‌توان برای محاسبه امتیاز تطابق کلی هر منطقه ترکیب نمود. توجه کنید که همه ساکنان آن منطقه، امتیاز یکسانی می‌گیرند. این امتیاز بیانگر نسبتی از جمعیت آن منطقه است که مطابق پروفایل مورد نظر می‌باشد.



شکل ۱۰-۳) مثالی از محاسبه تطابق با خوانندگان در سه منطقه سرشماری مانهاتان

استفاده از مشتریان فعلی برای یادگیری در مورد مشتریان احتمالی

یک راه مناسب برای یافتن مشتریان احتمالی، نگاه به همان مکانهایی است که بهترین مشتریان فعلی از آن جا می‌آیند. بنابراین باید راهی برای تشخیص بهترین مشتریان فعلی داشت. لازمه این کار، نگهداری سابقه نحوه دستیابی به مشتریان فعلی و وضعیت آنان در زمان دستیابی می‌باشد.

البته تکیه به مشتریان فعلی برای یادگیری در مورد مشتریان احتمالی، خطر انعکاس تصمیمات بازاریابی گذشته را دارد. مطالعه مشتریان فعلی، پیشنهادی برای جستجوی مشتریان احتمالی در مکانهای جدید نمی‌دهد. با این حال، عملکرد فعلی راه مناسبی برای ارزیابی کانالهای مشتری‌یابی موجود است. برای یافتن مشتریان احتمالی، مهم است بدانیم وقتی مشتریان فعلی، خودشان زمانی مشتری احتمالی بوده‌اند، چگونه به نظر می‌رسیدند. به‌طور ایده‌آل باید:

- ردگیری مشتریان را قبل از مشتری شدن شروع کرد.
- اطلاعات مشتریان جدید را در حین مشتری شدن آنان جمع کرد.
- رابطه بین داده‌های زمان مشتری شدن و نتایج مورد نظر آتی را مدل کرد.

- بخشهای بعدی، این بحث را تکمیل می‌کنند.

شروع ردگیری مشتریان قبل از مشتری شدن

خوب است اطلاعات مشتریان احتمالی قبل از مشتری شدن، ثبت شود. سایتهای وب می‌توانند با ایجاد یک کوکی[1] در اولین بازدید مشتری از سایت و سپس پروفایل کردن این مشتری بی‌نام با ثبت کارهایی که انجام می‌دهد، به این منظور برسند. وقتی بازدیدکننده از طریق همان برنامه پویشگر وب مانند *IE* و روی همان کامپیوتر قبلی، دوباره به سایت رجوع می‌کند، این کوکی شناخته شده و پروفایل به‌روز می‌شود. وقتی این بازدیدکننده تبدیل به مشتری یا کاربر ثبت‌نام کرده می‌شود، فعالیت منجر به انتقال بخشی از سابقه مشتری می‌گردد.

ردگیری پاسخها و پاسخ‌دهندگان در دنیای غیر وِبی نیز کار خوبی است. اولین اطلاعات مهم برای ثبت، پاسخ یا عدم پاسخ مشتری می‌باشد. مشخصات کسانی که پاسخ داده‌اند و آنهایی که پاسخ نداده‌اند، جزء لاینفک مدلهای پاسخ است. در صورت امکان، داده‌های اقدام بازاریابی محرک پاسخ، کانال دریافت پاسخ و زمان پاسخ را نیز باید ثبت کرد.

تشخیص پیام بازاریابی محرک پاسخ از میان پیامهای متعدد، می‌تواند دشوار و یا غیرممکن باشد. برای آسان کردن این تشخیص، فرمهای پاسخ و کاتالوگها دارای کُدهای تشخیص مناسب هستند. وب سایتها آدرس سایتی را که از آن مراجعه شده، ثبت می‌کنند. حتی می‌توان عملیات تبلیغات را از طریق تلفنهای جدا، بسته‌های پستی و یا آدرس سایتهای مختلف تفکیک نمود. بسته به طبیعت محصول یا خدمت، ممکن است پاسخ‌دهندگان ملزم به ارائه اطلاعات اضافه در فرم محصول یا ثبت نام باشند. اگر انجام خدمت مستلزم داشتن اعتبار باشد، اطلاعات اعتبار درخواست می‌شود. اطلاعات جمع‌آوری شده در ابتدای ارتباط با مشتری از هیچ گرفته تا معاینه کامل پزشکی برای بیمه عمر متغیر است. اکثر شرکتها جایی در وسط هستند.

جمع‌آوری اطلاعات از مشتریان جدید

وقتی یک مشتری احتمالی برای اولین بار مشتری می‌شود، فرصتی طلایی برای جمع‌آوری اطلاعات بیشتری پیش می‌آید. داده‌های مربوط به مشتری احتمالی قبل از تبدیل شدن به مشتری،

---

[1]- Cookie

از نوع جغرافیایی و یا دموگرافیک هستند. بعید است لیستهای خریداری شده شامل چیزی غیر از نام، آدرس تماس و منبع لیست باشند. با داشتن آدرس می‌توان اطلاعات دیگری در مورد مشتریان احتمالی بر پایه خصوصیات همسایگان به دست آورد. با داشتن نام و آدرس با هم می‌توان اطلاعات سطح خانوار را از فراهم‌کنندگان داده‌های بازاریابی خرید. این نوع از داده‌ها برای هدف‌گیری بخشهای عمومی مانند «مادران جوان» یا « نوجوانان حومه » مفید است ولی برای ایجاد رابطه فردی با مشتری به اندازه کافی دارای جزئیات نیست.

مفیدترین مشخصات قابل جمع‌آوری برای داده‌کاوی آتی، تاریخ خرید اولیه، کانال خرید اولیه، پیشنهاد پاسخ، محصول اولیه، امتیاز اعتبار اولیه، مدت تا پاسخ و محل جغرافیایی است. این مشخصات در عمل پیش‌بینی کننده خوبی برای مواردی مانند مدت مورد انتظار رابطه، بدحسابی[1] و خریدهای اضافی هستند.

پیش‌بینی نتایج آتی بر اساس متغیر زمان مشتری شدن

با ثبت همه اطلاعات ممکن در زمان مشتری شدن و سپس ردگیری مشتریان در طول زمان، می‌توان از داده‌کاوی برای مرتبط کردن متغیرهای زمان خرید به نتایج آتی مانند طول عمر مشتری، ارزش مشتری و ریسک پیش‌فرض استفاده کرد. این اطلاعات را می‌توان برای هدایت تلاشهای بازاریابی و تمرکز روی کانالها و پیامهای مؤثر (دارای بهترین نتایج) به کار برد. کشف اینکه برخی کانالها، مشتریانی با طول عمر دو برابر کانالهای دیگر می‌یابند، غیر معمول نیست. با فرض امکان تخمین ارزش ماهانه مشتری و داشتن طول عمر، می‌توان ارزش ریالی مشتری هر کانال را محاسبه کرد. برای نرخ‌گذاری کانالها، ارزش ریالی مشتری به اندازه هزینه تماس، مهم است.

۱۰-۱-۳- داده‌های مشتریان

اگر بخواهید از مشتری تصویر درستی داشته باشید باید از جهات مختلف به مشتری بنگرید و همه اطلاعات درباره مشتری را از بخش‌های مختلف سازمان در کنار هم بگذارید، دراین صورت است که این تصویر می‌تواند خصوصیات و رفتار صحیح مشتری را نمایان سازد. برای

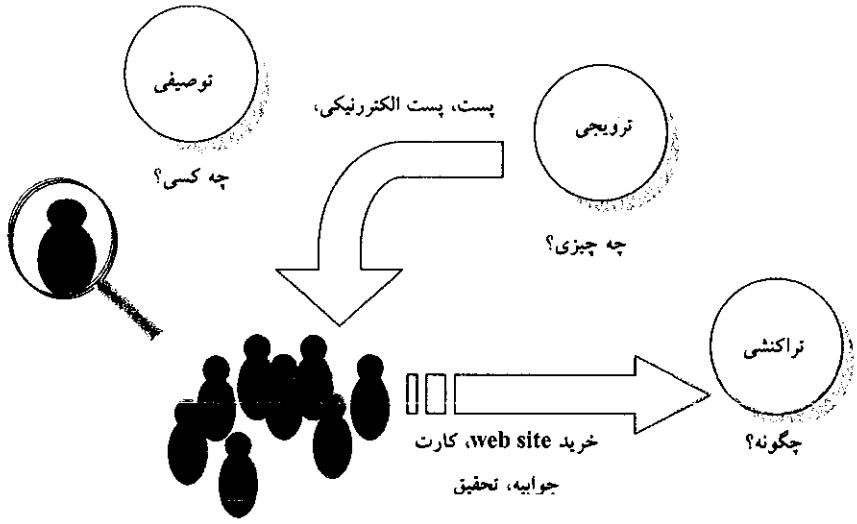
---

[1]- Bad Dept

مثال چون بخش فروش با مشتری ارتباط برقرار می‌کند، تصویری از نیازهای مشتری را دریافت می‌کند. زمانی که مشتری برای حل مشکلی با بخش خدمات تماس می‌گیرد در واقع تصویر دیگری از خود را به سازمان نمایان می‌سازد (رضایت یا عدم رضایت). هر کدام از این نقاط تماس[1] فرصتی برای تعامل هستند و سازمان می‌تواند از طریق آنها با مشتریانش در تماس باشد یا مشتریان با سازمان در تماس باشند، به همین دلیل به آنها نقاط تماس گفته می‌شود. در واقع نقاط تماس بهترین اطلاعات را درباره مشتریان تامین می‌کنند. ولی چون این اطلاعات به‌صورت پراکنده و مجزا وجود دارد، هر قسمت به تنهایی برای ارائه یک تصویرکامل از مشتری کافی نیست. درنتیجه یکپارچگی این داده‌ها که از منابع مختلفی می‌آیند در ارائه تصویر کامل از مشتری بسیار حائز اهمیت است [۱].

داده‌های مرتبط با مشتریان را می‌توان از منابع مختلفی به‌دست آورد شکل (٤-۱۰). در سیستم‌های داده‌کاوی مدیریت ارتباط با مشتری سه نوع اصلی داده وجود دارد که عبارتند از:

- توضیح اینکه مشتری کیست.
- توضیح اینکه چه بازاریابی یا تبلیغ فروشی برای مشتری انجام شده است.
- توضیح اینکه مشتری در مقابل این تبلیغات چه واکنشی نشان داده است.



شکل ۱۰-٤) برای فراهم کردن اطلاعات کافی برای داده‌کاوی باید به دنبال چه کسی، چه چیزی و چگونه باشید.

---

[1]-Touch Point

اگر شما این سه چیز درباره مشتریان یا حتی درباره افرادی که هنوز مشتری شما نیستند بدانید، در واقع داده کافی برای شروع پیش‌بینی را دارید. شما می‌توانید با کمک داده‌کاوی الگوهایی را از میان داده‌ها کشف کنید یا حتی با بهره‌گیری از این تجربیات، تراکنش‌های بازاریابی و فروش با این مشتریان را بهینه کنید. بدون دانستن اینکه مشتری کیست، چه کاری انجام می‌دهد و واکنش آن چیست، بهینه‌سازی یا بهبود سیستم امکان‌پذیر نمی‌باشد.

برای داشتن سودآورترین تراکنش با مشتری تا جای ممکن و بهینه کردن عملکرد سیستم *CRM*، باید توانایی تفکیک مشتریان خوب و بد و مشتریان سودآور و غیرسودآور را داشته باشید. باید بدانید آنها چه کسانی هستند و چه تفاوتهایی با هم دارند. به‌طور مشابه برای اینکه بدانید در تبلیغات و بازاریابی چگونه سرمایه‌گذاری کنید، نیاز به دانستن کارهایی که برای هر مشتری انجام می‌دهید و نتایج پیگیری این کارها دارید. شما باید تعداد زیادی تجربه‌های کوچک در امور تبلیغاتی و بازاریابی با مشتریان اصلی‌تان داشته باشید، و توجه کنید که تفاوت در هرتجربه بهترین راه برای پی بردن به کارهایی که باید انجام دهید و نباید انجام دهید، می‌باشد. برای یافتن ارزش واقعی سیستم، باید بتوانید نتایج را اندازه‌گیری کنید. اگر مشخص نشود که نتیجه تجربه خوب یا بد بوده است، پس واقعا چیز جدیدی که بتواند برای بهبود سیستم در دفعات بعد استفاده شود، حاصل نشده است. گروه‌بندی دسته‌های مختلف تجربه‌ها نیز مفید است چون باعث شکسته شدن داده‌های ذخیره شده در بانک اطلاعاتی به انواع مختلف داده می‌شود. همچنین این دسته‌بندیها توصیف خوبی از منابع داده ارائه می‌دهند. قراردادن داده در این سه دسته به شما برای داده‌کاوی موفق و تولید یک سیستم بهینه *CRM* کمک شایانی خواهد کرد.

داده توصیفی

داده توصیفی شامل توضیحاتی درباره مشتری یا مصرف‌کننده است. این نوع داده‌ها معمولاً به‌طور خلاصه در ستونهای مختلف جدول اطلاعات مشتریان در بانک اطلاعاتی ذخیره می‌شود. این‌گونه داده‌های توصیف‌کننده مشتری شامل اطلاعاتی مثل سن، جنسیت، موقعیت منزل، تعداد فرزندان، درآمد خانگی و درآمد فردی هستند. این اطلاعات قابل تغییر هستند ولی معمولاً زودتر از یکسال عوض نمی‌شود. البته اطلاعاتی مانند آدرس و تلفن نیاز به بروز رسانی فصلی یا

حداقل شش ماه یکبار در بانک اطلاعاتی دارند. این داده‌ها شامل موارد کلی ذیل هستند: دموگرافی، مالی و پروفایل.

رفتار مشتری است که مهم می‌باشد

هدف داده‌کاوی در *CRM*، اصلاح نسبت رفتارهای مثبت به رفتارهای منفی است. اطلاعات ایستا یا همان توصیفی مانند سن یا کُد پستی صرفاً جانشینی برای اطلاعات واقعی مهم یعنی رفتار مشتری هستند. پس چرا فقط از داده‌های رفتاری مانند سابقه خرید مشتری استفاده نمی‌کنیم؟ زیرا در بسیاری از اوقات اطلاعات کافی از سابقه رفتاری مشتریان موجود نیست در حالی که یک مشخصه ایستا مانند کد پستی مشتری، می‌تواند پیش‌بینی کننده خوبی برای بسیاری از رفتارهای مشتری باشد.

داده تبلیغاتی

داده تبلیغاتی شامل اطلاعات کارها و فعالیتهایی است که برای مشتری صورت گرفته است. قدرت این نوع داده معمولاً به پیچیدگی سیستم *CRM* بستگی دارد. این داده‌ها شامل موارد زیر است:

- لیستی از کارهایی است که برای تبلیغات صورت گرفته است مثل پست، کاتالوگ، نمونه‌های تبلیغاتی یا کارتهای تخفیف
- تبلیغات تعاملی مثل تبلیغ در تلویزیون، رادیو، روزنامه، و مجله‌های تبلیغاتی و غیره
- اطلاعات دقیقی مثل ارسال پست الکترونیکی و تعداد کلیکهای کاربران قابل شناسایی در سایتهای وب

انواع اطلاعاتی که می‌تواند جمع‌آوری شود عبارت است از:

- نوع تعامل[1]: فروش، بازاریابی از راه دور، تبلیغات چاپی، تبلیغات رسانه‌ای، تبلیغات وبی.
- توصیف تعامل: مانند رنگ کارت پستال.

---

[1]- Intervention

- رسانه: بازاری که تبلیغات در آن صورت می‌گیرد، وب سایتهایی که آگهی تبلیغاتی در آنها قرار دارد.
- زمان‌بندی: زمان تعامل.
- توصیف قصد و نیت: یک توصیف کامل از اینکه برای چه کسی تعامل معنی دارد و چرا؟ (مثلاً چرا این رنگ یا موسیقی زمینه باید انتخاب شود.)
- مالی: هزینه‌های ثابت و متغیر تعامل.

داده تراکنشی

به‌طورکلی داده‌های تراکنشی، داده‌هایی است که مربوط به تعامل با مشتریان می‌باشد. این داده‌ها می‌توانند هر چیزی از یک تماس تلفنی برای درخواست خدمات گرفته تا توضیح و توصیف محصولاتی که مشتری خریده‌است باشد. این داده‌ها هم مثل داده‌های تبلیغاتی، می‌تواند خیلی سریع در طول زمان تغییر کند. بنابراین طبیعی است که باید داده‌ها در ساختاری ذخیره شود که به سادگی قابل به روز رسانی و تغییر باشد. اطلاعات تراکنشی با اطلاعات توصیفی مشتریان که در طول زمان اساساً تغییر نمی‌کند متفاوت است.

تغییر آرایش داده‌های تراکنشی در فاصله زمانی کوتاهی می‌تواند خیلی چشمگیر باشد. برای مثال معرفی محصولات جدید یا مورد توجه قرارگرفتن محصولات قدیمی و فروش بیشتر آنها، الگوی محصولات فروخته شده را تغییر می‌دهد. این داده‌ها شامل موارد ذیل است: خرید، کلیک صفحه وب، تماس تلفنی، پست الکترونیک، بازدید از مغازه و پست فیزیکی.

لزوم تجمیع داده‌های تراکنشی

بسیاری از روشهای داده‌کاوی نیاز به یک رکورد اطلاعاتی برای هر نمونه آموزشی (در *CRM*، هر مشتری) دارند. داده‌های ایستایی مانند سن و جنسیت برای هر مشتری فقط یک عدد در هر رکورد هستند. درحالی‌که داده‌های واقعه‌ای یا همان تراکنشی مانند سابقه خرید برای هر مشتری ممکن یک یا چند رکورد می‌باشند. برای همین لازم است این داده‌ها از نظر زمانی یا مقداری تجمیع شوند. برای مثال محاسبه شود که هر مشتری از زمان شروع خریدش تا کنون، هر سه ماه چه مبلغی خرید داشته است یا از هر محصول چه تعداد خریداری کرده است. هر مشخصه تراکنشی غیر عددی (مانند نوع محصول) یا ترتیبی، نامزد خوبی برای تجمیع است.

بسته به کاربرد، راههای مختلفی برای تعریف تجمیع وجود دارد (مثلاً ماهانه یا فصلی، متوسط خرید یا حداکثر مقدار خرید) بنابراین می‌توان از روی داده‌های واقعه‌ای تعداد زیادی مشخصه برای هر مشتری ساخت. این موضوع، مسئله بُعد زیاد داده و لزوم استفاده از فنون کاهش بعد را ایجاب می‌کند که قبلاً بحث شده است.

۱۰-۲- برخی کاربردهای داده‌کاوی در مدیریت ارتباط با مشتری

داده‌کاوی ابزاری بنیادی است که برای آشکارسازی خصوصیات جمعیت‌شناختی مشتریان الزامی می‌باشد. از فنون داده‌کاوی می‌توان برای دستیابی به دامنه وسیعی از اهداف مختلف استفاده کرد. چند مثال از کاربردهای آن عبارتند از:

- شناسایی مشتریان سودآور و پروفایل آنان
- پیش‌بینی رفتار خرید مشتری
- رتبه‌بندی رویگردانی مشتری به منظور ارائه برنامه‌های مؤثر حفظ مشتری
- تمرکز تلاشهای بازاریابی بر مشتریان بالقوه‌ای که احتمال خرید کردن بیشتری دارند
- تخمین کارآمدی تبلیغات
- تخمین و اولویت‌بندی ریسک اعتباری مشتری
- برآورد میزان جدی بودن احتمالی مشتری
- فروش کناری[1] و بالاسری[2] به مشتریان بر اساس خرید محصولات قبلی
- هدفگیری مستقیم بازاریابی به سمت افرادی که بیشترین احتمال پاسخ را دارند
- پیش‌بینی کلاه‌برداری و تقلبات
- بهینه‌سازی سهم سبد خرید مشتری

به‌طور خلاصه با به‌کارگیری انواع روشهای داده‌کاوی، سازمان از مفهوم فروش صِرف به سوی خدمت‌رسانی به مشتریان حرکت می‌کند. برخی از این کاربردها به اختصار توضیح داده می‌شوند.

---

[1]- Cross-Selling

[2]- Up-Selling

مدلسازی حفظ و رویگردانی

حفظ باارزش‌ترین مشتریان و دانستن اینکه کدام‌یک در خطر رویگردانی هستند می‌تواند به‌طور چشمگیری سودآوری سازمان را تحت تأثیر قرار دهد. سازمان در این مدل باید از موارد زیر آگاه باشد:

- اینکه کدام مشتریان در حال رویگردانی به سوی رقیب هستند و دلایل آن.
- اینکه کدامیک از ارزشمندترین مشتریانِ سازمان در خطر هستند.
- اینکه آیا سازمان بودجه حفظ مشتریان را برای با ارزش‌ترین مشتریان خرج می‌کند یا خیر.
- اینکه آیا سازمان راههای متناوب معنی‌دار و مؤثری به‌منظور تماس با مشتریان دارد یا خیر.

از دست دادن مشتریان برای شرکتها تا حدی قابل اجتناب است. از طریق تجزیه و تحلیل داده‌ها می‌توان مدلهایی را به منظور پیش‌بینی احتمال خروج مشتریان و احتمال جذب آنان به دیگران از طریق تبلیغات فروش و عملیات تبلیغاتی ساخت.

با این مدلها سازمان می‌تواند با تهیه مشخصه‌های افرادی که در گذشته رویگردان شده‌اند، مشتریانی را تعیین کند که دارای بیشترین گرایش به کاهش یا قطع ارتباطند. مدل حفظ مشتری، پتانسیل یک مشتری را در ماندن با سازمان پس از رخ دادن اتفاقات احتمالی، بررسی می‌کند. مدل رویگردانی، احتمال توقف خریدهای مشتریان فعال را بررسی می‌نماید. با این اطلاعات می‌توان سیاستهای پیشگیرانه‌ای در نظر گرفت و مشتریان را به منظور توجه ویژه به آنان، فعالانه تعقیب یا علامت‌گذاری نمود.

مدلسازی پرهیز از ریسک

در این مدل سعی در پرهیز از کسب مشتریان غیرسودآور است. داده‌کاوی می‌تواند به‌طور تقریبی پیش‌بینی کند که کدام مشتریان بالقوه تبدیل به مشتریان بالفعل می‌شوند. ولی به‌طور خاص مفید است که تعیین شود کدام مشتریان سودآور خواهند بود. این موضوع مهم باید در نظر گرفته شود که به هر حال برخی مشتریان جدید بدهی خود را نخواهند پرداخت و شرکت مجبور به احتساب زیانهای وارده می‌باشد. این مدل، رفتار خرید، رفتار پرداخت، تاریخچه اعتباری و دیگر عوامل را بررسی می‌کند.

مدلسازی فروش جانبی

فروش جانبی و فروش بالاسری در *CRM* اموری محوری محسوب می‌شوند. تعامل سازمان با مشتریان فرصتی اساسی برای بازاریابی محصولات یا خدمات اضافی ایجاد می‌کند. در این زمینه می‌توان داده‌های متفاوتی را در نظر گرفت مانند اینکه مشتریان چه چیزی خریداری می‌کنند، علایق آنها چیست، چه کالاها یا خدماتی مد نظرشان است یا درباره کدام محصولات یا خدمات استعلام می‌کنند. در ابتدا باید در نظر گرفت که کدام محصولات و یا خدمات توان بالقوه‌ای در فروش جانبی دارند. این کار با قضاوت بر مبنای داده‌های فروش گذشته انجام می‌گیرد. به‌عبارت دیگر، تعیین می‌شود که یک مشتری منفرد کدام محصولات را بیشتر خریداری کرده است. سپس تعیین می‌گردد که پروفایل چه مشتریانی بیشترین تناسب را با گروه‌های متنوع محصول دارد. با این کار مشتریانی شناسایی می‌شوند که بیشترین تمایل را به خرید محصولات مشابه دارند ولی هنوز شرکت به این‌گونه مشتریان توجهی نشان نداده است. در این حالت سازمان فرصتی عالی برای تبلیغ آن محصولات به مشتریان فعلی خود پیدا می‌کند. تمام این فرایند در محدوده *CRM* می‌باشد.

مدلسازی سودآوری

در این روش، ارزش طول عمر *(LTV)*[1] مشتری اندازه‌گیری و تنظیم می‌شود تا معلوم گردد مشتریان در دوره طول عمر خریدشان چه چیزی را خریداری می‌کنند یا پتانسیل خرید چه چیزهایی را دارند. این روش یکی از روشهای تحلیلی داده‌کاوی است که از شیوه‌های محاسبه *LTV* و استفاده از آن برای دسته‌بندی کردن مشتریان استفاده می‌کند. فهم عناصر اصلی سودآوری کمک می‌کند تا سازمان درک نماید چه زمانی مشتریان سودآور خواهند شد و آیا اصلاً این اتفاق (سودآور شدن مشتریان) روی خواهد داد یا خیر.

مدلسازی تجزیه و تحلیل اینترنتی

رفتار اینترنتی، یعنی چگونگی گشت و گذار مشتری درون صفحات اینترنتی مربوط به شرکت را می‌توان به منظور فهم رفتار و ترجیحات مشتری ضبط و تجزیه و تحلیل نمود.

---

[1]- Life Time Value

ترجیحات مشتری شامل موارد زیر است: زمانی که مشتری صرف مشاهده یک صفحه می‌کند، کدام پیوندها را انتخاب می‌کند، به کدام آگهی‌ها توجه بیشتری دارد، از طریق کدام صفحه وارد سایت شرکت شده و از کدام صفحه خارج می‌گردد و نظایر آن.

همهٔ این موارد به اطلاعات آماری پایگاه داده‌های شرکت تبدیل می‌شوند. این داده‌ها از این نظر مفیدند که به کمک آنها شرکتها می‌توانند مکانیزم فروششان را بشناسند و بدانند که چگونه محصولات و دیگر اقلام جانبی را که ممکن است مورد علاقه مشتریان باشد مکان‌یابی نمایند. این روش همچنین درصد ترک صفحه را نیز نشان می‌دهد. یعنی معلوم می‌کند که بازدیدکنندگان از روی کدام صفحات می‌پَرند، از کدام صفحات اجتناب می‌کنند یا کدام صفحات موجب می‌شود که بازدیدکننده سایت وب شرکت را ترک کند.

تأکید این اطلاعات بر صفحاتی است که نیازمند ارزیابی بیشتر از نظر محتوا و سادگیِ استفاده می‌باشند. البته با اینکه نرم‌افزارهای قوی تجزیه و تحلیل اینترنتی فراوانند ولی هیچ‌کدام از آنها قادر نیستند به تنهایی تصویر کاملی از رفتارهای بازدیدکنندگان سایت ارائه دهند. مثلاً نمی‌توان به راحتی فهمید که آیا بازدیدکننده‌ها محصولات را برای خودشان می‌خرند یا به‌عنوان هدیه و برای دیگران می‌خرند، اینکه مشتریان سودآور هستند یا فقط در سایت به گشت و گذار می‌پردازند و غیره. بنابراین دموگرافی بازار هدف را نمی‌توان به تنهایی با استفاده از ابزار تجزیه و تحلیل اینترنتی انجام داد.

بازاریابی مستقیم

هدف تبلیغات آن دسته از مشتریان احتمالی است که در مورد تک‌تک آنان اطلاعات کافی نداریم. بازاریابی مستقیم حداقل نیاز به مقداری اطلاعات اضافه مانند نام و آدرس یا تلفن یا پست الکترونیک دارد. در بسیاری از کشورها، داده‌های قابل توجهی درباره بخش بزرگی از جمعیت در دسترس است. قبل از برنامه‌ریزی برای استفاده در بازاریابی، لازم است دسترسی به داده‌ها در بازار مورد نظر و محدودیتهای قاعدهی استفاده، بررسی شود. مشکل این است که حتی از طریق غربالهای بدیهی، نسبت مشتریان مورد بررسی به مشتریان احتمالی پاسخ‌دهنده، بسیار زیاد باشند. بنابراین یکی از کاربردهای اصلی داده‌کاوی برای یافتن مشتریان احتمالی،

«هدف‌گیری» است. یعنی به دنبال یافتن آن مشتریان احتمالی هستیم که احتمال بیشتری دارد به یک پیشنهاد خاص پاسخ دهند.

فعالیتهای بازاریابی مستقیم عموماً نرخ پاسخی کمتر از ۱۰٪ دارند. این مدلها با تشخیص آن دسته از مشتریان احتمالی که احتمال پاسخشان به مشتری‌یابی مستقیم بیشتر است، نرخهای پاسخ را بهبود می‌بخشند. مفیدترین مدلهای پاسخ، تخمینی واقعی از احتمال پاسخ می‌دهند. البته الزامی برای محاسبه احتمال واقعی پاسخ نیست بلکه داشتن مدلی که مشتریان احتمالی را به ترتیب بیشترین امتیاز پاسخ، رتبه‌بندی کند کافی است. با داشتن یک لیست مرتب شده می‌توان درصد پاسخگویان در یک فعالیت بازاریابی مستقیم را با پست به افراد بالای لیست یا تماس با آنها، زیاد کرد.

۱۰-۲-۱- لایه‌های کشف الگو

وقتی داده از انباره‌داده استخراج شد و مرحله آماده‌سازی را طی کرد، هرکدام از روشهای داده‌کاوی می‌تواند برای پاسخ به سؤالات کسب وکاری که در ذهن است استفاده شود. دسته‌بندی، خوشه‌بندی، رگرسیون، الگوهای مکرر[1]، قواعد تلازمی و سریهای زمانی مرسوم‌ترین روشهای داده‌کاوی هستند. همان‌طور که در جدول (۱۰-۳) نشان داده شده کشف الگو شامل چهار لایه مهم است [۲].

جدول ۱۰-۳) لایه‌های کشف الگو

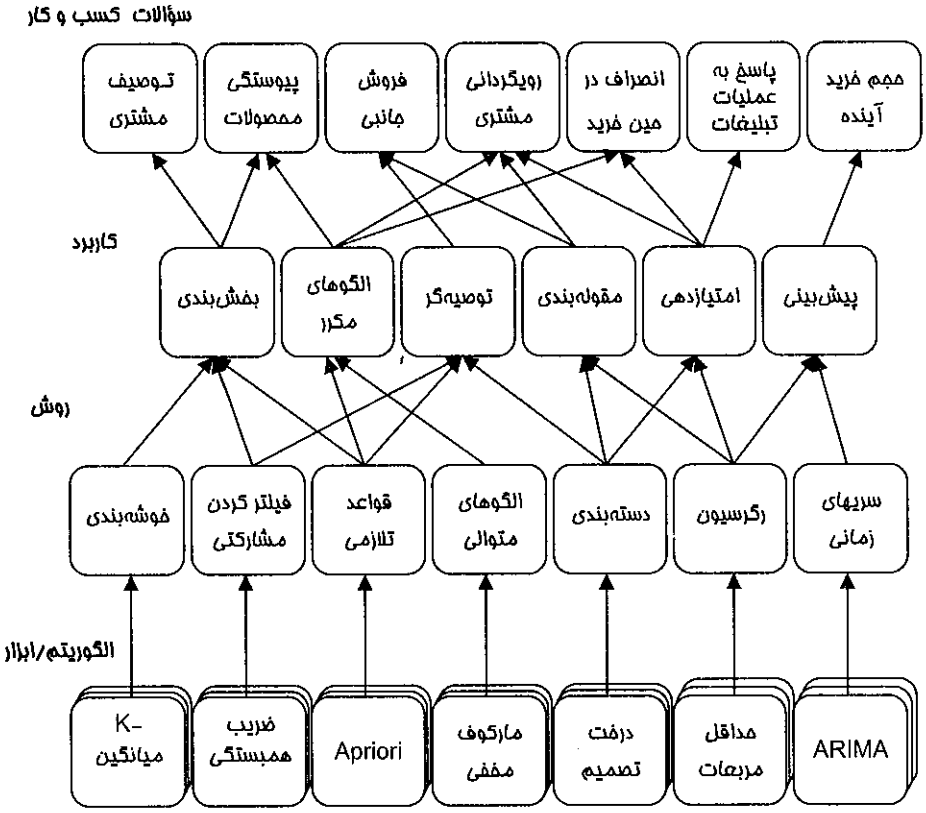
لایه اول	**سؤالهای تجاری** مانند توصیف مشتری
لایه دوم	**کاربردها** مانند امتیازدهی، پیش‌بینی برای حل یک نوع خاص سؤال کسب وکار.
لایه سوم	**روشها** مانند سریهای زمانی، طبقه‌بندی باتوجه به نوع داده خروجی که نتیجه فرآیند بر روی داده ورودی است
لایه چهارم	**الگوریتمها**

[1]- Frequent Pattern

سؤال استراتژیک و سؤال عملیاتی

سؤال استراتژیک، سؤالی است که جواب آن راهنما یا تأثیرگذار بر یک تصمیم است. یک مثال می‌تواند تحلیل برای تعیین محرکهای کلیدی رضایت مشتری باشد. در مثال دیگر می‌توان با خوشه‌بندی بر حسب کل دفعات خرید هر مشتری، مشتری با ارزش را تعریف کرد. سؤال عملیاتی، سؤالی است که نتایج آن مستقیماً برای افزایش وقایع سودآور استفاده می‌شود. برای مثال، در مدلسازی پاسخ بازاریابی مستقیم، خود امتیازهای مدل مهم‌ترین نتیجه هستند. اگر چه داشتن الگو یا مدل قابل توضیح، مطلوب است، ولی خود مدل هدف تحلیل می‌باشد.

درشکل (۱۰-۵) این لایه‌ها به‌طور جزئی‌تر نشان داده شده‌اند. البته امکان نمایش تمام حالات ممکن نمی‌باشد.



شکل ۱۰-۵) لایه‌های کشف دانش درمدیریت ارتباط با مشتری

دسته‌بندی (در مدیریت ارتباط با مشتری)

آیا می‌توان از روی خصوصیات ظاهری، تشخیص داد که یک فرد زن است یا مرد؟ این سؤالی است که دسته‌بندی به دنبال جواب آن است. چگونه می‌توان این کار را کرد؟ ابتدا باید تعدادی زن و مرد داشته باشید که خصوصیات فیزیکی و الگوهای رفتاری آنها ثبت شده باشد. حالا اگر فرد جدیدی را ببینید، می‌توانید از روی شباهت خصوصیات او را به یکی از دو دسته زن یا مرد، تخصیص دهید. برخی اوقات به دلیل کامل نبودن اطلاعات و یا غیرعادی بودن فرد، امکان اشتباه نیز وجود دارد.

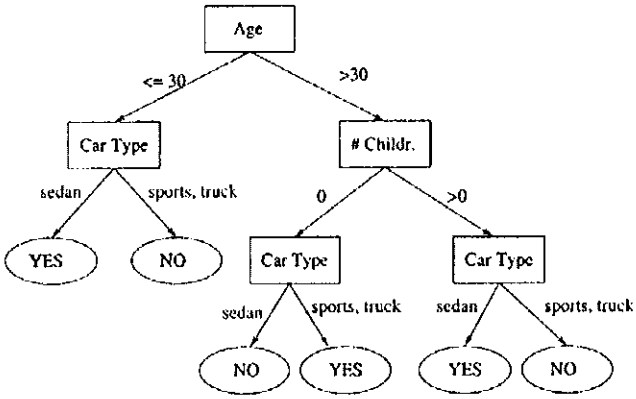
اگر دسته‌ها از قبل مشخص باشند (مثلاً اینکه خرید کرده یا نکرده) می‌توان به کمک روش دسته‌بندی ابتدا به‌وسیله داده‌های موجود (داده‌های آموزشی) مدلی ساخته و از آن برای پیش‌بینی موارد جدید استفاده کرد.

درخت تصمیم (در مدیریت ارتباط با مشتری)

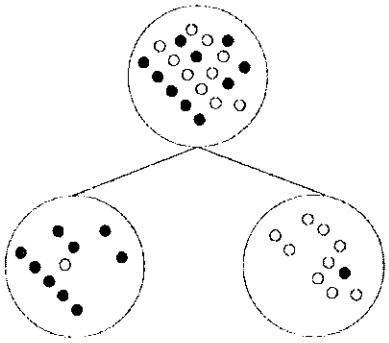
درخت تصمیم به دلیل سادگی بیان دانش یافته شده و قدرت پیش‌بینی مناسب، یکی از پرکاربردترین الگوریتمهای دسته‌بندی می‌باشد. دانش یافته شده به شکل قواعد اگر-آنگاه بیان می‌شود.

مسابقه ۲۰ سؤالی، نمونه‌ای از درخت تصمیم می‌باشد. در هر مرحله از مسابقه با پرسیدن یک سؤال، فضای جواب را محدودتر می‌کنیم تا در نهایت به جواب مشخصی که در ذهن طرف مقابل است برسیم. چه سؤالی را ابتدا انتخاب می‌کنید؟ سؤالی را ابتدا انتخاب می‌کنیم که فضای جواب را به خوبی به دو قسمت تقسیم کند.

مانند روشهای دیگر دسته‌بندی، هدف در درخت تصمیم، تعیین دسته (مثلاً خرید یا عدم خرید مشتری) از روی مشخصات می‌باشد. برای اینکار ابتدا مشخصه‌ای (متغیری) انتخاب می‌شود که با گذاشتن شرط بیشتر یا کمتر از یک مقدار روی آن، بهتر از همه متغیرهای دیگر قدرت تفکیک کلیه رکوردهای موجود را داشته باشد.



برای مثال هدف در شکل بالا، پیش‌بینی اشتراک افراد در یک مجله است. با بررسی روی تک‌تک مشخصات مشتری معلوم شده است که بهترین مشخصه تفکیک کننده مشتری، سن مشتری و آستانه تصمیم آن سن ۳۰ سال است. طبق شکل ذیل با بررسی شرط بیشتر بودن سن از ۳۰ سال، ۲۰ فرد موجود را که ۱۰ نفر از آنها مشتری هستند، به دو دسته تقسیم می‌شوند که در یکی از ۱۰ نفر، ۹ نفر مشترک هستند و در دیگر از ۱۰ نفر، ۹ نفر مشترک نیستند.



این فرایند برای شاخه‌های زیرین درخت دنبال شده و هر بار مشخصه‌های دیگری انتخاب می‌شود که بهتر از بقیه، داده‌ها را تفکیک می‌کند. این روند تا رسیدن به حد قابل قبولی از خطا یا رسیدن به دسته‌های ۱۰۰٪ خالص ادامه می‌یابد.

یکی از موارد مهم در دسته‌بندی، استحکام می‌باشد. همان‌طورکه متوجه شده‌اید امکان اشتباه در دسته‌بندی وجود دارد. بنابراین پس از ساختن مدل باید خطاهای دسته‌بندی را بررسی کرد تا در حد مقبولی باشند. ممکن است مدل برای داده‌های آموزشی خیلی خوب و کم‌خطا دسته‌بندی

کند ولی در مورد داده‌های جدید (مثلاً تصمیم‌گیری در مورد مشتریان جدید) دارای خطای زیادی باشد. معیار ارتباط خطای آموزش به خطای استفاده در عمل، استحکام نام دارد. مدلی خوب است که دارای استحکام بالایی باشد.

خوشه‌بندی (در مدیریت ارتباط با مشتری)

انسان به‌طور فطری تمایل به گروه‌بندی اشیاء و مفاهیم بر اساس شباهت یا تفاوت (فاصله) دارد. برای مثال کودک می‌آموزد که موجودات زنده دو دسته هستند: آنهایی که سبزند و راه نمی‌روند (گیاهان) و آنهایی که معمولاً قهوه‌ای‌اند، حرکت می‌کنند و احتمالاً خطرناکند (حیوانات)!

گروه‌بندی در ادبیات داده‌کاوی، خوشه‌بندی[1] یا خوشه‌یابی نامیده می‌شود. روش خوشه‌بندی برای توصیف گروه‌های مختلف در مجموعه داده به‌کار می‌رود و بر خلاف دسته‌بندی خوشه‌ها از قبل مشخص نیستند. معمولاً ابتدا خوشه‌بندی انجام شده و بعد خوشه‌ها به‌عنوان نام دسته‌ها برای دسته‌بندی به‌کار می‌روند.

پس از خوشه‌بندی، برای تعبیر خوشه‌ها، نماینده هر خوشه را در نظر می‌گیرند. نماینده می‌تواند یکی از مشاهدات میانه خوشه و یا میانگین همه مشاهدات یک خوشه در نظر گرفته شود. با توجه به تفاوت و شباهت نماینده هر خوشه به نماینده کل داده‌ها از نظر مشخصات (مانند سن، جنسیت، . . .) می‌توان هر خوشه را تعبیر کرد. مثلاً به یک خوشه نام«مادران جوان » و به خوشه دیگر نام « نوجوانان حومه » را داد. در *CRM* گروه‌بندی از آن جهت اهمیت دارد که ایجاد پروفایلی برای مشتریان مختلف را فراهم می‌کند و امکان برنامه‌ریزی استراتژیک روی گروه‌های مشتریان را می‌دهد. از طرف دیگر امکان انواع گروه‌بندیهای دیگر مانند محصولات مشابه را ایجاد می‌کند.

ممکن است بسیاری از خوشه‌های یافته شده، بدیهی و فاقد دانش جدید و مفیدی برای ما باشند. بنابراین لازم است خوشه‌های یافته شده از نظر جالب بودن[2] طبق نظر خبرگان بررسی شوند.

---

[1]- Clustering

[2]- Interestingness

گاهی اوقات، نشان دادن شباهت و تفاوت خوشه‌های مختلف با هم مفید می‌باشد. راه متداول برای اینکار شبکه‌های کوهونن یا همان *SOM* است.

خوشه‌بندی هدف‌گذاری شده (در مدیریت ارتباط با مشتری)

هر چند گروه‌های تشکیل شده بر اساس مشخصه‌هایی مانند سن، جنسیت یا رفتار قابل تمایز از هم هستند، ولی این ممکن است این گروه‌ها در بافت کسب و کار بامعنا نباشند. اگر همه گروه‌ها دارای عمر مشتری متوسطی بوده یا همه مقادیر یکسانی خرید کنند، این گروه‌بندی مفید نیست. خوشه‌بندی هدف به دنبال یافتن خوشه‌هایی است که با توجه به مقدار یک متغیر هدف خاص متفاوت هستند. این کار با افزایش وزن (اهمیت) متغیر هدف در معیار فاصله یا کُد کردن متغیرها با توجه به هدف انجام می‌شود. در حالت حدی که فقط متغیر هدف در معیار فاصله در نظر گرفته می‌شود، خوشه‌بندی شبیه دسته‌بندی می‌شود. فرق مهم خوشه‌بندی هدف و دسته‌بندی، این است که مقوله از پیش تعریف شده‌ای برای خوشه‌بندی هدف وجود ندارد.

رگرسیون و سریهای زمانی (در مدیریت ارتباط با مشتری)

روش رگرسیون سعی در پیش‌بینی یک خروجی پیوسته با استفاده از یک تابع دارد که میزان خطا از الگو را در داده‌ها نشان می‌دهد. روشهای رگرسیون را می‌توان برای کاربرد دسته‌بندی دوتایی و همچنین برای امتیازدهی و پیش‌بینی استفاده کرد. روشهای سری‌زمانی مثل رگرسیون مقادیر پیوسته را پیش‌بینی می‌کنند، ولی در طول زمان، گرایشها و چرخه رفتار را نیز مدل می‌کند.

قواعد تلازمی (در مدیریت ارتباط با مشتری)

قوعد تلازمی و الگوهای مکرر، وقایعی که در مجموعه داده اتفاق می‌افتد را توصیف می‌کنند. قواعد تلازمی، ارتباط بین موارد موجود در یک مجموعه داده است بدون اینکه ترتیب زمانی یا ترتیب خاصی داشته باشند.

فیلتر کردن مشارکتی[1]

فیلتر کردن مشارکتی وقایع را برای یافتن مجموعه‌هایی که شبیه یکدیگرند تحلیل می‌کنند. در مفهوم *CRM*، وقایع معمولاً بر اساس مشتریان گروه‌بندی می‌شود. ایده اصلی این است که اگر مشتریان با سابقه خرید مشابه کالای خاصی را بخرند، ممکن است مشتری جدید دارای همان مشخصات نیز همان کالا را بخرد. برای هر مشتری در مجموعه داده، با روش فیلتر کردن مشارکتی یک گروه از مشتریان پیدا می‌شوند که گزارشات وقایعشان به یکدیگر شبیه هستند. وقتی گروه شکل می‌گیرد، وقایع می‌تواند با چیزی که در کل گروه عمومیت دارد، امتیازدهی شود. خروجی فیلتر کردن مشارکتی می‌تواند هم مشتریان شبیه به هم و هم وقایع امتیاز داده شده باشد. برای مثال، اگر واقعه، خرید محصولات باشد، امتیازدهی وقایع می‌تواند برای توصیه محصولات به کار برده شود. هر محصولی که بیشترین امتیاز را بیاورد توصیه می‌شود. فیلتر کردن مشارکتی مشابه خوشه‌بندی برای کشف گروه‌های طبیعی در مجموعه داده است. تفاوت مهم فیلتر کردن مشارکتی این است که گروه‌های متفاوتی گرداگرد یکدیگر در مجموعه داده شکل می‌گیرند. در واقع هر مشتری مرکز یک گروه واحد است و مشتری A ممکن است در گروه مشتری B ظاهر شود ولی عکس آن درست نباشد. در نتیجه روش فیلتر کردن مشارکتی برای شخصی‌سازی سیستم به کار برده می‌شود.

---

[1]- Collaborative Filtering

منابع

1) *Berson A. , Smith S. ,Thearling K. (2001) "Bulding Data Mining Application for CRM" , McGraw-Hill.*

2) *Ye N. (2003) "THE HANDBOOK OF DATA MINING", LAWRENCE ERLBAUM ASSOCIATES , PUBLISHERS Mahwah, New Jersey London.*

3) *Rygielski C. , Wang J. C. , Yen D. C. (2002), "Data mining techniques for customer relationship management", Technology in Society 24 (2002) 483–502.*

4) *Berry M. J. A. , Linoff G. S. (2004) Data Mining Techniques for Marketing, Sales, and Cutomer Relationship Management, Second edition, Wiley.*

۵) سمیرا ملک محمدی (۱۳۸۵)، سمینار کارشناسی ارشد، کاربرد داده‌کاوی در مدیریت ارتباط با مشتری دانشگاه علم وصنعت، استاد راهنما دکتر مهدی غضنفری.

واژه نامه

ا

آزمونهای ردیف	Run tests
آشفتگی و آلودگی	Dirty
آماره انحراف	Deviance Statistic
آنتروپی	Entropy
ابتکار سیاسی اجتماعی محلی	Grass- Root Political Initiative
اَبَر مکعب	Hypercube
ابرمتنها	Hypertext
اتصال جانبی	Lateral
اجتماع کاوی	Community Mining
اجتماعات برخط	On Line Forums
اجتماعات وبی	Web Communities
ارتباط رکورد	Record Linkage
ارتباط مکانی	Topologic
ارزش شبکه‌ای	Network Values
LTV ارزش طول عمر	Life Time Value
استخراج	Extract
استخراج، تبدیل، بارگذاری	ETL (Extraction/Transformation/Loading)
استدلال مبتنی بر مورد	Case Base Reasoning
استراتژی تمرکز بر مشتری	Customer Focused Strategy
استقراء مدل	Model Induction
استنتاج درخت تصمیم	Decision Tree Induction
اسمی	Nominal
اشیاء	Objects
اطلاعات	Information
اطمینان	Confidence
اغتشاش	Noise

افزونگی	Redundancy
افزونه	Redundant
اقدام بازاریابی	Marketing Action
التهاب مغزی	Meningitis
الحاق ترجیحی	Preferential Attachment Model
الگوها	Patterns
الگوهای مکرر	Frequent Pattern
امتیاز گره‌ها	Nodes Degrees
انبارهٔ بنگاه	Enterprise Warehouse
انباره داده ها	Data Warehouse
انباره مجازی	Virtual Warehouse
انتخاب مشخصه‌ها	Feature Selection
انتخاب و استخراج رابطه	Relation Selection & Extraction
انتقال	Transform
انحراف	Deviation
ایده ها	Notions

ب

بارگذاری	Load
بازارچه داده‌ها	Data Mart
بازاریابی انبوه	Mass Marketing
بازاریابی مستقیم	Direct Marketing
بازاریابی ویروسی	Viral Marketing
بازجانشانی	Resubstitution
بازیافت اطلاعات	Information Retrieval
باستان شناسی داده ها	Data Archaeology
بخش	Segment
بخش بندی	Segmentation
بدون اتلاف	Lossless
بدیع	Novel

بدیع بودن	Novelty
برازندگی	Fitness
برچسب	Label
برچسب دسته	Class Label
برنامه‌ریزی منطقی قیاسی	Inductive Logic Programming
بزرگترین زیردنباله‌های مشترک	Longest Common Subsequences Similarity LCSS
بن بست	Gridlock
به خوبی قابل درک	Well Understood
بهترین واحد انطباق	BMU: Best Matching Unit
به‌روزرسانی	Refresh
به‌موقع	Timely
بهینهٔ سراسری	Global Optimum
بهینهٔ محلی	Local Optimum
بی‌ثبات	Non Stationary
بیش‌برازش	Over Fit

پ

پاکسازی داده ها	Data Cleaning
پایایی	NON Volatile
پایداری	Robustness
پراکندگی	Dispersion
پردازش تحلیلی برخط	On_line Analytical Processing
پروفایل	Profile
پس پردازش	Post-Process
پس‌انتشار	Back Propagation
پشتیبان	Support
پویاپذیری	Time Variant
پیش پردازش	Preprocess

پیشبینی	Prediction
پیوند مرکزی	Centroid
پیام آنی	Instant Messaging
پیش‌بینی	Predictive
پیشگویی تصادفی	Random Predictor
پیوند	Linkage
پیوند تکی	Single Link
پیوند کامل	Complete Link
پیوند متوسط	Average Link
پیوندکاوی	Link Mining
پیوندهای خروجی	Out-Links

ت

تاباندن زمانی پویا	Dynamic Time Warping: DTW
تأثیر مثبت گفتار شفاهی	Word of Mouth
تبدیل	Transformation
تبدیلات داده	Data Transformations
تپه توردی	Hill Climbing
تثبیت	Consolidation
تجزیه فوریه	Discrete Fourier Transform: DFT
تجزیه مقدار منفرد	Singular Value Decomposition: SVD
تجزیه موجک	Discrete Wavelet Transform (DWT)
تجزیه‌ای یا تقسیمی	Divisive
تجمیع	Aggregation
تجمیع مشخصه‌ها	Feature Aggregation
تجمیعی	Agglomerative
تحلیل سبد بازار	Market Basket -Basket Data
تحلیل مولفه‌های اصلی	Principal Component Analysis: PCA
تخمین قطعه‌ای خط	Line Segment Approximation
تخمینهای خطی قطعه‌ای	Piecewise Linear Approximations: PLA

ترتیب حروف الفبا	Lexicographic Order
ترکیب	Combination
تشخیص الگو	Pattern Recognition
تشخیص مغایرت	Discrepancy Detection
تشکیل مجدد شماتیک	Reformulation Schema
تصمیم‌گیری علمی	Decision Science
تصویر کردن	Data Projection
تصویر کلی	Global Representation
تصویر آنی	Snapshot
تصویر کردن	Projections
تصویر کردن تصادفی	Random Projection
تضادها	Conflicts
تطبیق	Adaptation
تعمیم	Generalization
تفسیر پذیری	Interpretability
تلخیص توصیفی داده‌ها	Descriptive Data Summarization
تنبل	Lazy
توابع فعال سازی	Activation Function
توزیع‌های دم‌پهن	Heavy-Tailed
توصیفی	Descriptive
توضیحات ویژگی – ارزش	Attribute Value Description

ث

ثروتمند، ثروتمندتر می‌شود	Rich-Get-Richer

ج

جاذب چگالی	Density Attractors
جالب بودن	Interestingness
جامع	Exhaustive
جامعیت	Integrated

جدول احتمال شرطی	Conditional Probability Table
جدول تصادفی	Contingency Table
جرم حیاتی	Biomass
جریان کلیکها	Clickstream
جعلی	Spurious
جمع شدن قطر	Shrinking Diameter
جمع و جور بودن مدل	Compactness

چ

چارکها	Quartiles
چندک	Quantile
چندک چندک	Quantile –Quantile (Q-Q)

ح

حساسیت	Sensitivity
حقایق	Facts

خ

خطاهای عمدی	Deliberate Errors
خلاصه سازی	Summarization
خوشه بندی	Clustering
خوشه‌بندی فضایی بر پایه چگالی برای داده‌های مغشوش	Density- Based Spatial Clustering of Applications with Noise

د

داده	Data
داده کاوی بر مبنای شبکه‌های عصبی مصنوعی	Neural Network Data Mining
داده های پرت	Outliers

داده‌روبی	Data Scrubbing
داده‌های مفقوده	Missing Data
دادهٔ کافی	Sufficient Data
دامنه میان چارکی	Interquartile range (IQR)
دانش	Knowledge
درجه جذابیت	Degree of Interest
درو کردن داده ها	Data Harvesting
دسته‌بندی	Classification
دسته‌های ازپیش‌تعیین‌شده	Predefined Classes
دسته‌های گسسته	Discrete Classes
دقت	Precision
دقیق	Accurate
دنیای کوچک	Small World

ر

رتبه‌ای	Ordinal
رقابت	Competition
رگرسیون	Regression
روش بسته‌بندی	Binning
روش شبکه اطلاعات آماری	A Statistical Information Grid Approach
روش نگاشت سریع	FastMap
روشهای انتخاب نقطه انشعاب	Split Selection
روشهای سلسله مراتبی	Hierarchical
روشهای مبتنی بر چگالی	Density based
روشهای مبتنی بر ناخالصی	Impurity – Based

ز

زبان پرس و جوی استاندارد	Standard Query Language
زمان تأخیری	Time- Lagged
زیر گرافها	Subgraph

س

ساخت مشخصه‌ها	Feature Construction
ساخت ویژگی	Attribute Construction
ساختاری	Structural
سادگی	Simplicity
سازگار	Consistent
سایتهای شراکت دانش	Knowledge- Sharing Sites
سربار شدن فیلد	Field Overloading
سرعت	Speed
سطلها	Bucket
سود اطلاعاتی	Information Gain
سوگیری	Biased
سیستم مدیریت پایگاه داده	Data Base Management System
سیستمهای اطلاعات مدیریت	Management Information System (MIS)
سیستمهای پیچیده	Complex Systems
سیستمهای عملیاتی یا میراثی	Legacy
سیستمهای منبع	Source System
سیگموئید	Sigmoid Scaling

ش

شاخص جینی	Gini Index
شبکه‌های وابستگی احتمالی	Probabilistic Dependency Network
شبکه‌های عصبی مصنوعی	Artificial Neural Networks
شبکه اجتماعی چندرابطه‌ای	Multi Relational Social Network
شبکه اعتماد	Web of Trust
شبکه چندلایه پیشخور	Feed Forward
شبکه رقابتی	Competitive Network
شبکه نردبان	Lattice
شش سطح جدایی	Six Degrees of Separation

شفافیت	Specificity
شمارش	Enumeration

ص

صحت	Accuracy

ط

طبقه ای	Categorical

ع

عرض ثابت	Equal-width
عمق ثابت	Equal-depth
عملیات هرس	Pruning

ف

فاصله دسترسی	Reach ability- Distance
فاصله متوسط	Average Distance
فاصله مرکزی	Core- Distance
فاصله‌ای	Interval
فایده	Usefulness
فرایندی تکراری	Iterative
فرض دنیای باز در مقابل دنیای بسته	Closed Versus Open World Assumption
فروش بالا سری	Up-Selling
فروش کناری	Cross-Selling
فعالیت تبلیغی	Campaign
فوق خطی	Superlinearly
فهم داده	Data Understanding
فهم کسب و کار	Business Understanding
فیلتر کردن مشارکتی	Collaborative Filtering

ق

قابل دسترس چگال	Density Reachable
قانون تراکم توانی	Densification Power Law
قانون تهی	Null Rule
قانون رشد توانی	Growth Power Law
قانون یکتایی	Unique Rule
قطر مؤثر	Effective Diameter
قواعد استنتاج شده	Induced Rules
قواعد تلازمی	Association Rules
قواعد قوی	Strong Rule
قواعد مکرر	Frequent Rule

ک

کارت	CART
کاملا متصل	Fully Connected
کاهش	Reduction
کاهش بُعد	Dimensionality Reduction
کاهش تعدد نقاط	Numerical Reduction
کسب دانش	Knowledge Acquisition
کشف دانش و داده‌کاوی	Knowledge Discovery and Data Mining (KDD)
کلمات ابَرپیوندی	Hyperlink
کمّی	Quantitative
کیفیت زندگی	Quality-of-Life
کتاب‌شناسی	Bibliography
کمی‌سازی برداری	Vector Quantization
کوپن	Coupon

گ

گراف خطی	Line Graph
گراق بدون دور	Acyclic Graph
گرایش مرکزی	Central Tendency
گروه‌های خبری	Use Net Groups
گره سفیر	Ambassador
گسسته سازی	Discrimination
گواهی	Evidence
گوناگونی	Diversity

ل

لایروبی داده ها	Data Dredging
لفاف	Wrapper

م

ماتریس اغتشاش	Confusion Matrix
ماشینهای بردار پشتیبان	Support Vector Machine (SVM)
متدلوژی CRISP	Cross Industry Standard Process for Data Mining
متصل چگال	Density Connected
متقابلاً ناسازگار	Exclusive
متن لنگر	Anchor Text
مثبت درست	True Positive
مثبت غلط	True Negative
مجموع تمام مسیرها	Ensemble of All Paths
مجموع مربع خطاها	Sum Square Error
مجموعه داده‌های ناهمگن	Heterogeneous
مجموعه کاهش یافته	Reduced Set
محدودیت‌های شیب	Slope Constraints

مخزن فراداده	Metadata Repository
مخزن مرکزی	Central Repository
مد	Mode
مدل آبشاری	Water Fall
مدل استدلال بر مبنای حافظه	Memory Based Reasoning
مدل برف‌دانه	Snow-Flake Schema
مدل پاسخ	Response Model
مدل ستاره‌ای	Star Schema
مدل شبکه ستاره‌ای	Star Net
مدل صورت فلکی	Fact Constellation
مدل کپی‌کننده	Copying Model
مدل مارپیچی	Spiral
مدیریت ارتباط با مشتری تحلیلی	Analytical CRM
مدیریت ارتباط با مشتری عملیاتی	Operational CRM
مدیریت ارتباط با مشتری مشارکتی	Collaborative CRM
مدیریت ارتباط با مشتری	Customer Relationship Management (CRM)
مرتب‌سازی نقاط برای شناسایی ساختار خوشه‌بندی	Ordering Points To Identify the Clustering Structure
مسئله تعمیم	Generalization Problem
مستقیماً قابل دسترس چگال	Directly Density Reachable
مشتاق	Eager
مشتری محوری	Customer Driven
مشتری مداری	Customer Centric
مشتریان احتمالی	Prospects Customers
مشتریان وفادار	Loyal Customer
مشخصه	Feature
مصالحه	Reconciliation
مصرف گذشته	Data Decay
مصورسازی	Visualization
مصورسازی داده‌ها	Data Visualization

مصیبت بُعد	The Curse Of Dimensionality
معنی‌دار	Significant
مغشوش	Noisy
مقادیر مفقوده	Missing Value
مقیاس بندی اعشاری	Normalization by decimal scaling
مقیاس‌بندی چندبعدی	Multidimensional Scaling: MDS
مکعب داده	Data Cube
ممیزی داده	Data Auditing
منفی درست	True Negative
منفی غلط	False Negative
موضوع محور	Subject Oriented
میان دامنه	Midrange
میان رشته ای	Interdisciplinary
میانگین	Mean
میانه	Median
میانگین امتیاز ثابت	Constant Average Degree Assumption

ن

ناسازگار	Inconsistent
ناقص	Incomplete
نحوی	Syntactical
نرمال‌سازی	Normalization
نرون	Neuron
نزدیک‌ترین همسایگی	K Nearest Neighborhood
نسبتی یا نسبی	Ratio
نقاط	Minpts
نقشه فقر	Poverty Map
نقشه‌های خودسازمان یا خودسازمانده	Self-Organizing Maps: SOM
نقشه‌های کوهونن	Kohonen Maps
نقشه‌های مشخصه خودسازمان	SOFM: Self-Organizing Feature Maps

نگاشت شماتیک	Schema Mapping
نمادین	Symbolic
نمودار پراکنش	Scatter Plot
نمودار چندک	Quantile Plot (Q-P)
نمودار دندانه‌ای	Dendogram
نمودار ستونی	Bar Chart
نمودار کلوچه‌ای	Pie Chart
نمودار لوئس	Loess Curve
نمونه برداری	Sampling
نمونه‌ها	Instances

و

وابستگی	Dependency
وابستگیهای احتمالی	Probabilistic Dependencies
ورودی نامناسب، خروجی نامناسب	Garbage in Garbage Out
وزن اتصال	Connection Weight
وزن‌دهی به شیب	Slope Weighting

ه

هرزنگاری	Spam
هرس	Prune
هرس اولیه	Prepruning
هرس ثانویه	Postpruning
هستان‌شناسی	Ontologies
همبستگیها	Correlation
هم‌پوشانی	Overlapping
همسایگان مشترک	Common Neighbors
همکاری	Cooperation
هموار	Smooth
هموارسازی	Smoothing

ی

یادگیر مبتنی بر نمونه	Instance Based Learner
یادگیری ماشین	Machine Learning
یک فایل آماده برای کار	Work File
یادگیری رابطه‌ای	Relational Learning
یالها	Edges
یک خوشه مرکزمحور	Center Defined Cluster
یکی‌سازی جمعی	Collective Consolidation