دانشگاه پیام نور

ده علوم تربیتی و روانشناسی؛دانش

گروه علم اطلاعات و دانش شناسی

جزوه ترسیم نقشه علم

مربوط به رشته علم سنجی

تألیف

دکتر فرامرز سهیلی، دکتر محمد توکلی زاده راوری،

دکتر افسانه حاضری و ندا دوست حسینی

مقدمه

پیشرفت علم در حوزه‌های گوناگون مرهون تلاش دانشمندان پیشین است. پژوهشگران در یک حوزه‌ی علمی به منظور دیدن فراسوهای دانش در حوزه‌ی تخصصی خود، معمولاً آثار دانشمندان پیش از خود را مرور می‌کنند؛ به عبارت دیگر، پژوهشگران با اتکاء به گذشته علم، آینده علمی حوزهٔ تخصصی خود را پیش می-برند. یکی از راه‌هایی که پژوهشگران را برای رسیدن به اهداف پژوهشی در حوزه تخصصی خود کمک می کند، داشتن درک و نمایی کلی از چارچوب علمی حوزه مورد نظر است. در این راستا دیداری‌سازی اطلاعات[1] یا ترسیم نقشه و ترسیم ساختار علمی آن حوزه، ضروری به نظر می‌رسد. عملکرد نقشهٔ علمی یک حوزهٔ علمی شبیه به نقشه‌های معمولی راه، شهر، مترو و مانند آن است که به کمک نمادهای گرافیکی قصد یاری رساندن به کاربران را دارند. در یک نقشه علمی که بر اساس بروندادهای علمی–پژوهشی دانشمندان یک حوزهٔ علمی ترسیم می‌شود، نویسندگان تأثیرگذار، خوشه‌های موضوعی شکل گرفته در طول زمان و آثار مهم و تأثیرگذار، تعیین و معرفی می‌شوند. از دیگر ویژگی‌های دیداری‌سازی اطلاعات که حاصل آن ترسیم نقشه-های علمی است، امکان مطالعهٔ تاریخ علم است. در نقشه‌های علمی به وضوح ظهور حیطه‌های جدید و توقف برخی حیطه‌های علمی اشباع شده قابل ملاحظه و مطالعه است. به بیانی ساده نقشه علمی به تصویر کشیدن نتایج برآمده از تجزیه و تحلیل انتشارات یک حوزه علمی از زوایای مختلف و ارائه نگرشی کلی از آن حوزه است. همان‌گونه که تاریخ هر شهر، کشور و منطقه را افراد، رویدادها و حوادث برجسته و غیره آن منطقه شکل می‌دهند، تاریخ علم هر حوزه علمی، کشور و یا دانشگاه را نیز نویسندگان و مقالات برجسته آنها تشکیل می‌دهند (سهیلی و عصاره، ۱۳۸۹).

در تعریف نقشه‌های علمی اعتقاد بر این است که نقشه علمی، نمایی از حوزه‌های علمی است که با تجزیه و تحلیل کمّی اطلاعات کتابشناختی تهیه می‌شود. عناصر تشکیل دهنده‌ی نقشه‌های علمی، بروندادهای حوزه‌های پژوهشی هستند. در این نقشه‌ها، حوزه‌های علمی که دارای ارتباط مفهومی قوی‌تری هستند، در کنار همدیگر و حوزه‌هایی که ارتباط ضعیف‌تری دارند در فاصله‌ی دورتری قرار می‌گیرند (نویونز[2]،۱۹۹۹، نقل در محمدی، ۱۳۸۷ الف).

1 Information Visualization (Iv)

2 Noyons

مطالعات تاریخی علم، به منظور مشخص نمودن مؤثرترین ایده‌ها، خاستگاه‌ها و بنیان‌های نظری در یک حوزه‌ی علمی، می‌تواند دید روشنی از گذشته‌ی یک علم را پیش رو قرار دهند. در مطالعات تاریخی علم، طی یک سیر زمانی، پارادایم‌ها و جریان‌هایی که در توسعه‌ی یک حوزه علمی خاص تأثیرگذار بوده‌اند، مشخص می‌شوند و مورد تحلیل قرار می‌گیرند. استنادها و ارجاعات دو پایه‌ی اساسی چنین بررسی‌هایی هستند. مک‌کین[1] (۱۹۹۴)، شیفرین[2] و برنر[3] (۲۰۰۴) بیان می‌کنند که حوزه‌ی ترسیم ساختار علم، حوزه‌ی مهمی برای پی بردن به رشد پژوهش علمی، دنبال کردن پویایی یک حوزه، کشف حوزه‌های جدید پژوهشی و ایجاد یک تصویر بزرگ از ساختار علمی و ایجاد رقابت است. سودمندی نقشه‌ی علمی یک حوزه برای متخصصان، در بررسی روندها و تصدیق پیش‌بینی‌ها و برای غیرمتخصصان نقطه‌ی ورودی به یک حوزه و پاسخ به پرسش‌های مخصوص آن حوزه است (بویاک[4]، ۲۰۰۴، نقل در رید[5] و چن[6]، ۲۰۰۷).

ترسیم ساختار علم از طریق گروه‌بندی دسته‌ها با استفاده از تحلیل دسته‌ها (خوشه‌ها) میسر است (کاکل[7]، ۱۹۷۶، نقل در عصاره، ۱۳۷۷). اسمال[8] (۱۹۹۹) نقشه علمی را یک نمایش فضایی از چگونگی ارتباط ارتباط رشته‌ها، حوزه‌ها، تخصص‌ها و مقاله‌های مؤلفان با یکدیگر تعریف می‌کند و معتقد است نقشه‌های علمی می‌توانند ادراک کاربران را از ارتباطات و گسترش‌های فکری تسهیل بخشند. یک نقشه علمی می‌تواند بینشی را در مورد وضعیت جاری دانش فراهم کند.

اسمال (۱۹۹۹) یکی از قدیمی‌ترین تلاش‌ها برای نمایش بصری گسترش‌های علمی را روش تاریخ نگاشتی[9] به کارگرفته شده توسط گارفیلد می‌داند که نمایش ترسیمی از الگوهای استنادی است که ارتباط بین مقاله‌ها در طول زمان را به منظور ردیابی اندیشه‌ها ترسیم می‌کند.

گارفیلد، پودوکین[10] و ایستومین[11] (۲۰۰۳) پیرامون بررسی تاریخ نگاشتی یک حوزه علمی بیان می‌کنند که "این روزها، صحبت از پارادایم‌ها رایج شده است، اما پارادایم در اصل مدل یک رشته است. تاریخ نگاشتی عمیقاً با درک تغییرات پارادایم‌ها در ارتباط است. اگر ما بخواهیم بدانیم چگونه یک پارادایم تغییر کرده است، شما باید آثار اولیه موضوع را مشخص کنید؛ بنابراین با مشاهده تغییرات در استنادهای آثار کلیدی رشته، شما می‌بینید که چگونه مفاهیم پایه یا درک از پارادایم‌ها تغییر یافته است". همچنین آن‌ها این فرض که

1 Maccain
2 Shiffrin
3 Borner
4 Boyak
5 Reid
6 Chen
7 Cawkell
8 Small
9 Historiography
10 Pudovkin
11 Istomin

اطلاعات کتابشناختی در برگیرنده‌ی مجموعه‌ای از مقاله‌های علمی منتشر شده است و برای هدف بازآفرینی ساختار تاریخ نگاشتی رشته کفایت می‌کند را به عنوان فرض پایه در نظر گرفتند. به علاوه، به دلیل این‌که نمایه‌های استنادی از آثار مورد استناد قرارگرفته‌ی هزاران پدیدآور شکل گرفته‌اند، فرض بر این است که روی هم رفته این نمایه‌ها بخش اعظم آثار اصلی در هر رشته‌ای را منعکس می‌کنند و آن آثار می‌توانند تصویر نسبتاً کاملی از پس زمینه تاریخی موضوع را ارائه می‌کنند.

این اندیشه که علم می‌تواند به صورت فضایی نمایش داده شود، دیر زمانی است که قابل ردیابی است. برای مثال بوش در سال ۱۹۴۵، «ترسیم سه بعدی رشته‌های علمی» را توصیف کرد. همچنین در دهه ۱۹۷۰ به اندیشهٔ ترسیم ساختار علم در علوم اجتماعی و جغرافیای انسانی و نیز در جامعه‌شناسی اشاره شده است. تعداد زیادی از دانشمندان، ساختار علم در حوزه‌های علمی مختلف را بر اساس متون علمی آن حوزه و با استفاده از تحلیل خوشه‌ای، تحلیل عاملی و مقیاس چندبعدی ترسیم کرده‌اند (گولد و وایت ۱۹۷۴؛ کرین، ۱۹۷۴ به نقل از عصاره، ۱۳۷۷).

ترسیم ساختار علم برای رشته‌های مختلف و پیگیری آخرین تغییرات آن‌ها، موضوع مورد توجه دانشمندان، کتابداران، فیلسوفان، دولتمردان و ناشران است و متون علمی، ماده‌های اصلی برای این ترسیم محسوب می‌شوند. در ترسیم ساختار علم سه جزء در نظر گرفته می‌شوند: عناصر فردی، عناصر مرتبط با یکدیگر که یک شبکه را به وجود آورده‌اند و تفسیر روابط بین عناصر (پائول[1]، ۲۰۰۱، نقل در پشوتنی زاده و عصاره، ۱۳۸۹).

دانشمندان مختلف با استفاده از ابزارها و روش‌های متفاوت اقدام به دیداری سازی و ترسیم ساختار علم در رشته‌های علمی مختلف کرده‌اند، با این حال نرم‌افزار «هیست سایت» در ترسیم نقشهٔ علمی از تقدم زمانی نیز استفاده کرده است (گارفیلد، پوداوکین و ایستومین، ۲۰۰۲)، با توجه به مزیت‌های برشمرده، برخی از فواید رسم نقشه‌های علمی و نیز تاریخ نگاشتی علمی به کمک این نرم‌افزار را می‌توان به شرح زیر دانست:

مشخص شدن جریان اطلاعات از گیرنده به فرستنده و بالعکس؛

نمایش توسعه و پیشرفت علم؛

مقایسهٔ رشد علم در سال‌های مختلف؛

نمایش تاریخ علم؛

مشخص شدن آثار مهم و اثرگذار در هر رشته؛

نشان دادن موضوعات جدید و زمان مطرح شدن و رشد آن‌ها.

1 Paul

ترسیم ساختار علم

پرایس[1] مطالعه علم، با استفاده از روش‌های علمی را پیشنهاد کرد، از آن زمان به بعد پژوهش در حوزه کتاب‌سنجی و علم‌سنجی فنونی را برای تحلیل مجموعه‌ای از داده‌ها، توسعه داده‌اند. بسیاری از آثار اولیه بر شناسایی شبکه‌ها یا خوشه‌های نویسندگان، مقاله‌ها یا ارجاعات، متمرکز بود. روش‌های جایگزینی بر اساس تحلیل هم واژگانی برای شناسایی موضوعات معنایی توسعه داده شد. پیشرفت در قابلیت‌های رایانه‌ها، تحلیل مجموعه‌های از اسناد با مقیاس بالا را تسهیل نمود، پیشرفت‌های اخیر در فنون دیداری‌سازی، توانایی به تصویر کشیدن حوزه‌های دانش[2] را اضافه نموده است.

به‌طور کلی یک نقشه علم شامل مجموعه‌ای از عناصر همراه با روابط بین این عناصر است. این عناصر می‌توانند حوزه‌ها یا رشته‌های علمی، مجله‌ها، مقاله‌ها، یا سایر واحدهایی که قسمتی از علم را نمایش می‌دهند، باشد. ویژگی‌هایی که یک نقشه را از یک نظام رده‌بندی ساده متمایز می‌کند عبارت‌اند از: ۱- دیداری‌سازی عناصر،۲- پیوند صریح جفت‌های از عناصر بر اساس روابط بین آنها. از چشم‌انداز ترسیم[3]، رده‌بندی اغلب به عنوان گامی همراه با راهی برای ایجاد یک نقشه دیداری در نظر گرفته می‌شود، اما اگر روابط بین رده‌ها صریحاً مشخص نشده باشد دیداری‌سازی با نقشه کردن مساوی نیست. نقشه علم به‌طور معمول به عنوان نمودارهای گره- لبه[4]، شبیه به آن‌هایی که در شبکه علم بکار برده می‌شود دیداری‌سازی شده است (کلاونز و بویاک[5]، ۲۰۰۹).

بر خلاف ترسیم علم[6]، ترسیم دانش[7]، بیش از حد بر پرسش هستی‌شناسی، یا دانش چیست و چگونه دانش ممکن است رده‌بندی شود، متکی است. به علاوه ترسیم دانش تعریف متفاوتی از کلمه "رسم کردن " دارد. در ترسیم دانش، مفهوم ترسیم با ارتباط بین نظام رده‌بندی و پدیده مورد بحث سر و کار دارد.

در ترسیم علم، مفهوم یک نقشه، از نقشه‌کشی گرفته می‌شود، زیرا نقشه‌ها با نقشه اشکوب (کف) یک کتابخانه قابل مقایسه هستند، جایی که کتاب‌ها در داخل اتاق‌هایی گذاشته می‌شوند (یعنی طرح‌های رده‌بندی) و اتاق‌ها طوری قرار دارند که پژوهشگران حداقل مسافت را طی می‌کنند (یعنی نواحی مرتبط نزدیک هم هستند). نقشه‌های دانش به سطوح حساس هستند- برای مثال بیماری عفونی و پزشکی به سادگی دو رده، به

1 Price
2 Knowledge Domains
3 Mapping
4 Node-Edge Diagram
5 Klavansa And Boyack
6 Science Mapping
7 Knowledge Mapping

عنوان یک موضوع عملی در کتابخانه هستند، شخصی ممکن است یک اتاق را به بیماری عفونی و اتاق دیگر به کتاب‌ها و مجله‌هایی درباره سایر زیر مجموعه‌های پزشکی اختصاص دهد (کلاونز و بویاک، ۲۰۰۹).

نقشه علمی نمایشی فضایی است، از اینکه چگونه رشته‌ها، حوزه‌ها، متخصصان و مقاله‌های انفرادی یا نویسندگان به همدیگر مرتبط هستند، همان‌طوری که توسط مجاورت فیزیکی و مکان‌های نسبی‌شان نشان داده می‌شوند و با راهی که نقشه‌های جغرافیایی، روابط جنبه‌های فیزیکی یا سیاسی زمین را نشان می‌دهند، قابل مقایسه‌اند. در مورد نوشته‌های علمی، نمایش فضایی می‌تواند درک ما را از روابط مفهومی و توسعه آنها تسهیل کند. نقشه علمی می‌تواند بینشی به درون وضعیت معاصر دانش فراهم کند(اسمال[1]، ۱۹۹۹).

حوزه‌ی فرعی از علم‌سنجی که بر ترسیم نقشه علم تمرکز دارد، به‌طور گسترده‌ای بر رویکرد تحلیلی هم استنادی که در سال ۱۹۷۳ توسط اسمال معرفی شد، متمرکز است (اسمال، ۱۹۷۳). دیداری‌سازی کل دانشِ علمی و ردگیری آخرین توسعه‌ها در علم و فناوری، نسل‌هایی از دانشمندان، فیلسوفان، کارکنان دولت، کتابداران و ناشران را کنجکاو نموده است. پیشرفت در دیداری‌سازی اطلاعات ابزارهای نوید بخشی را برای ارائه ساختار دانش و گسترش آنها در یک سبک شهودی رو به رشد عرضه نموده است (چن و پائول[2]،۲۰۰۱).

تولیدات علمی عناصر مناسبی برای دیداری‌سازی دانش هستند. پژوهشگران عموماً بر الگوهای ساختاری برجسته در کشف دانش، بازیابی اطلاعات و سایر روش‌هایی که ممکن است نگرش‌هایی را به درون ماهیت روابط مکنون و آشکاری همانند روابطی که بین نویسندگان انتشارات علمی، اسناد و مجله‌ها وجود دارد، تمرکز دارند.

روش‌ها و ابزارهای ترسیم ساختار علم

برای ترسیم نقشه‌های علمی از روش‌های مختلفی از جمله هم‌استنادی نویسندگان و هم‌رخدادی واژگان، هم‌استنادی مجله‌ها، کشورها و مانند آن، استفاده می‌شود؛ اما در ترسیم نقشه علمی یک حوزه به صورت تاریخ نگاشتی از خود استنادها استفاده می‌شود.

شیباتا[3] و همکاران (۲۰۰۹) بیان می‌کنند که استنادها در کشف زمینه‌های پژوهشی پیشین در مقایسه با هم‌استنادی بهتر عمل می‌کنند. همچنین آن‌ها بیان می‌کنند که مقالاتی که با استناد مستقیم به یکدیگر مرتبط شده‌اند تمایل خوشه شدن بیشتری دارند که بیان کننده این است که آن‌ها در مقایسه با خوشه‌های تشکیل شده توسط هم‌استنادی، شباهت بیشتری دارند.

1 Small
2 Chen And Paul
3 Shibata

گارفیلد و همکاران (۲۰۰۲) در شرح روش تاریخ نگاشتی خود بیان می‌کنند که برخلاف بسیاری از فنون ترسیم نقشه و دیداری‌سازی موجود، هیچ کدام از آن‌ها برای ایجاد تاریخ نگاشتی به کار برده نشده است. در واقع هیچ کدام از مؤلفان در حوزه نقشه‌های هم‌استنادی ارتباط معنی‌دار قابل توجهی بین نمایش تاریخی و نقش بالقوه آن در ارزیابی بروندادهای پژوهشی در نمایه استنادی علوم، مدلاین، چکیده نامه شیمی و غیره در نظر نگرفته‌اند. در واقع هدف اصلی در یک نقشه تاریخ نگاشتی نمایش گسترش تاریخی از موضوعات و حوزه‌ها، از مقاله نخستین به جلو و سال به سال است.

این نکته را هم باید در نظر داشت که تحلیل‌های کتاب‌سنجی و علم‌سنجی بستگی به موفقیت بازیابی اطلاعات مرتبط در مورد مقالات موضوع انتخاب شده دارد. در گذشته انجام این‌گونه تحلیل‌ها با روش‌های دستی انجام می‌پذیرفت و از دقت پایین‌تری نیز برخوردار بود. خوشبختانه امروزه منابع پیوسته‌ای برای تحلیل‌های کتاب‌سنجی وجود دارد. بزرگترین و معمول‌ترین منبعی که برای مطالعات کتاب‌سنجی و علم‌سنجی پذیرفته و استفاده می‌شود پایگاه "وب آو ساینس[1] " است که بخشی از پایگاه‌های "وب آو نالج[2] " یا "تامپسون رویترز[3] " است. "وب‌آوساینس " از سه پایگاه نمایه استنادی علوم، نمایه استنادی هنر و علوم انسانی و نمایه استنادی علوم اجتماعی تشکیل شده که بیشتر تحلیل‌های مربوط به علم‌سنجی و کتاب‌سنجی با استفاده از این پایگاه‌ها صورت می‌پذیرد. گارفیلد، پودوکین و ایستومین (۲۰۰۳) بیان می‌کنند که حتی قبل از ظهور نمایه استنادی علوم به صورت چاپی، استفاده از داده‌های استنادی برای کمک به نوشتن تاریخ علم در ۱۹۶۴ در گزارش "استفاده از داده‌های استنادی در نوشتن تاریخ علم" مورد بحث قرار گرفته است. این گزارش شامل ترسیم تاریخ‌نگاری از مقالات علمی در حوزه DNA است.

همچنین امروزه علاوه بر این پایگاه‌های پیوسته، نرم‌افزارهای بسیاری برای کمک و تسهیل تحلیل‌های کتاب‌سنجی و علم‌سنجی طراحی شده و در اختیار پژوهشگران این حوزه‌ها قرار گرفته است. از جمله این نرم‌افزارها می‌توان به "هیست سایت " اشاره کرد. آن‌گونه که گارفیلد و همکاران (۲۰۰۳) بیان می‌کنند ایده‌ی نوشتن برنامه‌ای کامپیوتری که بتواند چنین نقشه‌هایی را مستقیماً از فایل‌های الکترونیکی از نمایه استنادی علوم ایجاد کند، در طی پروژه ترسیم نقشه علم نگاشتی پژوهش‌های حوزه DNA شکل گرفت که منجر به تولید نرم‌افزار "هیست سایت" در سال ۲۰۰۲ شد. "هیست سایت"، نرم‌افزاری برای تحلیل و دیداری‌سازی ارتباط‌های استنادی مستقیم بین مقالات علمی است. ورودی‌های آن پیشینه‌های کتاب‌شناختی شامل منابع مورد استناد قرار گرفته از پایگاه "وب آو ساینس" یا دیگر پایگاه‌ها است. خروجی آن جدول‌های مختلف و

1 Web of Science

2 Web of Knowledge

3 Tamson Ruters

نقشه علم نگاشتی در مورد موضوع دانشی مورد مطالعه است (گارفیلد، پاریس[1] و استاک[2]، ۲۰۰۶). ترسیم ساختار هر شاخه‌ای از علم، مبتنی بر آثار تأثیرگذار آن حوزه است و با توجه به اهمیت آن امروزه نرم‌افزارهای گوناگونی همچون SPSS، Pathfinder، HistCite™، UCINET، Pajek و مانند آن برای انجام این امر در دسترس هستند. ولی هیچ یک از آنها همچون HistCite™ نمی‌توانند سیر تحول تاریخی یک حوزه علمی را نشان دهند، زیرا این نرم‌افزار قابلیت ترسیم نقشه آثار علمی را به ترتیب سال نشر آنان دارد. در این ترسیم دوبعدی دو گروه از آثار وجود دارند. آثاری که به لحاظ اهمیتشان، بسیار مورد استناد قرار گرفته‌اند (دایره‌های بزرگ) و آثاری که استناد داده‌اند.

نقشه علمی یا علم نگاشت

نقشه علمی، بازنمونی فضایی از چگونگی پیوند رشته‌ها، حوزه‌ها، متخصصان و مقاله‌های آنها به وجود می‌آورد، این نقشه‌ها را می‌توان به نقشه‌های جغرافیایی که رابطه‌های سیاسی یا جنبه‌های فیزیکی را بر روی زمین نشان می‌دهند، تشبیه کرد (اسمال، ۱۹۹۹). شبکه استنادی که در قالب این نقشه‌های فضایی نشان داده می‌شود، نموداری جهت‌دار و پیچیده است که رأس‌های آن می‌توانند به ترتیب زمانی مرتب شوند و خط‌های مرزی موجود در این نگاشت‌ها رأس‌های قدیمی را به رأس‌های جدید وصل می‌کنند. این شبکه الگوهای ارتباطی و همچنین چگونگی همکاری علمی و روند استنادهای ملی و جهانی پژوهشگران را نشان می‌دهند. نمایش فضایی، درک پژوهشگران را از رابطه‌های مفهومی و پیشرفته تسهیل می‌نماید. همچنین، نگاهی عمیق به وضعیت دانش معاصر در قلمروهای گوناگون علمی فراهم می‌کند.

ریپ[3] (۱۹۸۸) نقشه علم را به عنوان دیداری‌سازی مکان‌شناسی روابط بین عناصر یا جنبه‌های علم تعریف کرده است. نقشه‌کردن علم معمولاً به یکی از این سه دلیل دنبال می‌گردد: (۱) برای پشتیبانی از بازیابی اطلاعات، (۲) برای درک بهتر علم و توسعه‌ی آن و (۳) به دلایل سیاست‌گذاری‌های علم جهت مطلع کردن تصمیماتی درباره تخصیص منابع و پاداش‌ها.

حوزه‌ای فرعی از کتاب‌سنجی که در نقشه‌کردن علم به کار گرفته می‌شود، دیداری‌سازی متون[4] یا دیداری‌سازی اطلاعات نامیده می‌شود (میلوجویک[5]، ۲۰۰۹). در ساختار علوم بعضی از ارتباطات و پدیده‌ها به صورت انتزاعی برای ذهن قابل درک است؛ در صورتی که همین روابط به صورت فیزیکی برای چشم ملموس نیستند، پژوهشگران حوزه علم اطلاعات در تلاش‌اند که روابط و پدیده‌های نامرئی موجود در

1 Paris
2 Stock
3 Rip
4 Visualization Of Literature
5 Milojevic

ساختار علم را کشف نموده و با زبان گرافیکی به صورت چند بعدی در قالب نقشه‌های علم ترسیم نمایند (محمدی،۱۳۸۷).

اگر ساختار یک علم را به عنوان یک کشور تصور کنیم، حوزه‌های موضوعی تحت پوشش آن علم را می‌توان شهرهای آن کشور به حساب آورد. به راحتی می‌توان در ساختار یک حوزه از علم روابط مختلفی را کشف نمود. ازجمله این ارتباطات می‌توان به ارتباط بین پژوهشگران یک علم و تعامل بین حوزه‌های موضوعی فرعی یک علم اشاره کرد. حال ممکن است این نوع روابط به صورت انتزاعی برای ذهن قابل درک؛ ولی برای چشم نامرئی باشند. از این رو، کاربرد زبان گرافیک برای ترسیم پدیده‌های نامرئی در ساختار علوم موجب تشکیل یک علم جدید از ارتباطات بصری شده است که پس از تصاویر و نشانه‌ها از آن به عنوان زبان سوم نام می‌برند (مویا انگون[1]، وارگس کوواسادا[2] و دیگران، ۲۰۰۴ به نقل از محمدی،۱۳۸۷).

پل اتله[3] در سال ۱۹۸۵ متوجه گردید که رده‌بندی دهدهی دیویی می‌تواند به عنوان یک نقشه علمی از حوزه‌های گوناگون علم باشد. بدین جهت، اتله مطالعات خود را به روی رده‌بندی دهدهی دیویی برای تبدیل به رده‌بندی دهدهی جهانی آغاز کرد. در سال ۱۹۱۸ اتله اظهار داشت که رده‌بندی دهدهی جهانی می‌تواند به عنوان یک نقشه کلی از حوزه‌های مختلف علوم به حساب آید (اتله، ۱۹۸۸). جان برنال[4]، یکی از مشهورترین فیزیکدانان جهان، تاریخ‌نگار و جامعه‌شناس علم، در سال ۱۹۳۹ یکی از اولین نقشه‌های علم جهان را ترسیم کرد (برنال، ۱۹۳۹). اسمال و گارفیلد[5] اظهار می‌دارند که شاید برادفورد یکی از اولین کسانی باشد که به‌طور غیرمستقیم به ترسیم ساختار علم اشاره کرده است. دویل[6] در سال ۱۹۶۱ با تأکید بر نقش رایانه در ترسیم نقشه‌های علمی، چگونگی ساخت و ترسیم این نوع نقشه‌ها را برای ایجاد تصویر بزرگی حوزه‌های جامع علمی پیشنهاد داد (دویل، ۱۹۶۱). گارفیلد در سال ۱۹۶۳ در مقاله‌ای علاقه واضح و روشن خود را در مورد ساخت نقشه‌های تاریخی علم براساس استناد را نشان داد (گارفیلد، ۱۹۶۳ به نقل از محمدی ۱۳۸۷).

ترسیم تصویری بزرگ از دانش علمی به دلایل متعددی مورد علاقه بوده است. از جمله این‌که رویکردهای سنتی ماهیتاً نیروی قهریه[7] هستند که پژوهشگران را مجبور می‌کند تا از بین حجم انبوهی از متون، متون، پژوهش‌هایشان را به پیش ببرند. به‌طور واضحی این کار زمان‌بر، به سختی قابل تکرار و ذهن‌گرا است. این کار در عین عظیم بودنش، پیچیده است. انتخاب از میان مدارکی که اخیراً منتشر شده‌اند، جهت یافتن موردی که بعدها به عنوان مورد مهمی شناخته خواهد شد پرزحمت و وقت‌گیر است. رویکردهای سنتی

1 Moya-Anegon
2 Varges-Quesada
3.Paul Otlet
4 John Bernal
5 Small And Garfield
6 Doyle
7 Brute-Force

به‌طور فزآینده‌ای مشکل هستند تا با رشد اطلاعات هماهنگ شوند. وقتی که به سمت حوزه‌های چند رشته‌ای مطالعاتی پیش می‌رود، حفظ دید کلی از آنچه که در حال انجام است نسبتاً مشکل‌تر می‌شود (بارنر[1]، چن[2] و بویاک[3]، ۲۰۰۳).

از اولین مطالعات دیداری‌سازی حوزه بر مبنای داده‌های استنادی، ایجاد نقشه تاریخی پژوهش‌های DNA است که به‌طور دستی در ابتدای دهه ۱۹۶۰ انجام شد. پس از آن پرایس در اثر کلاسیک خود در مورد ترسیم شبکه‌های علمی داده‌های مشابهی را مطالعه کرد. در دیداری‌سازی حوزه، روابط بین پیشگامان پژوهش[4] از طریق "نمایش فضایی[5]" ارائه می‌شوند. چنین نمایشی سه‌بعدی به استفاده کننده‌ها فرصت می‌دهد دهد تا متون علمی را بر اساس الگوهای فضایی ترسم شده ناوبری کنند.

گارفیلد (۱۹۹۴) هم‌چنین مفهوم رسم نقشه طولی[6] را مطرح کرد. در نقشه‌های طولی، یکسری از نقشه-های تاریخی متوالی می‌توانند برای نمایان ساختن پیشرفت‌های علمی استفاده شوند. تحلیل‌گران و متخصصان حوزه می‌توانند از نقشه‌های طولی برای پیش‌بینی گرایش‌های نوظهور برای حوزه‌های موضوعی استفاده کنند. چون دیداری‌سازی حوزه نوعاً آثار کلیدی را در یک رشته مورد ارجاع قرار می‌دهد، ابزار مناسبی است تا افراد مبتدی را قادر سازد که با این حوزه از طریق شناسایی آسان مقالات و کتاب‌های برجسته و هم‌چنین اعضاء دانشگاه نامرئی یا متخصصان آشنا شوند (بارنر، چن و بویاک، ۲۰۰۳).

ایده رسم نقشه در کتاب‌سنجی[7] بیشتر توسط پژوهشگران هلندی بخصوص نویونز و وان ران[8] مطرح شده است. نویونز و وان ران فنون ریاضیاتی خاصی را برای نقشه کتاب‌سنجی ایجاد کردند. فرضیه اصلی این است که هر حوزه پژوهشی می‌تواند توسط سیاهه‌ای از مهم‌ترین کلمات کلیدی توصیف شود.

به‌طور کلی یک نقشه علمی از مجموعه‌ای از عناصر همراه با روابط میان عناصرش تشکیل گردیده است. این عناصر می‌توانند حوزه‌ها یا رشته‌های علمی، مجله‌ها، پروانه‌های ثبت اختراع، مقاله‌ها یا هر بخش دیگری که قسمتی از علم را ارائه می‌کند، باشد.

شکل‌های اساسی نقشه‌های علم

نقشه‌های علمی دارای اَشکال متعددی هستند که از آن جمله می‌توان به موارد زیر اشاره کرد:

1 Borner
2 Chen
3 Boyack
4 Research Fronts
5 Spatial Representation
6 Longitudinal Mapping
7 Bibliometric Mapping
8 Noyons And Van Raan

۱-شکل سلسله مراتبی[1] (توسط برخی از نویسندگان به عنوان یک مدل خطی طراحی شده است) که در آن اکثریت رشته‌های علمی در یک توالی خطی پیوند خورده‌اند. اگر چه سطح پایینی از شاخه گزینی در شکل سلسله مراتبی می‌تواند باشد، اکثریت رشته‌ها به وسیله یک ساختار خطی به هم متصل گشته‌اند.

۲-شکل مرکزی[2]، این نوع شبکه، شامل یک رشته در مرکز شبکه و شبکه‌ای از نوع پره‌ای[3] که در آن درجه‌ی بالایی از شاخه‌گزینی از گره مرکزی، وجود دارد، است. قابل توجه است که همه نقشه‌ها با شکل مرکزی رشته واحدی در مرکز دارند.

۳-شکل غیر مرکزی[4]، شکل سوم که نه سلسله مراتبی است و نه مرکزی، اما به‌طور معمول در ساختاری ساختاری مدور یا شبیه حلقه ایجاد می‌شود. تنها تفاوتش با فرم سلسله مراتبی در این است که دو انتهای سلسله مراتب به روشنی به هم متصل شده‌اند، بنابراین شبیه یک حلقه است (کلاونز و بویاک، ۲۰۰۹).

طراحی نقشه‌ها در قالب یکی از فرم‌های سلسله مراتبی، مرکزی و غیر مرکزی براساس ترکیب عناصر به وسیله نویسندگان اصلی و تفسیر ما از نقشه‌های حقیقی ارائه شده در مقالات ارجاعی هستند.

نقشه‌ها می‌توانند، نقشی حیاتی در ارتباطات داشته باشند، در فضایی که کشفیات قدیم و جدید همراه با هم واقع شده‌اند. اساساً کشفیات جدید اغلب در فضاهای سفید یا نواحی از نقشه که بدون فعالیت یا کم فعالیت هستند، واقع می‌شوند. نواحی سفید مکان‌هایی هستند که ارتباط پژوهشگران در آنجا ایجاد نشده است. کشفیات جدید اغلب در آن مکان‌ها رخ می‌دهند.

نقشه‌های غیر مرکزی می‌توانند گزارش واحدی درباره کشفیات جدید ارائه نمایند. در بیشتر موارد جایی که بیشتر رشته‌ها به دور دایره‌ای به هم متصل‌اند، کشفیات جدید می‌توانند پیرامون دایره مکان‌یابی شوند. کشفیات کنونی احتمالاً در مقایسه با گذشته بیشتر بین‌رشته‌ای هستند، این کشفیات جدید می‌توانند نزدیک‌تر به مرکز دایره واقع شوند. نقشه‌های غیر مرکزی می‌توانند ارزش بالایی برای کشفیات جدید و پژوهش‌های بین‌رشته‌ای ایجاد کنند.

به عبارتی دیگر، نقشه‌های مرکزی و سلسله مراتبی به‌طور ذاتی اشاره به این نکته دارند که وضعیت‌های[5] های[5] بیشتری در بعضی از حوزه‌های علوم نسبت به دیگر حوزه‌ها وجود دارد. نقشه سلسله مراتبی، بیشتر بر نقش رشته‌های برتر در سلسله مراتب تأکید دارد (معمولاً ریاضیات). یک نقشه مرکز محور، به بالاترین وضعیت‌ها (موارد برجسته) در نواحی مرکزی پژوهش ارجاع می‌دهد و نیز نشان می‌دهد که حوزه‌های علمی در انتهای شاخه‌ها در نقشه از موقعیت‌های پایین‌تری برخوردارند یا شاید حتی بتوان گفت بی روح و مرده‌اند.

1 Hierarchical Form
2 Centric Form
3 Spokes Type Of Network
4 Non-Centric Form
5 Status

فرآیند دیداری‌سازی حوزه‌های دانش

وایت و مک کین[1] (۱۹۹۷) پنج مدل از نوشته‌ها را تعریف کرده‌اند: ۱)کتابشناختی[2]، ۲) ویراستاری[3]، ۳) کتاب‌سنجی[4]، ۴) کاربر[5] و ۵ ترکیبی[6]. با استفاده از ابزارهای رایانه‌ای و فنونی که امروز وجود دارد، خطوطی بین این مدل‌های سنتی می‌تواند محو گردد[7] (بورنر چن و بویاک،۲۰۰۳). مدلی که امروزه توسط بسیاری از پژوهشگران مورد استفاده قرار می‌گیرد، ممکن است به عنوان فرا مدل کاربر[8] توصیف شود.

این یک مدل کاربر است که در آن کاهش نوشته‌ها بر اساس جستجوهای کاربر، پرس و جوها، پروفایل‌ها یا فیلترها، اغلب به‌طور سریعی با استفاده از دسترسی رایانه‌ای به منابع داده‌ای تولید و فرمول‌بندی می‌شوند.

برای فراهم کردن پاسخ به سؤالات خاص مانند نویسندگان، عنوان‌ها، اصطلاحات توصیفگرها، تاریخ‌ها و غیره، این مدل نقش مدل کتابشناختی یا فرا مدلی که شامل فراداده‌های نویسندگان، عنوان‌ها، اصطلاحات توصیفگرها، تاریخ‌ها و غیره می‌شود و از آنها برای تعریف روابط مرتبط از طریق ترسیم نقشه، استفاده می‌کند و ویژگی‌های داده‌ها را در نقشه‌های دیداری‌سازی پیشرفته نمایش می‌دهد ایفا می‌کند. این داده‌ها که اغلب داده‌های کتابشناختی را در بر دارند یا می‌توانند برای تولید داده‌های کتابشناختی مورد استفاده واقع شوند، عبارت‌اند از: شمارش استنادها، توزیع اصطلاحات، توزیع داده‌ها به تفکیک سال نشر، عامل تأثیر و غیره که به آسانی می‌توانند از طریق دیداری‌سازی اطلاعات نمایش داده شوند و تفسیر نقشه را تسهیل نمایند. از جمله ویژگی‌های کتاب‌سنجی می‌توان به کاربرد آستانه‌ها و رتبه‌بندی‌ها اشاره کرد که از آنها برای محدود کردن داده‌ها به مرتبط ترین‌ها و شناسایی مهم‌ترین‌ها توسط کاربر، مورد استفاده قرار می‌گیرد (بورنر چن و بویاک، ۲۰۰۳).

فرا مدل کاربر دقیقاً مرتبط با فرآیندی است که توسط آن نقشه‌های حوزه یا دیداری‌سازی تولید می‌گردد. گام‌های اساسی در این فرآیند عبارت‌اند از: ۱- استخراج داده‌ها ۲- تعریف بخش تحلیل ۳- انتخاب سنجه‌ها ۴- محاسبه شباهت بین بخش‌ها ۵- تخصیص مختصات برای هر بخش و ۶- استفاده از نتایج دیداری‌سازی برای تحلیل و تفسیر. گام‌های چهار و پنج در این فرآیند اغلب در یک عملکرد جمع می‌گردند که می‌تواند به عنوان آرایش[9] داده‌ها توصیف شوند.

1 White And Mccain
2 Bibliographic
3 Editorial
4 Bibliometric
5 User
6 Synthetic
7 Become Blurred
8 User Meta Model
9 Layout

اولین گام در فرآیند ترسیم هر نقشه‌ای استخراج داده‌های مناسب است. به‌طور کلی کیفیت ترسیم هر نقشه‌ای یا دیداری‌سازی اطلاعات ضرورتاً به کیفیت داده‌های در برگرفته بستگی دارد.

گام دوم، مربوط به انتخاب یک بخش تحلیل، مرتبط با پرسشی است که شخص تصمیم دارد به آن پاسخ بدهد.

متداول‌ترین بخش‌ها در ترسیم نقشه متون، مجله‌ها، اسناد، نویسندگان و اصطلاحات توصیفگر یا واژگان هستند. هر کدام از این موارد بخش‌های متفاوتی از یک حوزه را نشان می‌دهند و انواع متفاوتی از تحلیل را فراهم آورند. به عنوان مثال، یک نقشه از مجله‌ها می‌تواند برای به دست آوردن دیدی کلان از علم، نشان دادن جایگاه مناسب و روابط بین رشته‌های اصلی، مورد استفاده قرار گیرد (باسسکولارد و زیت[1]، ۱۹۹۹). نقشه مجله‌ها همچنین می‌تواند در مقیاس‌های کوچک‌تر برای نشان دادن تمایزهای ظریف بین یک رشته مورد استفاده قرار گیرد. اسناد، شامل مقاله‌ها، پروانه‌های ثبت اختراعات و مانند آن، رایج‌ترین بخش‌هایی هستند که برای رسم نقشه یا دیداری‌سازی یک حوزه دانش به کار می‌روند. این نقشه‌ها برای اهداف مختلفی از قبیل بازیابی اسناد، تحلیل حوزه، تصمیم‌گیری در سیاست‌گذاری اطلاعات، ارزیابی کارایی پژوهشی[2] و مدیریت علم و فناوری یا هوش رقابتی به کار روند (بورنر چن و بویاک، ۲۰۰۳).

نقشه‌های مبتنی بر نویسنده نیز نسبتاً رایج هستند و به دو شکل رخ می‌دهند. نقشه‌های هم استنادی نویسندگان که به‌طور معمول برای پی بردن به ساختار فکری یک حوزه به کار برده می‌شوند، در مقابل نقشه-های هم نویسندگی برای نشان دادن شبکه‌های اجتماعی از رشته یا گروه به کار برده می‌شوند که در بخش-های مربوطه توضیح داده خواهد شد.

نقشه‌های معنایی[3]، نیز وجود دارند که اغلب به عنوان تحلیل‌های هم واژگانی شناخته می‌شوند و برای درک ساختار شناختی یک حوزه به کار برده می‌شوند. واژگان‌ها از منابع متنی مختلفی شامل واژگان منفردی که از عنوان مقاله‌ها، اصطلاحات توصیفگر، یا توصیفگرهای اختصاص یافته توسط ناشر که توسط پایگاه اطلاعاتی کارگزار فراهم می‌گردد، استخراج می‌گردند.

دیداری‌سازی حوزه دانش[4] نوعی از دیداری‌سازی اطلاعات است که ساختار یک رشته علمی یا دانشگاهی را نمایش می‌دهد. دیداری‌سازی حوزه دانش نویسندگان، اصطلاحات یا مقاله‌ها و روابط بین آنها را برجسته می‌کند (بورنر، چن و بویاک، ۲۰۰۳). دیداری‌سازی حوزه دانش دو هدف اصلی دارد: تحلیل و ارتباطات. دیداری‌سازی حوزه دانش ابزاری برای پژوهشگران است تا تاریخ یا وضعیت جاری یک رشته را

1 Bassecoulard & Zitt
2 Assessing Research Performance
3 Semantic Maps
4 Knowledge Domain Visualization (Kdv)

تحلیل کنند. پژوهشگران می‌توانند از دیداری‌سازی حوزه دانش برای ارزیابی تکامل حوزه در طی زمان (وایت و مک کین[1]، ۱۹۹۸)، یا برای شناسایی حوادث مهم در حوزه از قبیل تغییرات پارادایمی[2] نقاط بحرانی فکری[3] (چن[4]، ۲۰۰۴) استفاده نمایند. با بررسی زمان ظاهر شدن مقوله‌ها در دیداری‌سازی و چگونگی شکل گرفتن گروه‌ها و ارتباطشان، پژوهشگران می‌توانند نواحی موضوعی داغ و کم اهمیت‌تر را شناسایی و شاید ارتباطی را بین آن‌ها کشف کنند که قبلاً شناخته نشده است (چن، ۲۰۰۴).

دیداری‌سازی حوزه دانش

دیداری‌سازی‌های حوزه دانش به صورت سنتی به صورت آفلاین با استفاده از ترکیبی از فنون مبتنی بر رایانه و دستی ایجاد می‌شد. به‌طور کلی فرآیند ایجاد دیداری‌سازی یک حوزه از دانش موارد زیر را در بر می‌گیرد. گام اول شناسایی مجموعه‌ای از مجله‌ها، مقاله‌ها یا نویسندگان و انتخاب یک دوره زمانی برای مطالعه است. پژوهشگران دست‌اندرکار دیداری‌سازی حوزه دانش، از پایگاه‌های اطلاعاتی نظیر وب آو ساینس یا دایالوگ برای به دست آوردن داده‌های استنادی یا هم استنادی جهت مقوله‌های برگزیده شده استفاده می‌کنند. فرض اصلی این رویکرد این است که استنادها شاخصی از اهمیت یا اعتبار هستند. مقالاتی که در متون خیلی مورد استاد قرار گرفته شده باشند، نسبت به آن‌هایی که خیلی کم مورد استناد واقع می‌شوند مهم‌تر و معتبرترند. با شمارش تعداد استنادهایی که هر مقوله در دورهٔ زمانی انتخاب شده دریافت می‌کند، پژوهشگران می‌توانند اهمیت هر مقوله را تشخیص دهند و بررسی کنند که اهمیت آن افزایش یا کاهش پیدا می‌کند.

دیداری‌سازی حوزه دانش عموماً اهمیت را با استفاده از نماد اندازه (به عنوان مثال مقوله‌های مهم با استفاده از نمادهای بزرگ نمایش می‌دهند) یا رنگ (به عنوان مثال مقوله‌های مهم با رنگ‌های تند و مقوله‌هایی کم اهمیت‌تر با رنگ روشن) نمایش می‌دهد. به‌طور ویژه به نظر می‌رسد که اندازه یک ارتباط مستقیم، اهمیت دارد، یعنی مقوله‌هایی که به‌طور متناوب در یک حوزه مورد استناد واقع می‌شوند به عنوان بزرگ درک می‌شوند (سینستودت[5] و چن، ۲۰۰۵). در دیداری‌سازی حوزه دانش، معمولاً از یک آستانه استنادی (به عنوان مثال ۵۰ نویسنده‌ای که بالاترین استناد را دریافت کرده‌اند) استفاده می‌شود که تعیین می‌کند که کدام مقوله‌ها در دیداری‌سازی ظاهر شوند (آلندورفر[6]، ۲۰۰۶).

فرضیه مهم دیگر در دیداری‌سازی حوزه دانش، این است که هم رخداده‌ها، شاخص‌های ارتباط هستند. رایج‌ترین نوع هم رخداده‌ای که در دیداری‌سازی حوزه دانش به کار می‌رود هم استنادی است. هم استنادی

1 White & Mccain
2 Paradigm Shifts
3 Intellectual Turning Points
4 Chen
5 Synnestvedt
6 Allendoerfer

زمانی اتفاق می‌افتد که دو مقاله توسط مقاله سومی مورد استناد واقع می‌شوند (عصاره، ۲۰۰۳). هم استنادی داده‌ها می‌تواند در سطح مقالات انفرادی مورد تحلیل واقع شود یا می‌تواند به بررسی نویسندگان، مجلات، مؤسسات یا دیگر اطلاعات، خلاصه شود. مقوله‌هایی که به‌طور مکرر هم استناد شده‌اند نسبت به آن‌هایی که به ندرت هم استناد شده‌اند مرتبط‌تر در نظر گرفته می‌شوند.

نوع دیگری از هم رخداده‌ها که در دیداری‌سازی حوزه دانش به کار می‌رود، هم رخداده‌های بین مقاله-ها، پروانه‌های ثبت اختراع و کلیدواژه‌ها یا اصطلاحات هستند. اگر یک مقاله و یک اصطلاح با هم در دیگر مقالات به‌طور زیادی به کار روند و مورد بحث قرار گرفته باشند، تصور می‌شود که مقاله و اصطلاح با هم مرتبط هستند. این نوع از هم رخداده دقیقاً از هم استنادی صحبت نمی‌کند، اما مفهوم آن اغلب مشابه هم استنادی است. هم استنادی و سایر داده‌های هم رخدادی می‌تواند با استفاده از مقیاس چند بعدی، تحلیل شبکه و تحلیل خوشه‌ای تحلیل شوند. این فنون نقشه‌ها، نمودارهای شبکه یا سایر دیداری‌سازی‌ها که روابط بین مقوله‌ها را نشان می‌دهد، ایجاد می‌کنند. دیداری‌سازی حوزه دانش به‌طور معمول مرتبط ترین را با استفاده از مجاورت[1] (به عنوان مثال مقوله‌های مرتبط نزدیک‌تر به هم نمایش داده می‌شوند)، رابط‌ها[2] (به عنوان مثال مقوله‌های مرتبط درون مرز خوشه قرار می‌گیرند[3]) نشان می‌دهد (آلندورفر، ۲۰۰۸).

نقشه‌های دیداری سازی ایستای حوزه دانش اغلب در مجله‌های کتابداری و علم اطلاعات منتشر می-گردند یا در مجلات یا مجموعه مقالات کنفرانس‌های این حوزه دیداری‌سازی می‌شوند (چن و دیگران، ۲۰۰۵، پیلکینتون و مردیت، ۲۰۰۹). نقشه‌ها به‌طور معمول برچسب گذاری و توسط پژوهشگر دیداری‌سازی و تفسیر می‌شوند. این عمل اغلب با همکاری متخصصان موضوعی در حوزه‌ی مربوطه، در قالب گروه‌بندی‌ها و خوشه‌ها شرح داده، یا تأیید می‌شوند. روش دیداری‌سازی ایستای حوزه دانش مانند هر روش عینی دیگر مزایا و معایبی دارد. با همه اهمیتی که این روش در ارائه اطلاعات مهم، معتبر و مرتبط از میان متون دارد لیکن در درون متون کتابداری و علم اطلاعات، مخالفت‌هایی[4] برای استفاده از فنون هم استنادی و هم رخدادی مخصوصاً برای سنجش اهمیت یا تأثیر وجود دارد. به عنوان مثال هر استناد یا هم استنادی نمی‌تواند شاخص معتبری برای اهمیت یا مرتبط بودنش باشد (گارفیلد، ۱۹۷۹). به عبارتی شمارش استنادهای مقاله‌ای که خود استنادی‌های زیادی دارد، استنادهای منفی (اینجا مثالی از یک مقاله‌ای نامعتبر)، یا استنادهایی از منابع محدود، ممکن است اهمیت واقعی آن مقوله را منعکس نکند. با این وجود، استناد و هم استنادی توسط تعداد

1 Proximity
2 Connectors
3 Related Items Are Contained Within A Cluster Boundary
4 Objections

زیادی از کتابسنج‌ها[1] تأیید شده است که روش مفیدی برای سنجش اهمیت و مرتبط بودن است (گارفیلد، ۱۹۷۹، ۲۰۰۵)؛ و این فنون پایه‌های بسیاری از مطالعات دیداری‌سازی حوزه دانش را شکل می‌دهند.

کاربرد نقشه‌های موضوعی

نقشه های موضوعی، عملاً مجموعه ای از مفاهیم و اصطلاحاتی هستند که روابط بین آن‌ها از طریق پارامترهای حاصل از فنون مرتبط با تحلیل واژگانی، مانند تحلیل شبکه‌های اجتماعی یا خوشه‌بندی نشان داده می‌شود. از پیش نیازهای ترسیم نقشه‌دانش علم، در اختیار داشتن اصطلاحنامه‌ی آن حوزه دانش است (ناصری جزه و همکاران، ۱۳۹۱).

اصطلاح نقشه، در ادبیات مربوط به پروانه‌های ثبت اختراع با عنوان‌های نگاشت و نقشه به کار رفته است. به کارگیری هر یک از این دو اصطلاح مبتنی بر نوع ساختار موجودیت‌های موجود در پروانه‌های ثبت اختراع است. یک پروانه ثبت اختراع، موجودیت‌های زیادی دارد که برای تجزیه و تحلیل مناسب است. برخی از این موجودیت‌ها، ساختاریافته و برخی غیرساختاریافته هستند. منظور از ساختاریافته این است که از لحاظ ظاهری در تمام پروانه‌های ثبت اختراع وضعیت مشابهی دارند، مانند شماره پروانه‌های ثبت اختراع ، شماره ثبت یا صاحب امتیاز. در برابر موجودیت‌های غیرساختاریافته، به‌صورت متنی با اندازه و محتوای گوناگون هستند مانند بخش‌های دعوی حقوقی، چکیده یا توصیف اختراع. نتایج دیداری‌سازی براساس تحلیل پروانه‌های ثبت اختراع، اگر بر اساس موجودیت‌های ساختاریافته باشد به آن‌ها نگاشت گفته می‌شود. در برابر، اگر از بخش‌های غیرساختاریافته متن باشد به آن نقشه پروانه ثبت اختراع می‌گویند. البته در کل ممکن است به هر دو آن‌ها نقشه گفته شود (تسنگ و همکاران، ۲۰۰۷).

ویژگی اصلی تحلیل هم‌واژگانی، دیدارسازی ساختار منطقی یک حوزه خاص از طریق ترسیم نقشه مفهومی و تولید نقشه‌های علمی و فنی است. ترسیم نقشه‌هایی که حاصل از واژه‌های کلیدی پروانه‌های ثبت اختراع به عنوان یک موجودیت غیرساختاری هستند، می‌توانند به عنوان یک روش مهم در کشف شبکه مفاهیم حوزه‌های مختلف فناوری مورد استفاده قرار بگیرند.

1 Bibliometricians

کاربرد شبکه‌های موضوعی تنها به علم‌سنجی و کتابسنجی محدود نمی‌شود. از کاربردهای آن می‌توان از جمله در زمینه فناوری زیستی (ریپ و کورشیال[1]، ۱۹۸۴)، مهندسی نرم‌افزار (کولتر؛ مونارچ و کوندا[2]، ۱۹۹۸)؛ رفتاردرمانی (ژانگ[3] و همکاران، ۲۰۱۲)؛ بازیابی اطلاعات (دینگ[4] و همکاران، ۲۰۰۱) و غیره نام برد.

کاربرد نقشه‌های موضوعی در علم سنجی متنوع است و دسته بندی کاربردهای این نقشه‌ها در این حوزه، به سادگی ممکن نیست. در زیر از طریق تحلیل و مطالعه متون مختلف به گوشه‌هایی از آن اشاره شده است:

۱- یافتن ارتباطات پنهان در یک حوزه از علم (هی، ۱۹۹۹): یکی از کارکردهای این نقشه‌ها، کشف همکاری‌های بین حوزه‌های پژوهشی در علم و نشان‌دادن پیوندهای علمی است که کشف آن‌ها، ممکن است ازروش‌های دیگر دشوار باشد (بردلت[5]، ۲۰۰۶؛ سوانسون و اسمالسر[6]، ۱۹۹۹).

۲- کشف تکامل تدریجی مفاهیم یک حوزه از علم یا فناوری (مان و بورنر[7]، ۲۰۰۴): تحلیل هم‌واژگانی، ابزار قدرتمندی برای نشان‌دادن ساختار و تکامل تدریجی شبکه‌های اجتماعی ـ شناختی[8] است (بردلت، ۲۰۰۶).

۳- آشکارسازی گرایش‌های یک حوزه خاص: این کار از طریق سنجش روابط اصطلاحات یک حوزه صورت می‌گیرد (وانگ و اینابا، ۲۰۰۶).

۴- کشف الگوهای ارتباطی بین موجودیت‌ها: هدف از تصویرسازی این است که به کاربر این امکان را بدهد که روابط میان عناصر را کشف کند. نقشه‌های علمی نمادهای یک حوزه علمی را به تصویر درمی‌آورند. عناصر این نقشه‌ها، هم‌نشینی بین موضوعات و یا مفاهیم است. در این نقشه‌ها عناصر مرتبط با یکدیگر در مجاورت هم و عناصر متفاوت دورتر از هم قرار می‌گیرند (آسف و روریسا، ۲۰۱۳). از این رو، کشف الگوهای ارتباطی ممکن می‌شود.

۵- درک ساختار شبکه‌های موضوعی (وانگ و اینابا، ۲۰۰۹؛ بردلت، ۲۰۰۶): روش تحلیل هم‌واژگانی این امکان را فراهم می‌کند که ساختار روابط درونی و بیرونی عامل‌های موضوعی، به صورت عینی و بدون جرح و تعدیل نمایش داده شود (وانگ و اینابا، ۲۰۰۹) و این باعث می‌شود که درک ساختار روابط موضوعی آن‌ها ممکن شود.

---

1. Rip & Courtial
2. Coulter., Monarch & Konda
3. Zhang
4. Ding
5. Bredillet
6. Swanson & Smallheiser
7. Mane & Börner
8. Socio – Cognitive

۶- شناسایی موضوعات برجسته، اصلی و مهم یک حوزه (هی، ۱۹۹۹؛ آسف و روریسا، ۲۰۱۳؛ کومار و جان[1]، ۲۰۱۲): یکی از اهداف تحلیل هم‌واژگانی از طریق فنونی مانند "تحلیل شبکه‌های اجتماعی وخوشه‌بندی، شناسایی گروه‌های برجسته است."(وانگ و همکاران، ۲۰۰۶). بسیاری از کارهای اولیه در زمینه کشف‌دانش بر تحلیل فراوانی و تحلیل هم‌واژگانی تمرکز داشته‌اند. در تحلیل فراوانی، کلمات کلیدی شمارش می‌شوند و آنهایی که فراوانی بالایی دارند، به عنوان زمینه‌های برجسته پژوهش محسوب می‌شوند. از طریق تحلیل هم‌واژگانی می‌توان موضوعاتی را مشخص کرد که در یک دوره زمانی برجسته بوده‌اند، سپس کم رنگ شده‌اند و یا همچنان برجسته باقی مانده‌اند (کومار و جان، ۲۰۱۲).

۷- اشاعه یک ایده در یک دوره زمانی (آسف و روریسا، ۲۰۱۳): تحلیل هم‌واژگانی یک روش کیفی و عینی می‌باشد. این روش، بر ماهیت کلماتی استوار هستند که حامل ایده، دانش یا مفاهیم علمی مهم هستند (وانگ و اینابا، ۲۰۰۶).

۸- کشف موضوعات مورد علاقه و مفاهیم غالب در آثار پژوهشگران (رایان و برنارد، ۲۰۰۳؛ کومار و جان، ۲۰۱۲).

به طور کلی می توان بیان کرد که ترسیم نقشه‌های حاصل از تحلیل هم‌واژگانی، یک رویکرد علمی برای کشف‌دانش است. نکته بسیار مهم این است که ترسیم نقشه‌های علمی حاصل از تحلیل هم‌واژگانی، به تنهایی هدف نیست بلکه پس از ترسیم نقشه‌های مورد نظر باید تحلیل و تفسیر آن‌ها را انجام داد که مرحله‌ای بسیار مهم است.

نقشه‌های کتابشناختی

نقشه‌های کتابشناختی نمایش بصری شبکه کتابشناختی هستند که در آن مجموعه‌ای از موجودیت‌های کتابشناختی و روابط میان آنها نمایش داده می‌شود و هدف آن ارائه نمایی از ساختار متون علمی در یک حوزه مورد نظر است (ون ایک و والتمن[2]،۲۰۱۰). نقشه‌های کتابشناختی می‌توانند حوزه‌های پژوهشی در یک رشته علمی را نشان دهند. همچنین با کمک این نقشه‌ها می‌توان میزان ارتباط حوزه‌های پژوهشی را نیز تعیین کرد. این نقشه‌ها اغلب زمانی که درک روابط میان مجموعه‌ای از داده‌ها مورد نظر است، بسیار کارآمد هستند. نقشه-های کتابشناختی ابزارهای مفیدی برای سیاست‌گذاری علم هستند. به این ترتیب که مطالعه موجودیت‌ها به

1. Kumar & Jan
2. Van Eck

تعیین شاخص‌ها کمک می‌کند و شاخص‌ها مبنای تصمیم‌گیری سیاست‌گذاران قرار می‌گیرند (نویون[1]،۲۰۰۱؛ فرانکلین و جانسون[2]، ۱۹۸۸؛ هیلی[3] و دیگران، ۱۹۸۶).

واحدهای تحلیل و اندازه‌گیری ارتباط در نقشه‌های کتابشناختی

واحد تحلیل موجودیتی است که در نقشه نشان داده می‌شود. به‌طور معمول مدارک، مؤلفان، مؤسسات یا واژه‌ها واحد تحلیل در نقشه‌های کتابشناختی هستند. اسمال اولین کسی است که به ترسیم خوشه‌های مدارک پرداخت (اسمال و گریفیث۴ ۱۹۷۴؛ اسمال و سوییني۵ ۱۹۸۵). قدم اساسی در رسم نقشه‌های حوزه‌ای را اسمال و مارشاکوا[6] برداشتند. هر کدام از آنها هم استنادی میان مدارک را متغیر مناسبی برای نقشه علمی قلمداد کردند. اسمال معتقد بود که از طریق دسته‌بندی مدارکی که بسیار با هم استناد شده‌اند می‌توان ساختار علم را مطالعه کرد (اسمال، ۱۹۷۳؛ مارشاکوا، ۱۹۷۳). حرکت اسمال و مارشاکوا شروع ساخت نقشه‌های علمی شد که اغلب از نمایه استنادی علوم به عنوان منبع اطلاعات و هم استنادی استفاده می‌کردند. پژوهشگران دانشگاه درکسل (مک کین۷، ۱۹۹۰ و ۱۹۹۱؛ وایت۸ و گریفیت، ۱۹۸۱) نیز پیشگام در رسم نقشه‌های مؤلفان و مجلات مجلات بودند. فرانسویان (کالون، کورتایل، ترنر و باین۹، ۱۹۸۳) از دیگر پیشگامان ترسیم نقشه‌های واژگان بودند که بعدها توسط پیترز و وان ران، (۱۹۹۳) نیز پیگیری شد. در دهه ۹۰ روش‌های جدیدی در بازیابی اطلاعات و فنون نو در تجزیه و تحلیل اطلاعات، ترسیم و حالت فضایی[10] اطلاعات بر اساس فنون برای ترسیم ساختار حوزه‌های علمی مورد مطالعه قرار گرفت (مویا-آناگان[11] و دیگران، ۲۰۰۴). نکته مهم در مورد روش‌شناسی همه پژوهشگران این بود که گروهی از مدارک که زمینه موضوعی مشترکی داشتند مشخص می‌شدند. این امر گواه و شاهدی بود بر آنکه علم شبکه‌ای است از حوزه‌های خاصی که با هم ارتباط درونی داشته و از طریق تحلیل‌های کمی تولیدات مکتوب می‌توان به آن ارتباطات درونی دست یافت. علم ماهیتی پویا دارد، جریان مداوم تولیدات علمی تغییرات مداومی نیز در ساختار علم ایجاد می‌کند. نقشه‌های کتابشناختی روند رشد و تغییرات حوزه‌های علمی را ترسیم می‌کنند. به‌طور معمول مدارک، مؤلفان، مؤسسات، یا واژه‌ها واحد تحلیل در نقشه‌های کتابشناختی هستند. برای ترسیم نقشه کتابشناختی باید به نوع روابط میان موجودیت‌ها دقت کرد. بنابراین باید مقیاسی وجود داشته باشد تا میزان روابط میان موجودیت‌ها را بسنجد. به بیان دیگر باید

1 Noyons
2 Franklin & Johnston
3 Healey
4 Small & Griffith
5 Sweeney
6 Marshakova
7 Maccain
8 White
9 Callon, Courtial, Turner, & Bauin
10 Spatial Positioning
11 Moya-Aneg.N

دید دو موجودیت تا چه اندازه با هم مرتبط هستند. درک این ارتباط نیازمند یک مقیاس است. ذکر این نکته ضروری است که مقیاس اندازه‌گیری ارتباط برای واحدهای تحلیل (موجودیت‌ها) ممکن است متفاوت باشد (ون ایک و والتمن[1]، ۲۰۰۷).

رابطه مجلات، مدارک و مؤلفان اغلب بر اساس ارتباط استنادی بررسی می‌شوند. استناد مستقیم، هم استنادی و زوج‌های کتابشناختی سه مقیاس برای سنجش ارتباط استنادی میان موجودیت‌ها هستند (اسمال، ۱۹۹۹). روش دیگر برای سنجش ارتباط میان داده‌های کتابشناختی، روابط هم نویسندگی است (وایت و گریفیت، ۱۹۸۱؛ وایت و مک کین، ۱۹۹۸). این مقیاس می‌تواند ارتباط میان مؤلفان، دانشگاه‌ها و کشورها را تعیین کند. این ارتباط بر اساس اعداد تألیفات مشترک میان آنها سنجیده می‌شود. روش معمول دیگر برای سنجش ارتباط میان مدارک و مؤلفان بررسی میزان ارتباط میان واژگان آنها است (کالون، ترنر، لاویل[2]، ۱۹۹۱). به این ترتیب می‌توان ارتباط دو موجودیت را براساس تعداد هم رخدادی واژگان آنها سنجید. انواع مختلفی از نقشه‌های کتابشناختی را می‌توان ترسیم کرد و ساختار یک حوزه علمی را با آن نشان داد. برخی از نقشه‌ها بر پایه روابط هم استنادی و برخی بر اساس هم رخدادی واژه‌ها شکل می‌گیرند. خوشه مدارک هم استناد نمایانگر دانش پایه، مفاهیم کلیدی، روش‌ها یا تجارب پژوهشگرانی است که آن خوشه را ساخته‌اند. در تحلیل هم واژگانی، رخدادهای واژه‌ها در عنوان، چکیده یا متن مدارک بررسی می‌شود.

هم رخدادی واژه‌ها میزان ارتباط شناختی میان یک مجموعه مدارک را نشان می‌دهد. مقیاس‌های مذکور نوع روابط میان موجودیت‌ها را آشکار می‌کند. این ارتباط را می‌توان به شکل فاصله یا گراف در یک نقشه دیداری‌سازی کرد. نقشه‌هایی که بر پایه فاصله هستند ارتباط را با دوری و نزدیکی موجودیت‌ها نشان می دهند. هر چه فاصله دو موجودیت کمتر باشد ارتباط آن‌ها قویتر است. در نقشه‌های بر پایه گراف خطوط میان موجودیت‌ها ارتباط آنها را تبیین می‌کند. این نقشه‌ها بر اساس فنون ترسیم گراف فراچرمن و رینولد[3] (۱۹۹۱) و کامادا کاوایی[4] (۱۹۸۹) شکل گرفت. در ادامه هر کدام از این سنجه‌ها و مقیاس‌ها به‌طور مفصل مورد بحث قرار خواهند گرفت.

زوج‌های کتابشناختی[5]

1 Waltman
2 Laville
3 Fruchterman & Reingold
4 Kamada & Kawai
5 Bibliographic Coupling

دو روش عمده در تحلیل استنادی وجود دارد که عبارت‌اند از زوج‌های کتابشناختی و هم استنادی. این دو روش در واقع کاملاً متفاوت عمل می‌کنند، اما هر دو روش می‌توانند در ترسیم محتوای موضوعی پژوهش های یک حوزه به کار گرفته شوند.

در تحلیل استنادی چنین فرضی وجود دارد که اگر دو مقاله دارای مراجع یکسانی باشند، نوعی رابطه محتوایی بین آن مقالات برقرار است. این رابطه اولین بار توسط کسلر[1] (۱۹۶۳) مطرح شد، اشتراک در مأخذ که در برخی از متون تحت عنوان زوج‌های کتابشناختی نیز از آن یاد شده است، یعنی دو یا چند اثر به آثار مشترکی استناد داده باشند و مأخذ مشترکی داشته باشند. به این نوع رابطه میان آثار علمی گوناگون، ارجاع مشترک یا مأخذ مشترک نیز گفته می‌شود.

زوج‌های کتابشناختی توسط کسلر با بررسی تعدادی مقاله در ابتدای دهه ۱۹۶۰ و طی دو گزارش مؤسسه فناوری ماساچوست معرفی شد. این روش ابتدا برای گروه‌بندی مقالات علمی و فنی به کار رفته تا تهیه اطلاعات علمی و بازیابی مدارک را تسهیل بخشد (ژارنوینگ، ۲۰۰۶). بر اساس گزارش مؤسسه فناوری ماساچوست تعریف زوج‌های کتابشناختی چنین آمده بود: وجود یک عنوان مرجع مشترک میان دو مقاله، واحد زوجیت دو مقاله محسوب می‌شود که آن را اشتراک در متن نیز می‌نامند. بر این اساس می‌توان زوج‌های کتابشناختی را در دو گروه دسته‌بندی کرد. زوج‌هایی که تنها یک واحد مأخذ مشترک دارند و زوج‌هایی که بیش از یک واحد مأخذ مشترک دارند؛ بنابراین هر چه تعداد واحدهای مشترک دو مقاله در مأخذ بیشتر باشد، این دو زوج از لحاظ محتوایی به هم نزدیک‌ترند. پژوهش کسلر نشان داد مقاله‌هایی که تعداد بیشتری ارجاع-های کتابشناختی مشترک دارند، به احتمال بسیار ارتباط موضوعی نزدیکتری نسبت به سایر مقاله‌هایی که اشتراک مأخذ ندارند، دارا هستند. وی این مقیاس را بر مبنای ارجاع‌های مشترک، زوج‌های کتابشناختی یا اشتراک مأخذ نامید.

هنری اسمال (۱۹۷۸) معتقد بود که مأخذ یک مقاله نمادهایی از مفاهیم آن مقاله هستند و از طریق ردگیری این مأخذ می‌توان رابطه مفهومی مقالات و حرکت مفاهیم را در درون یک حوزه یا میان حوزه‌های مختلف علمی دنبال کرد.

هم استنادی

تحلیل هم‌استنادی[2] اولین بار توسط اسمال و گریفیث[3] (۱۹۷۴) برای بررسی دو فرضیه مرکزی و اصلی به وجود آمد. فرضیه اول این بود که ساختار علم را متخصصان تشکیل می‌دهند که می‌توان با روش‌های

1 Kessler

2 Co-Citation

3.Small And Griffith

عینی[1] آن را تبیین کرد و دوم این‌که استنادها به‌طور خاص، میزان سرمایه‌های فکری مشترک بین دو مدرک را می‌سنجند که این کار خود راهی عملی برای توصیف ساختار علم است (گارفیلد، ۱۹۷۹). اولین گام در تحلیل استنادی، مشخص کردن نمونه‌ای از اسناد منبع است. این کار باید از حوزه مورد علاقه استخراج گردد. هر سند منبع، شامل استناد به آثار قبلی است که نویسنده به عنوان پیوندی به پژوهش خود دریافت کرده است. با شناسایی دو یا چند اثر قبلی به عنوان آثار مرتبط با پژوهش کنونی، نویسنده پیوندی بین آثار گذشته با استنادکردن به آنها با همدیگر، شناسایی می‌کند. تحلیل هم استنادی تأکیدش بر پیوندهای خلق شده بین آثار مستند واقع شده است؛ که حاصل آن با هم مورد استناد واقع شدن نویسندگان منابع است. زمانی که نمونه‌ای از اسناد منبع انتخاب شدند، هر اثری که به سند منبع استناد داده باشد، زوجی است تا زوج‌های هم استنادی را شکل دهد. فرآیند زوج‌شدگی، برای همه اسناد منبع تکرار می‌شود و سپس فراوانی رخداد هر زوج واحد، محاسبه می‌شود. هر چه رخداد یک زوج هم‌استنادی بیشتر باشد، پیوند بین زوج‌های مقالات قوی‌تر می‌شود. این فراوانی هم‌استنادی، قدرت هم‌استنادی نامیده می‌شود که نشان می‌دهد چندین بار ایده‌های موجود در دو اثر که قبلاً منتشر شده‌اند، در اسناد بعدی پیوند داده شده و به این ترتیب نشان دهنده میزان توافق و مشابهت یک جفت مدرک با هم استناد شده در یک حوزه علمی است (لوک[2] و پررا[3]، ۲۰۰۱).

هم استنادی یا اشتراک در مأخذ، زمانی اتفاق می‌افتد که دو یا چند مدرک مکرراً و با هم در فهرست منابع مدارکی ظاهر شوند. در چنین حالتی گفته می‌شود که آن مدارک با هم رابطه‌ای اعم از موضوعی، روش-شناسی و غیره دارند (اسمال، ۱۹۷۳). اسمال تعریفی نمادین از هم استنادی ارائه می‌دهد: اگر A گروهی از مقالات است که مدرک A را در مأخذ خود دارد و B گروهی از مقالات است که مدرک B را، آنگاه $(A \cap B)$ مجموعه است که هر دو دو مدرک A و B با هم در فهرست مأخذشان آمده است. تعداد موارد $A \cap B$ که $n(A \cap B)$ خوانده می‌شود بسامد هم استنادی را نشان می‌دهد.

اسمال بیان می‌کند که برعکس زوج‌های کتابشناختی که متون را به هم مرتبط می‌کند، هم استنادی، مأخذ یا مدارک استناد شده را مرتبط می‌کند و بنابراین مشابه اندازه‌گیری توصیفگر یا قرابت واژگانی است. حد اشتراک مدارک هم استناد را جمعیت مدارکی تعیین می‌کند که به دو مدرک هم استناد، استناد دهند؛ بنابراین اشتراک زوج هم استناد پیوندی است که نویسندگان جدید میان آنها برقرار کرده‌اند. بسته به ارتباط نویسندگان الگوهای هم استنادی می‌تواند در سیر زمانی تغییر یابد؛ همان‌طور که هم رخدادی کلمات به مرور زمان تغییر

1 Objective Means
2.Locke
3.Perera

می‌کند. خوشه مدارک هم استناد معرف دانش پایه، مفاهیم کلیدی، روش‌ها و تجارب پژوهشگران آن خوشه است (اسمال، ۱۹۷۷، ۱۹۷۸).

اسمال (۱۹۷۳) معتقد بود که می‌توان رابطه مفهومی مدارک و حرکت مفاهیم را در درون یک حوزه یا میان حوزه‌های مختلف علمی از طریق ردگیری مآخذ دنبال کرد. هرگاه بپذیریم که میان یک متن و مراجع آن ارتباط وجود دارد، آنگاه می‌توان حدس زد اگر دو مقاله دارای مراجع مشترک باشند باید میان آن دو مقاله و همچنین میان مراجع مورد نظر مشابهت موضوعی وجود داشته باشد. اسمال این مقالات را هم استناد نامید. در این روش، واحد اندازه‌گیری حد اشتراک، مقاله جدیدی است که دو مقاله پیشین را در مأخذ خود به کار گرفته باشد. تعداد مقالات جدیدی که این دو مأخذ را مورد استفاده قرار داده باشد، معیار نزدیکی آن دو تلقی می‌شود؛ بنابراین هم استنادی، بسامد تعداد دفعاتی است که دو مقاله همراه هم در مقاله بعدی مورد استناد قرار گیرند. این روش ربط میان شبکه استنادها ابزار مفیدی برای تعیین حدود و ثغور تخصص‌های مختلف است. اسمال (۱۹۹۹) همچنین تأکید می‌کند اگر امکان آن باشد متونی را که به‌طور غیرمستقیم به هم مرتبط می‌شوند در یک محیط دیداری به نمایش گذاشت، می‌توان چشم‌اندازهای جدیدی در ترکیب دانش ایجاد کرد.

اسمال (۱۹۷۳) تفاوت میان تحلیل هم استنادی مدارک و زوج‌های کتابشناختی را چنین بیان می‌کند: برای آنکه زوجی قویاً هم استناد باشند تعداد زیادی از مؤلفان باید آنها را هم زمان مورد استناد قرار دهند؛ بنابراین هم استنادی (مدرک) رابطه‌ای است که توسط مؤلفان استناد دهنده ایجاد می‌شود. برای اندازه‌گیری میزان هم استنادی دو مدرک، میزان ارتباطی را که جمعیتی از مؤلفان استناد دهنده ایجاد کرده‌اند، مورد اندازه‌گیری قرار می‌گیرد. از آنجا که رفتار مؤلفان استناد دهنده است که نتایج را تغییر می‌دهد، الگوهای هم استنادی نیز در طول زمان تغییر می‌کند. زوج‌های کتابشناختی یک ارتباط ثابت و دائمی است، زیرا وابستگی به مراجع دو مدرک دارد. وقتی دو مدرک منتشر می‌شوند زوجیت میان آن‌ها نیز از طریق مراجعشان تعیین شده است؛ در حالی که درجه هم استنادی میان مدارک در طول زمان قابل تغییر است. تفاوت قابل توجه میان این دو دسته زوج‌های (هم استنادی و کتابشناختی) وقتی است که دو مقاله به دفعات با هم مورد استناد قرار می‌گیرند. هر مقاله خود نیز لزوماً مورد استناد قرار می‌گیرد. این بدین معنا است که برخی جنبه‌های کیفیت، اهمیت یا دیده شدن باید برای مدارکی که برای تحلیل هم استنادی انتخاب شده‌اند در نظر گرفته شود. این در حالی است که برای تحلیل روابط زوج‌های کتابشناختی توجه به این جنبه ضروری نیست. زوج‌های کتابشناختی تحلیل گذشته‌نگر است و هم استنادی آینده‌نگر است.

قوت اصلی تحلیل هم استنادی عینیت و دسترس‌پذیری داده‌ها است. برای شناسایی ساختار یک رشته، همچنین می‌توان از بررسی پژوهشگران آن حوزه درباره روابط میان نواحی موضوعی به عنوان جایگزین

استفاده نمود؛ اما در یک رشته‌ی علمی با دامنه‌ی گسترده‌ای از متون، برای پژوهشگران خیلی مشکل است که چشم‌انداز "تصویری بزرگ" از ساختار آن رشته را شکل دهند[1] و وقتی که از آنها سؤال می‌شود تا این عمل را انجام دهند، آنها لزوماً ممکن است پاسخ هایشان را به سمتی سوق دهند که در آن حیطه مشغول‌اند – نه به خاطر سوگیری‌های عمدی، بلکه به خاطر آشنایی بیشتر با آن حوزه، همچنین به دست آوردن میانگین پاسخ‌ها به سؤالات نظرسنجی روی الگوی ارتباطات در یک حوزه علمی مشکل است.

تحلیل هم استنادی از داده‌هایی که بدون قصد[2] توسط پژوهشگران تولید شده‌اند استفاده می‌نماید و ابزار انتقال چنین داده‌هایی به درون نقشه‌ی یک رشته یا تخصص را فراهم می‌کند. مطالعاتی که درصد بودند تا تأیید نمایند که ساختار هم استنادی در مقایسه با ساختارهایی که توسط متخصصان در آن حوزه تولید می-شوند چگونه است، در اکثر موارد متوجه شدند که متخصصان، با روابطی که توسط فن تحلیل هم استنادی شناسایی شده، هم رأی هستند. وقتی دو نوشته با یکدیگر هم‌استنادی داشته باشند، از این جهت حائز اهمیت است که نشانگر نوعی رابطه موضوعی، روش‌شناسی و غیره بین آن دو مدرک است؛ به عبارت دیگر آنها اشتراکی در حوزه موضوعی، روش‌های مورد استفاده و اطلاعات مورد علاقه دارند که باعث شده است این دو در کنار هم در مدرک سوّمی ظاهر شوند.

تحلیل‌های هم استنادی یک روش کتاب‌سنجی است که دانشمندان علم اطلاعات برای نقشه کردن ساختار فکری یک حوزه پژوهشی به کار می‌برند و این عمل شامل شمارش اسناد یک حوزه انتخاب شده از اسناد هم استناد شده[3] یا اسناد زوج شده[4] است که به‌طور متناوب در سیاهه ارجاعات کتابشناختی اسناد مورد استناد ظاهر می‌شوند. مطالعات هم استنادی، شمارش هم استنادها را در شکل یک ماتریس و از لحاظ آماری آنها را برای به دست آوردن تصویری لحظه‌ای در نقطه‌ای ناهمسان در زمان گردآوری می‌کنند و این نشان دهنده آن است که ساختار دانش حقیقتاً دچار تغییر و تکمیل می‌شود (اسمال[5]، ۱۹۹۳).

در سال‌های اخیر پژوهشگران روابط میان جنبه‌های مختلف دانش مکتوب (به عنوان مثال، نویسندگان، انتشارات یا مؤسسه‌ها) را اغلب به کمک بازنمایی شبکه تصویری[6] تحلیل می‌کنند (ژائو و استورتمن[7]، ۲۰۰۸). ۲۰۰۸). تحلیل و دیداری سازی شبکه‌های دانش می‌تواند به‌طور مؤثر در کشف دانش جدید و در مدیریت و استفاده از منابع دانش موجود کمک کند (اسمال، ۱۹۹۹). این قابلیت امروزه به‌طور فوق‌العاده مورد توجه و مطالعه قرار گرفته است، زیرا دانش ثبت شده به‌طور فزآینده‌ای به صورت رقومی در دسترس است و این‌چنین

1 Formulate A ‘‘Big Picture’’ View Of The Structure
2 Unselfconsciously
3 Co-Cited Documents
4 Paired Documents
5 Small
6 Visual Netwrok Recentralization
7 Zhao & Stroman

شکل‌های رقومی، حجم و تنوع زیادی از داده‌ها را برای مطالعات شبکه دانش فراهم می‌کند، همچنین امروزه قدرت محاسباتی مناسب به آسانی در اختیار دانشمندان اجتماعی جهت تحلیل و دیداری ساختن شبکه عظیمی از اطلاعات موجود است (بورنر، مارو و گلدستون، ۲۰۰۴، بویاک، بورنر و کلاونس، ۲۰۰۷).

تحلیل هم استنادی مدرک[1]

تحلیل هم استنادی مدرک به بررسی شبکه‌ای از مدارک (مقاله، مجله، کتاب و مانند آن) پرداخته و هر زوج مدرکی را که با هم مورد استناد قرار گرفته تحلیل می‌کند (اسمال، ۱۹۸۰). اسمال معتقد است مدارک استناد شده نماد اندیشه‌های علمی، روش‌ها و تجارب هستند. از دیدگاه نماد مفهومی، بررسی شبکه استنادی بر تحلیل خوشه‌های مدارک استناد شده و روابط میان آنها تأکید دارد.

تحلیل هم استنادی نویسندگان

تحلیل هم استنادی نویسنده زیر شاخه‌ای از کتاب‌سنجی است. این روش معمول‌ترین روش برای بررسی و ردگیری ساختار فکری یک حوزه علمی است. در تحلیل هم استنادی نویسنده بسامد استناد هر نویسنده یا نویسنده دیگر که با هم در مدارک استناد شونده حضور یافته‌اند شمارش می‌شود (بیر، اسمارت و مک لاین، ۱۹۹۰). تأکید تحلیل هم استنادی نویسنده بر نویسندگان هم استنادی است که به وسیله پیوندهای هم استنادی به هم مرتبط شده‌اند. اغلب مطالعات هم استنادی نویسنده بر نویسنده اول در مأخذ تأکید داشته‌اند، زیرا اطلاعات همه نویسندگان همکار در دسترس نبوده است. تحلیل هم استنادی نویسنده فنی مناسب برای تشخیص نویسندگان برجسته در حوزه‌های مختلف پژوهشی است (مک کین، ۱۹۹۰). تحلیل هم استنادی یک حوزه علمی نماینده ساختار فکری آن حوزه است. تحلیل هم استنادی نویسنده و هم استنادی مدرک فرآیند مشابهی دارند و تنها تفاوت آنها در واحدهای تحلیل مورد بررسی است؛ اما انتظار می‌رود که تحلیل هم استنادی مدارک الگوهای خاص‌تری را نسبت به تحلیل هم استنادی نویسنده آشکار سازد، زیرا مأخذ استناد شده اطلاعات بیشتری نسبت به نویسندگان استناد شده در خود دارد. ماهیت استفاده از مأخذ استناد شده در شبکه هم استنادی مدارک ابهام کمتری نسبت به شبکه هم استنادی نویسنده دارد (چن، ۲۰۱۰).

یکی از روش‌هایی که برای دیداری سازی اطلاعات و کشف روابط میان جنبه‌های مختلف حوزه‌های مختلف دانش مکتوب به کار می‌رود تحلیل هم استنادی نویسنده[2] است. تحلیل هم استنادی نویسنده مجموعه‌ای از فنون جمع‌آوری، نمایش تحلیلی و گرافیکی داده‌ها است، که می‌تواند برای تولید نقشه تجربی

1 Document Co-Citation Analyses
2 Author Co-Citation Analyses (Aca)

نویسندگان برجسته در حوزه‌های مختلف دانش به کار رود. درون یک نقشه معین، مجاورتِ نقاط نمایانگرِ نویسنده، انعکاس‌دهنده شباهت محسوس در برخی از ابعاد است. با آزمایش کردن توزیع نویسندگان و خوشه‌های نویسنده درون فضای ذهنی[1] دو یا سه بعدی از یک نمایش نقشه شده، سایر جنبه‌های ساختاری می‌توانند توصیف شوند. خوشه‌های نقاط می‌توانند با حوزه‌های موضوعی، متخصصین موضوعی، مکاتب فکری، سبک‌های فکری مشترک یا گره‌های زمانی یا جغرافیایی شناسایی شوند (مک کین[2]، ۱۹۹۰).

تحلیل هم استنادهای نویسنده شکل ویژه‌ای از تحلیل استنادی است که مبتنی بر شمارش زوج‌های با هم استنادیِ بالای آثار برجسته[3] است، یعنی مجموعه از نوشته‌های یک نویسنده یا اولین نویسنده در همکاری‌های گروهی است (وایت و گریفیث[4]، ۱۹۸۲). در تحلیل هم استنادی نویسنده، از نویسنده به عنوان بخشی از تحلیل استفاده می‌شود و از روش‌های تحلیل چند متغیری همانند تحلیل عاملی، تحلیل خوشه‌ای و مقیاس چند بعدی به عنوان روش‌های آماری برای آشکار ساختن ساختار اصلی روابط میان نویسندگان استفاده می‌شود و نویسندگان پر استناد و نویسندگان هم استناد بخش تحلیل را تشکیل می‌دهند (سهیلی، عصاره و خادمی، ۱۳۹۱).

مطالعات هم استنادی نویسندگان با سیاهه متمرکز شده‌ای از نویسندگان که احتمالاً، در برخی از حوزه-های موضوعی پر استناد واقع شده‌اند، آغاز شد (عصاره و مک کین[5]، ۲۰۰۸). در پژوهش‌های مرتبط با تحلیل هم استنادی نویسندگان، شمارش استنادها را در ماتریسی با اسامی نویسندگان که به‌طور یکسان در ردیف‌ها و ستون‌هایی منظم شده‌اند، جمع می‌کنیم. ستون‌ها در ردیف‌هایی که قسمت بالائی و قسمت پائینی نیمه ماتریس را پر می‌کنند، جای می‌گیرند. تعیین اینکه در سلول‌های مورب بین نیمه ماتریس پایینی و ماتریس بالائی چه گذاشته شود، مشکل است. باید در این قسمت، کل استنادهای خود نویسنده قرار داده شود. اگر این چنین باشد، باید آن را به عنوان تعداد آثاری که به دو اثر توسط نویسنده استناد شده است، مد نظر قرار داد، یعنی هم استنادی یک نویسنده با خودش، آنچه مشکل بیشتری ایجاد می‌کند، آن است که برای یک نویسنده برجسته، عدد موجود در سلول وی ممکن است خیلی بالاتر از سایر عددهای سلول‌های موجود در ماتریس باشد (مک کین، ۱۹۹۰). برای جلوگیری از این مشکل، نویسندگان مختلف روش‌های متفاوتی را به کار برده‌اند. به عنوان مثال وایت و گریفیث،۱۹۸۲ ارزش‌های سلول‌های مورب را با جایگزین کردن یک ارزش بر اساس شمارش بالاترین هم استنادی سایر سلول‌ها برای هر نویسنده کاهش دادند.

1 Intellectual Space
2 Mccain
3 Highly Co-Cited Pairs Of Oeuvres
4 White & Griffith
5 Osareh And Mccine

همچنین مک کین در سال ۱۹۹۰، ارزش این سلول‌های مورب را به عنوان داده‌های گمشده[1] مد نظر قرار داد. سهیلی و عصاره (۱۳۸۶) عدد صفر را برای سلول‌های مورب در نظر گرفتند. عصاره و مک کین (۲۰۰۸)، سهیلی، عصاره و خادمی (۱۳۹۱) میانگین هم استنادی نویسندگان هر ستون را تقسیم بر تعداد ردیف‌های آن ستون کرده و میانگین آنها را به عنوان هم استنادی در سلول‌های مورب هر ستون قرار دادند.

بعد از اینکه ماتریس مربوط به هم استنادی نویسندگان ایجاد گردید. اولین گام در نقشه کردن یا خوشه کردن نویسندگان هم استناد شده، تبدیل داده‌های خام ماتریس به ماتریسی از ارزش‌های تقریبی است، که شباهت یا عدم شباهت زوج‌های نویسنده را نشان می‌دهد. در اکثر مطالعات هم استنادی نویسندگان، ضریب همبستگی پیرسون برای سنجش شباهت‌ها به کار برده شده است (مک کین، ۱۹۹۰).

ایجاد ماتریس همبستگی حداقل دو مزیت عمده دارد: اول اینکه برای هر زوج نویسنده، ضریب همبستگی به عنوان یک سنجه عمل می‌کند، نه تنها برای اینکه فقط چند بار این دو نویسنده با هم مورد استناد واقع شده‌اند، بلکه برای میزان شباهتی که پروفایل هم استنادیشان با هم دارد. دو نویسنده‌ای که اغلب اوقات به میزان بالایی توسط نویسندگان سومی مورد استناد واقع می‌شوند و به ندرت با دیگر نویسندگان، همبستگی مثبت بالایی خواهند داشت و می‌توان گفت که در یک معنا توسط جامعه استنادکنندگان درک شده‌اند که در برخی از مفاهیم به هم شبیه یا مرتبط هستند. همان‌طوری که قبلاً اشاره گردید برای تحلیل داده‌های هم استنادی نویسنده از تحلیل عاملی[2] استفاده می‌گردد. تحلیل عاملی تعداد زیادی از متغیرهای مشاهده شده را به تعداد کمی کاهش می‌دهد؛ به عبارت دیگر، کاربرد اصلی تکنیک‌های تحلیل عاملی شامل:

۱- کاهش تعداد متغیرها؛

۲- کشف ساختار در ارتباط بین متغیرها، یعنی طبقه‌بندی متغیرها (عصاره، ۲۰۰۳).

نویسندگانی که در ایجاد ساختار فکری یک زیر حوزه مشارکت دارند، تمایل به بارگیری شدن[3] در عامل‌های یکسان را دارند (مک کین، ۱۹۹۰). تحلیل عاملی به کمک نرم‌افزار آماری SPSS صورت می‌گیرد. به منظور استخراج عامل‌های از تحلیل مؤلفه‌های اصلی و دوران[4] استفاده می‌شود. معمولاً بارگیری‌های بالاتر از |.4| در نظر گرفته می‌شوند.

همچنین در مطالعات مرتبط با هم استنادی می‌توان به زوج‌های هم‌استنادی اشاره نمود. زوج‌های هم-استنادی عبارت است از مطالعه مدارکی که با هم مستند واقع می‌شوند؛ دسته‌بندی مواد هم استناد در ترسیم شبکه مجلات به کار می‌رود؛ همچنین، شناسایی نویسندگان کلیدی یک رشته نیز از طریق بررسی زوج‌های

1 Missing Data
2 Factor Analysis
3 Loading
4 Rotation

هم‌استنادی میسر است. از انواع تحلیل زوج‌های هم‌استنادی می‌توان مؤلفان هم‌استنادی[1]، مجلات هم‌استنادی[2] و کشورهای هم‌استنادی[3] را برشمرد.

هم نویسندگی

توسعه و گسترش علم فرآیندی اجتماعی است که از طریق شبکه‌ای از پژوهشگران که مجامع را تشکیل می‌دهند شکل می‌گیرد. پژوهشگران درون یک جامعه‌ی علمیِ خاص برای مشارکت در پایگاه دانش کلی آن جامعه با همدیگر تعامل برقرار کرده و همکاری می‌کنند. در واقع شتاب پیشرفت‌های علمی در چند دهه اخیر عمدتاً منسوب به تأسیس مجامع علمی و بهبود ارتباطات میان پژوهشگران معاصر است (چانگ و هارینگتون[4]، ۲۰۰۵). رشد دانش از طریق همکاری‌های رسمی بین مجامع دانشگاهی و ارتباطات غیررسمی با استفاده از شبکه‌های اجتماعی درون این انجمن‌ها تسهیل گردیده است. کارآمدی چنین جوامعی بستگی به قدرت و وسعت روابط بین اعضای آن دارد؛ بنابراین تحلیل این جوامع فرصتی را برای بررسی ساختار روابط بین یک جامعه علمی فراهم می‌کند (گالیسون[5]، ۱۹۹۷، به نقل از راچرلا و هو، ۲۰۱۰).

در پژوهش‌های دانشگاهی، نادر است که پژوهشگری بتواند بروندادهایی را بدون ارتباط با متون جوامع پژوهشی، تولید نماید. یافته‌های جدید معمولاً از متون جوامع پژوهشی، یعنی از جمع‌آوری پژوهش‌های پیشین یا روابط مشترک در حوزه‌ی پژوهشی استنتاج می‌شوند؛ بنابراین لازم است که فعالیت پژوهشگران در برخی حوزه‌ها به منظور درک ویژگی‌های آن حوزه در تولید دانش تحلیل شود، اما تحلیل و ارزیابی فعالیت-های هر پژوهشگر به تنهایی کافی نیست، بلکه باید جایگاه وی را در ساختار معنوی[6] (فکری) آن حوزه نیز در در نظر گرفت (یوشیکانه، نوزاوا و تسوجی، ۲۰۰۶). یکی از راه‌های رسم ساختار فکری یک حوزه علمی استفاده از تحلیل هم نویسندگی[7] است. در ادامه این فصل ابتدا به بررسی همکاری علمی[8] و در بخش‌های بعدی تحلیل هم نویسندگی مورد بررسی قرار خواهد گرفت.

همکاری علمی

1 Author Co-Citation
2 Journal By Journal Co-Citation
3 Country By Country Co-Citation

4 Chang & Harrington
5 Galison
6 Intellectual Structure
7 Co-Authorship Analyses
8 Scientific Collaboration

در طول تاریخ، پدیدهٔ همکاری در میان نوع بشر، همواره موضوعی مهم در جامعه‌شناسی بوده است. این نوع همکاری‌ها که در قالب‌های مختلف شامل تعاملات علمی، اقتصادی، حمایتی، سیاسی و مانند آن وجود داشته‌اند، به‌صورت شبکه‌های همکاری نمود پیدا کرده‌اند. همکاری علمی، یکی از نمودهای عینی همکاری در میان نویسندگان است و پدیدهٔ پیچیده‌ای است که از اشتراک توانمندی‌های آنان، حاصل و تولید دانش علمی جدید را بهبود می‌بخشد (سهیلی و عصاره، ۱۳۹۱).

تولید علمی گروهی به معنای همکاری افراد مناسب، در زمان مناسب، برای انجام کار علمی مناسب است که موجب افزایش نوآوری، خلاقیت، پیشرفت و ترقی گروه می‌شود. این امر، مستلزم ارتباط مداوم، متقابل و نزدیک بین اعضای گروه است (نوروزی و ولایتی، ۱۳۸۸، ص ۱۶). منشأ و بنیان همکاری علمی را باید با مطالعهٔ تاریخ گذشته آغاز کرد. همکاری علمی، ویژگی مهم ساختار پژوهش علمی معاصر است، البته کار گروهی علمی از قرن بیستم نمود پیدا نکرده و در واقع همکاری علمی مفهومی تازه نیست، بلکه واکنشی در برابر پدیدهٔ «حرفه‌ای شدن» علم است. حرفه‌ای شدن، فرآیندی است که گروهی از پژوهشگران را با مجموعه‌ای از گرایش‌ها – گرایش‌هایی که هم فراگیر و هم منحصر به فرد هستند – سازمان می‌دهد، به این معنی که حرفه‌ای شدن، قواعد، حقوق و راه و رسم دسترسی به یک گروه را در بر می‌گیرد؛ چه به این صورت که اعضای یک گروه را گرد هم آورد و یا اینکه آنها را از سایر افراد در جامعهٔ بزرگ‌تری جدا سازد (رحیمی و فتاحی، ۱۳۸۶).

همکاری علمی که یکی از نمودهای بارز آن هم نویسندگی است، از سوی یک شیمی‌دان فرانسوی طی سال‌های ۱۸۳۰–۱۸۰۰ مطرح و رواج پیدا کرد (عصاره، ۲۰۰۶). کاربرد این پدیده در متون تا جنگ جهانی اول رشد آرامی داشت و بعد از آن رشد سریع‌تری پیدا کرد. مطالعات نشان می‌دهد که در سال‌های اخیر همکاری علمی و به ویژه هم نویسندگی در میان نویسندگان و پژوهشگران رشد افزایشی داشته است. شاید بتوان علت این رشد فزاینده را به مزایایی که همکاری‌های علمی برای نویسندگان و آثارشان دارند، نسبت داد، که برخی از این مزایا شامل تبادل ثمربخش بودن ایده‌ها، کیفیت و اعتبار بالای آثاری که حاصل همکاری علمی می‌باشند، دریافت استنادهای بیشتر و به ویژه فوایدی که این همکاری‌ها برای کشورهای در حال رشد پدید می‌آورند، می‌شود (عصاره، ۱۳۸۸). همکاری علمی را می‌توان بازتاب فعالیت‌ها و رویکردهای جامعه علمی به شمار آورد. مطالعه و بررسی این مقوله می‌تواند به جامعه‌شناسی علم نیز کمک کند (رحیمی و فتاحی، ۱۳۸۶).

همکاری علمی به پژوهشگران فرصت می‌دهد تا قابلیت‌ها و توانایی‌های رشته‌های مختلف علمی را با هم ترکیب کنند، امری که انجام آن به صورت انفرادی امکان‌پذیر نیست. همکاری علمی با استفاده از تحلیل

الگوهای هم نویسندگی در مقالات منتشر شده و نیز با استفاده از تحلیل شبکه‌های اجتماعی قابل بررسی است. هدف پژوهشگران در همکاری علمی بررسی، تفسیر و بازبینی دانش جهانی است. گروهی از دانشمندان که با هم کار می‌کنند، شبکه‌ای اجتماعی را شکل می‌دهند. همکاری علمی ممکن است برای تولید پیشنهاد مالی، هم نویسندگی در خلق یک اثر علمی و یا اشتراک اندیشه‌ها از طریق بحث‌های غیررسمی صورت گیرد. از همکاری علمی اغلب برای اجرای پژوهش‌های بین‌رشته‌ای استفاده می‌شود. هدف پژوهش‌های بین‌رشته‌ای این است که نقاط قوت رشته‌های متعددی را برای ایجاد رشته‌ای جدید با هم ترکیب کنند (سهیلی و عصاره، ۱۳۹۱). پژوهش‌های بین‌رشته‌ای، فاصله‌های موجود میان واژگان، رویکرد و روش‌شناسی را پر می‌کند و رویکرد جدیدی را برای اندیشیدن ایجاد می‌کند. به‌علاوه همکاری علمی به پژوهشگران فرصت می‌دهد تا به منظور غلبه بر هزینه‌های گزاف تجهیزات و آموزش متخصصان، منابع خود را با هم به اشتراک بگذارند (بلانکا[1]، ۲۰۰۹).

امروزه پژوهشگران زیادی شکل‌ها و نقش‌های مختلف همکاری علمی را در حوزه‌های مختلف علمی بیان کرده‌اند. بررسی‌های این پژوهشگران را می‌توان با تحلیل در سه سطح زیر خلاصه کرد. در سطح خرد[2] (افراد)، سطح میانی[3] (مؤسسات) یا سطح کلان[4] (کشورها) (گلنزل[5]، ۲۰۰۲، کرشمر[6]، ۲۰۰۴). در واقع می‌توان توان به بررسی همکاری علمی بین پژوهشگران به صورت انفرادی یا همکاری علمی که بین مؤسسه‌های مختلف صورت می‌گیرد و بخش اعظمی از این نوع همکاری شامل ارتباط با صنعت در دانشگاه‌ها و مؤسسات پژوهشی است، پرداخت. یا به بررسی همکاری بین کشورهای مختلف که معمولاً در کشورهای صنعتی بیشتر مصداق دارد پرداخت و الگوهایی را که در سطوح مختلف همکاری وجود دارند، استخراج نمود. این الگوها که اغلب از طریق نگاشت‌هایی که حاصل برونداد نرم‌افزارهای مورد استفاده هستند، نمایش داده می‌شوند.

الگوهای همکاری که از طریق نگاشت نمایش داده می‌شوند می‌توانند از منظر کلان (شبکه محور) یا خرد (عامل محور) بررسی شوند. ساختار کلان یک نگاشت فرد را از عملکرد احتمالی ساختارهای اجتماعی برخاسته از فیزیک ارتباط‌های او آگاه می‌کند؛ نقش‌آفرینانی که در این شبکه جاسازی شده‌اند، ممکن است به طور کامل از این ساختار بی‌اطلاع باشند؛ مانند، شبکه‌هایی که اکثر نقش‌آفرینان آنها دارای ارتباط‌هایی با فاصله‌ی کوتاه به دیگر نقش‌آفرینان هستند، توانایی اشاعه اطلاعات بیشتری را دارند. البته لازم به ذکر است که

1 Bellanca
2 Micro Level
3 Meso
4 Macro Level
5 Glänzel
6 Kretschmer

احتمال انتقال اطلاعات توسط نقش‌آفرینان مستتر در این شبکه‌ها، بیش از شبکه‌هایی نیست که تراکم نقش‌آفرینان در آن‌ها کمتر است. ساختار خرد یک نگاشت، فرد را از موانع افتراقی[1] و فرصت‌هایی که هر نقش‌آفرین به تنهایی با آن مواجه است و رفتار اجتماعی آنها را شکل می‌دهد، مطلع می‌نماید. برای مثال نقش‌آفرینی که ارتباط‌های خیلی زیادی نسبت به دیگر نقش‌آفرینان دارد ممکن است، خیلی بانفوذ و جایگاه اجتماعی بالاتری داشته باشد. جنبه‌های زیادی از رفتارهای کلان شبکه‌ها، تحت تأثیر افکاری هستند که از ساختار آن‌ها تبعیت می‌کنند. توانایی پاسخگویی سریع به محرک، میزان و کامل بودن اشاعه، توانایی شناسایی و ایجاد راه حل‌های بدیع برای مشکلات جدید و نهادینه کردن همکاری میان نقش‌آفرینان، همگی متأثر از الگوهای ارتباطی میان نقش‌آفرینان هستند (لی- چون[2] و دیگران، ۲۰۰۶).

مقالات و گزارش‌های علمی و پژوهشی عمدتاً نتیجه کار نویسندگان متعدد هستند (آندرس[3]، ۲۰۰۹) و همان‌طور که پوزنر[4] اشاره می‌کند آثار علمی دانشگاهیان بیش از پیش نتیجه کارهای گروهی است. از آنجا که پیشرفت علم نتیجه فعالیت‌های جمعی است (پوزنر، ۲۰۰۱)، مطالعه کم و کیف همکاری میان دانشمندان، موضوع جالب توجهی است که طی چندین دهه مورد توجه پژوهشگران حوزه علم‌سنجی قرار داشته است. همکاری علمی فرآیندی است که طی آن دو یا چند نویسنده با هدف خلق اثری مشترک، منابع و استعدادهای خود را به اشتراک می‌گذارند. بررسی متون، گویای این امر است که همکاری علمی در قالب پدیده هم نویسندگی تجلی می‌یابد و یکی از شکل‌های همکاری علمی، هم نویسندگی است، که همکاری در تولید علم از قبیل مقاله، یا سایر قالب‌های آثار علمی را شامل می‌شود. در سال‌های اخیر نیز مجله‌های معتبر ترجیح می دهند مقاله‌هایی را چاپ کنند که حاصل تلاش مشترک چند نویسنده باشند (حریرچی، ملین و اعتماد، ۲۰۰۷؛ به نقل از حریری و نیکزاد، ۱۳۹۰).

همکاری علمی، پدیده‌ی پیچیده‌ای است که اشتراک توانمندی‌ها و تولید دانش علمی جدید را بهبود می‌بخشد. همکاری علمی با افزایش پیچیدگی دانش و به واسطه افزایش تقاضا برای تخصصی شدن بیشتر و مهارت‌های بین‌رشته‌ای در پژوهش ایجاد شده است. به عبارتی، همکاری علمی پدیده‌ای است که از طریق انواع مختلفی از تعاملات بهبود بخش ارتباطات، اشتراک توانایی‌ها و تولید دانش علمی مشخص می‌گردد. اغلب برای توصیف الگوهای همکاری علمی که توسط روابط هم نویسندگی تعریف شده‌اند، از تحلیل شبکه-های اجتماعی استفاده می‌شود (استفانو، جیوردانو و ویتیل[5]،۲۰۱۱).

1 Differential Constraints
2 Li-Chun
3 Andrés
4 Posner
5 Stefano, Giordano &Vitale

درک تنوع نظری درون حوزه علم اطلاعات نیازمند درک ساختار همکاری علمی در این حوزه است. هرچند حرکت مستقیم از فضای اندیشه به ساختار شبکه معمولاً شفاف نیست، ادعاهای مربوط به اجماع نظر در جامعه‌شناسی، سه ساختار متمایز همکاری را پیشنهاد می‌کند. با توجه به اینکه حوزه علم اطلاعات ارتباط فراوان و مبانی مشترک بسیاری با جامعه‌شناسی دارد به این سه ساختار اشاره می‌گردد. ساختار اول اینکه، خیلی‌ها اعتقاد دارند که جامعه‌شناسی هیچ نظریه جامعی[1] ندارد، اما در عوض از لحاظ نظری تجزیه شده[2] و از متخصصان پژوهشی متعددی که با هم ارتباطی ندارند، تشکیل شده است. نویسندگان عقیده دارند که واکنش در برابر کارکردگرایی[3]، رشد سریع، فشارهای سازمانی برای بهره‌وری، تغییر فنون پژوهش و یا تغییر در محیط تأمین بودجه برای علوم اجتماعی جهت تولید متخصصان پژوهشی خودمحور[4]، با فنون پژوهشی منحصربه‌فرد و استانداردهایی برای ارزیابی شواهد اثر متقابل داشته است (به نقل از مودی، ۲۰۰۴). این توصیف، شبکه اجتماعی کاملاً دسته‌بندی شده‌ای[5] را پیشنهاد می‌کند. دوم اینکه سایرین اظهار داشته‌اند که تولید علمی به‌طور قطعی به تعداد معدودی نخبه علمی بستگی دارد، نخبه‌های علمی که آثارشان دوره‌ی کوتاه مدتی از یک رشته را شکل می‌دهد.

نخبه‌های علمی سطح نامتناسبی از یافته‌های پژوهشی، ملاقات علمی بسیار زیاد و دانشجویان و همکاران خیلی زیاد را جذب می‌کنند. نظام‌های ستاره‌ای توزیع نابرابری از فراگیری را در شبکه‌های همکاری پیشنهاد می‌کنند. در نهایت تغییرات در شیوه پژوهشی بایستی با مرزهای نظری نفوذپذیر تعامل داشته باشد تا امکان همکاری‌های با دامنه گسترده را فراهم آورد که با تخصص پژوهشی محدود نمی‌گردد (آبوت[6]، ۲۰۰۱؛ هادسون[7]، ۱۹۹۶).

به عنوان مثال افزایش در روش‌های کمی پیچیده که به‌طور ذاتی خنثی[8] هستند، امکان همکاری بین افرادی دارای مهارت‌های فنی عام و افراد مشغول به کار روی پرسش‌های تجربی ویژه‌ای را پدید می‌آورد. این فرآیند شبکه‌ی همکاری منسجم از لحاظ ساختاری بسیار دسترس‌پذیر را ارائه می‌کند.

مدت‌هاست که نظریه پردازان اعتقاد دارند که ایده‌های یک شخص تابعی از جایگاه وی در زمینه اجتماعی است که به‌طور عمیقی توسط الگوهای تعاملی، ساختاربندی شده است (دورکیم[9]، ۱۹۸۴، مانهیم[10]،۱۹۳۶، سیمل[1]، ۱۹۵۰، به نقل از مودی، ۲۰۰۴). برای مثال، کوهن (۱۹۷۰) استدلال می‌کند که اعتقاد

1 Overarching
2 Fractured
3 Functionalism
4 Self-Contained
5 Clustred
6 Abbott
7 Hudson
8 Substantively Neutral
9 Durkheim
10 Mannheim

به روایی تجربی[2] یک نظریه می‌تواند مدتی طولانی پس از شواهد تجربی موجود، قدرت خود را حفظ کند، مشروط بر اینکه، دانشمندانی در جوامع پژوهشی باشند که به صورت خودکار داده‌ها را به روش‌های مشابهی تفسیر کنند. این چشم‌انداز تلویحی در اثر کرین[3] (۱۹۷۲) که توسعه سریع ایده‌های جدید را به ساختار اجتماعی دانشگاه‌های نامرئی کوچک پیوند داده، آمده است. کرین متوجه گردید که متخصصان پژوهشی توسط گروه‌های هسته، از دانشمندانی که با همدیگر همکاری می‌کنند و توده نامتناسبی[4] از ایده‌های جدید را تولید می‌کنند، مشخص می‌گردند (به نقل از مودی، ۲۰۰۴).

دانشمندانی که در شبکه‌های همکاری قرار می‌گیرند، ایده‌هایشان را به اشتراک می‌گذارند، از فنون و روش‌های مشابهی برای استخراج و تحلیل داده استفاده می‌کنند و به عبارت دیگر بر کارهای یکدیگر تأثیر می‌گذارند.

نحوه محاسبه میزان همکاری گروهی

کاربرد این پدیده در متون تا جنگ جهانی اول رشد آرامی داشت و بعد از آن رشد سریع‌تری پیدا کرد (بیور و روزن[5]، ۱۹۸۷) مطالعات نشان می‌دهد که در سال‌های اخیر همکاری علمی و به ویژه هم نویسندگی در میان نویسندگان و پژوهشگران رشد تصاعدی داشته است. به گونه‌ای که نرخ رشد آن، در مواردی حتی ۲/۸ بیش از نرخ رشد تولیدات علمی بوده است (عصاره و ویلسون،۲۰۰۲).

اقبال گستردهٔ نویسندگان از پدیدهٔ همکاری علمی موجب شد، که پژوهشگران برای گسترش کاربردهای آن تلاش کنند که از آن جمله ارائهٔ ضریب همکاری علمی[6] به منظور مطالعهٔ روند رشد هم نویسندگی را می‌توان نام برد. کاربرد این ضریب را می‌توان در مقالات عصاره (۲۰۰۶) و سهیلی و عصاره (۱۳۸۷) و مانند آن، ملاحظه کرد. در ایران نیز پدیدهٔ همکاری علمی مورد توجه واقع شده و نویسندگان آثار خود را به سه صورت با همکاری همکاران در درون سازمان، با همکاری همکاران در سازمان‌های دیگر یا با همکاری هم‌فکرانشان در خارج از کشور تولید و چاپ می‌کنند.

اما نکتهٔ قابل تأمل و تا حدودی نگران‌کننده در این است که هدف اصلی همکاری علمی کمتر مورد توجه نویسندگان واقع شود و فقط شکل صوری نویسندگی مورد استفاده قرار گیرد. نگارندگان این سطور شاهد چند مقاله با این شرایط بوده‌اند. به عبارت دیگر به جای همکاری نزدیک و علمی پدیدآورندگان در خلق آثار علمی مشترک، مثلاً چهار نویسنده، هرکدام، یک مقاله به صورت انفرادی تألیف کنند و سپس نام

---

1 Simmel
2 Empirical Validity
3 Crane
4 Disproportionate Volume
5 Beaver & Rosen
6 Collaboration Coefficient (Cc)

خود را در بالای هر چهار مقاله در توالی‌های متفاوت اضافه و ادعا کنند که آثار مشترک خلق کرده‌اند. این رویکرد غیر اخلاقی، علاوه بر اینکه کیفیت بالاتر و استنادهای بیشتری برای آن مقالات را در بر نخواهد داشت، کار نادرستی بوده و نقض کامل ایدهٔ ظریف و ایجادکنندهٔ همکاری علمی در میان نویسندگان است.

هم نویسندگی

هم نویسندگی[1] رسمی‌ترین جلوه همکاری فکری میان نویسندگان در تولید پژوهش‌های علمی است و عبارت است از مشارکت دو یا چند نویسنده در تولید یک اثر که منجر به تولید بروندادی علمی با کمیت و کیفیت بالاتری نسبت به تولید و انتشار فردی اثر، می‌شود (هادسون[2]، ۱۹۹۶).

یکی از کاربردهای ویژه تحلیل شبکه‌های اجتماعی بررسی شبکه‌های همکاری علمی است که به‌طور خاص شبکه‌های وابستگی[3] هستند که در آنها شرکت‌کنندگان در گروهی از یک گونه یا گونه‌های مختلف با هم همکاری می‌کنند و گره‌های بین یک جفت از نقش‌آفرینان به وسیله عضویت رایج گروه همانند باشگاه‌ها، گروه‌ها یا مدارس به وجود می‌آید. وقتی که فنون تحلیل شبکه‌های اجتماعی برای کاوش نوعی از شبکه وابسته ثبت شده در انتشارات مجله‌ها استفاده می‌گردد، هم نویسندگی نامیده می‌شود. هم نویسندگی در مقاله یک مجله، می‌تواند به عنوان مستند کردن همکاری بین دو یا چند نویسنده مطرح باشد (پیرسون[4]، ۱۹۹۶، نیومن[5]، ۲۰۰۴، جنست و ثایبولت[6]، ۲۰۰۱). مجموعه‌ای از این همکاری‌ها درون یا در بین مجله‌ها ممکن است شبکه هم نویسندگی را شکل دهند که در آن شبکه، نقش‌آفرینان نویسندگان هستند و گره بین دو نقش‌آفرین رابطه هم نویسندگی است که در مقاله‌های مجله‌ها ایجاد می‌شود. پژوهشگران هم نویسندگی را رابطه‌ای فکری و نیز بین فردی می‌دانند که فرصتی برای شناسایی، سنجش فعالیت‌های اجتماعی، نفوذ و اعتبار درون یک رشته خاص را فراهم می‌نماید (استوک و هارتلی[7]، ۱۹۸۹؛ پیترز و وان ران[8]، ۱۹۹۱).

بررسی گره‌های شبکه‌های هم نویسندگی میان نویسندگان بیانگر آن است که نویسندگانی که در حوزه-های شناختی مشابهی همانند علم اطلاعات کار می‌کنند، ممکن است تلاش‌های مشترکی در حجم و با انسجام (پیوستگی) مختلفی را نشان دهند- برخی به هم متصل، برخی هم از سایرین مجزا و تنها هستند. تحلیل این الگوها می‌تواند به پاسخگویی سؤالاتی از این قبیل کمک نماید. کدام نویسنده نقش مهم‌تری ایفا می‌نماید؟ و

---

[1] معادل‌های فارسی مختلفی برای این واژه (Co-Authorship) بکار برده شده است ازجمله آنها می‌توان به هم نویسندگی، تألیف مشترک، هم تألیفی، همکاری در تألیف و... اشاره کرد در این پژوهش به منظور رعایت یکدستی از واژه هم نویسندگی استفاده گردیده است.

2 Hudson
3 Typically Affiliation Networks
4 Persson
5 Newman
6 Genest & Thibault
7 Stokes & Hartley
8 Peters & Van Raan

چه کسی گروه‌های همکاری مختلف را به هم متصل می‌نماید؟ بنابراین روش‌های شبکه ممکن است چشم‌انداز مفیدی باشند که بتوان از طریق آن‌ها وضعیت یک حوزه علمی را بررسی نمود.

به‌طور کلی برای مطالعه همکاری‌های پژوهشی، دو روش به کار می‌رود، یکی از آنها تحلیل هم استنادی است، جایی که پیوندهای بین پژوهشگران از طریق ارجاع به پژوهش‌ها و انتشارات یکدیگر به وجود می‌آید. روش دیگر تحلیل داده‌های هم نویسندگی است. این دو روش از لحاظ دامنه به‌طور گسترده‌ای با هم متفاوت‌اند.

تحلیل استنادی ساختار شناختی جوامع علمی را ترسیم می‌کند و ضرورتاً ساختار اجتماعی و شبکه‌هایی را که به واسطه همکاری شکل گرفته‌اند منعکس نمی‌کند (راچرلا و هو،۲۰۱۰). همان گونه که استکوکس و هارتلی مشاهده کردند "استنادها در یک محیط دانشگاهی شرایط فکری را تأیید می‌کنند، نه شرایط شخصی نویسنده را " (استوک و هارتلی، ۱۹۸۹). از سویی دیگر هم نویسندگی که بین پژوهشگران ارتباط برقرار می-کند، هر دو جنبه‌ی فکری و دیون شخصی را تصدیق می‌کند و بنابراین فرصتی برای شناسایی و سنجش وسعت فعالیت‌های اجتماعی و تأثیرگذاری در متخصصان علمی فراهم می‌کند؛ به عبارت دیگر تحلیل استنادی ممکن است به شناسایی مقالات علمی مهم و مرکزی کمک نماید، در حالی که تحلیل هم نویسندگی مهم‌ترین دانشمندان را شناسایی می‌کند (راچرلا و هو، ۲۰۰۸).

هم نویسندگی در حالی که مانند شبکه‌های استنادی و هم‌استنادی است، اما شبکه‌های همکاری، گره-های اجتماعی قوی‌تری را شامل می‌شوند. استنادها ممکن است بدون اینکه نویسندگان همدیگر را بشناسند، صورت بگیرند و ممکن است در طی زمان ادامه یابد. در حالی که هم نویسندگی مستلزم وجود رابطه‌ای موقتی بین همکاران است و این سبب می‌گردد تا در دامنه تحلیل شبکه‌های اجتماعی قرار بگیرند؛ به عبارت دیگر هم نویسندگی مستلزم پیوندهای اجتماعی قوی‌تری نسبت به استناد است. در هم نویسندگی نویسندگان معاصر و با هم آشنا هستند (لیو و دیگران، ۲۰۰۵).

شبکه‌ها، برای کتاب‌سنجی پدیده‌ی جدیدی نیستند، زیرا کتاب‌سنجی حوزه‌ای است که تاریخچه‌ای طولانی از مطالعات شبکه‌های استنادی دارد، این شبکه‌ها با استفاده از استنادهای بین مقالات شکل می‌گیرند. شبکه‌های استنادی کاملاً از شبکه‌های هم نویسندگی متمایز هستند. گره‌های شبکه‌های استنادی، مقالات هستند نه نویسندگان و پیوندهای بین آنها استنادها هستند نه هم نویسندگی. حال آن که شبکه هم نویسندگی، شبکه‌ای است که جوامع دانشگاهی و علمی را به نمایش می‌گذارد، همچنان که شبکه‌ای است که نشان‌دهنده‌ی ساختار دانش است. شاید به همین دلیل، توجه بسیار کمتری در مقایسه با شبکه‌های استنادی به شبکه‌های هم نویسندگی شده است (نیومن، ۲۰۰۴).

نقطه شروع تحلیل الگوهای هم نویسندگی در علم اطلاعات و در کتاب‌سنجی است (اگه و روسو[1]، ۱۹۹۰). نویسندگان متعددی شبکه‌های هم نویسندگی را در دهه‌های گذشته مورد ملاحظه قرار داده‌اند. نیومن شبکه‌های هم نویسندگی در نواحی متعددی از پژوهش‌های علمی در مجموعه‌ای از مقالات، زیست‌پزشکی، فیزیک و حوزه‌های فرعی آن، ریاضیات و علوم رایانه مورد مطالعه قرار داده است (نیومن، ۲۰۰۱ و ۲۰۰۴). بارباسی و دیگران (۲۰۰۲) شبکه‌های هم نویسندگی در علوم ریاضیات و علوم عصب‌شناسی و مودی (۲۰۰۴) شبکه‌های هم نویسندگی علوم اجتماعی را مورد مطالعه قرار داده‌اند.

مولینز[2] (۱۹۷۳) چهار نوع از روابط اجتماعی را که ممکن است بین پژوهشگران وجود داشته باشد، شناسایی کرد: هم نویسندگی، ارزیاب معتمد[3]، همکلاسی[4]، شاگردی[5]. هم نویسندگی شامل اجتماعی به شدت شدت بسته است، جایی که دو یا بیش از دو پژوهشگر در یک پژوهش مشترک به کار مشغول‌اند. کاربرد تحلیل شبکه‌های اجتماعی در ارتباطات علمی، تحلیل هم نویسندگی نامیده می‌شود. هم نویسندگی مستند کردن همکاری بین دو یا چند نویسنده است. مجموعه‌ای از این همکاری‌های در سراسر مجله‌ها می‌تواند یک شبکه هم نویسندگی را شکل دهد که در آن نقش‌آفرینان نویسندگان هستند و یک گره بین دو نقش‌آفرین توسط هم نویسندگی مقالات مجله‌ها ایجاد می‌گردد. همان‌طوری که ذکر شد، هم نویسندگی که بین پژوهشگران ارتباط برقرار می‌کند و هم شرایط فکری و شخصی را مورد توجه قرار می‌دهد؛ بنابراین فرصتی برای شناسایی و سنجش فعالیت‌های اجتماعی و تأثیرگذاری درون یک رشته خاص را فراهم می‌کند (پیترز و وان ران، ۱۹۱؛ استوک و هارتلی، ۱۹۸۹). بررسی گره‌های شبکه‌های هم نویسندگی میان نویسندگان می‌تواند آشکار سازد که نویسندگانی که در حوزه‌های شناختی مشابه کار می‌کنند، ممکن است تلاش‌های مشترکی را به نمایش بگذارند. تحلیل این الگوها می‌تواند به پاسخگویی سؤالاتی از قبیل، کدام نویسندگان نقش‌های مهم-تری را ایفا می‌کنند و چه کسی مجموعه گروه‌های مختلف را در شبکه به هم وصل می‌کند، کمک کند؛ بنابراین روش‌های شبکه می‌توانند برای کسانی که وضعیت یک حوزه را بررسی می‌کنند چشم‌انداز مفیدی را ارائه نماید (چنگ، ۲۰۰۶).

یکی از کاربردهای تحلیل شبکه‌های اجتماعی در ارتباطات علمی، تحلیل هم نویسندگی است. همان‌طوری که ذکر شد، تحلیل هم نویسندگی بر این اصل تأکید دارد که وقتی دو یا بیش از دو پژوهشگر به صورت مشترک یک مقاله را می‌نویسند پیوندهای فکری و اجتماعی بین آنها به وجود می‌آید (استوک و هارتلی، ۱۹۸۹). تحلیل هم نویسندگی تعاملاتی را که بین نقش‌آفرینان رخ می‌دهد و از طریق انواع نگاشت به

1 Egghe & Rousseau
2 Mullins
3 Trusted Assessorship
4 Colleagueship
5 Apprenticeship

نمایش درمی‌آید، تحلیل می‌کند. ارتباط دادن اطلاعات جمعیت شناختی نویسندگان با نویسندگان مسئول مقاله‌ها و مجله‌ها، می‌تواند تصویری تفصیلی از جنبه‌های متعدد تعاملات علمی را فراهم کند؛ بنابراین تحلیل هم نویسندگی به شناسایی الگوهای اصلی فعالیت‌های پژوهشی کمک می‌کند، ضمن اینکه تصویر مفصلی از شبکه رسمی همکاری که در درون آن شبکه این تعاملات صورت می‌گیرد را فراهم می‌کند (پیرسون[1]، ۱۹۹۶ به نقل از چنگ، ۲۰۰۶).

هم نویسندگی یکی از ملموس‌ترین و مستندترین شکل‌های همکاری علمی است. اغلب، هر جنبه‌ای از شبکه‌های همکاری علمی می‌تواند به‌طور موثقی توسط تحلیل شبکه‌های هم نویسندگی با استفاده از روش-های کتاب‌سنجی ردیابی گردد (گلنزل و شوبرت[2]، ۲۰۰۴). همکاری علمی یک پدیده پیچیده اجتماعی در پژوهش است که به صورت نظام‌مند از ۱۹۶۰ مورد مطالعه قرار گرفته شده است و افزایش میزان رشد همکاری توسط پژوهشگران مختلفی گزارش گردیده است. سرمایه زیاد، کارهای گروهی، تغییر الگوهای ارتباطاتی و افزایش تحرک پذیری دانشمندان از عوامل تأثیرگذار بر افزایش همکاری‌های علمی هستند. این عوامل، همکاری را در حوزه‌های کم‌هزینه مانند ریاضیات محض و پژوهش‌های نظری در علوم اجتماعی را تهییج می‌کند.

ساده‌ترین پاسخ به این سؤال که چرا پژوهشگران با هم همکاری می‌کنند، این است که کار کردن با دیگران کیفیت آثار تولیدی را بهبود می‌بخشد. شواهدی در متون وجود دارد که نشان می‌دهد، بین همکاری علمی و کیفیت آثار علمی تولیدی ارتباط وجود دارد. به علاوه با توجه به پیدایش حوزه‌های بین‌رشته‌ای و گسترش سریع علوم، یک فرد به تنهایی قادر نخواهد بود که در بخش‌های ریزتر و فرعی یک حوزه علمی اشراف کافی داشته باشد. این امر نیز به نوبه خود سبب افزایش گرایش نویسندگان به سمت کارهای مشترک می‌گردد. در این راستا گرایش افراد به سمت نویسندگان کلیدی و هسته بسیار بیشتر از سایر نویسندگان است. نویسندگانی که در شبکه‌های اجتماعی نقش محوری و در مرکزیت نقش‌آفرینان قرار دارند، از موقعیت مهم-تری برخوردارند و سایر پژوهشگران تمایل بیشتری دارند تا با آنها در تولید آثار علمی مشترک همکاری نمایند. شناسایی این نویسندگان تأثیرگذار و ساختارهای موجود در شبکه‌های هم نویسندگی به سایر پژوهشگران کمک خواهد نمود تا با انتخاب راهبردهای مناسب، بهتر بتوانند راهبرد خاص هم نویسندگی خود را انتخاب نمایند (هارت، ۲۰۰۰).

1 Persson
2 Glänzel & Schubert

هم نویسنده‌ها بر اساس پیشینه‌های پیشین آنها به چهار نوع طبقه‌بندی می‌شوند: گروه‌های پایدارها[1]، زودگذرها[2]؛ تازه واردها[3] و پایان‌بخش‌ها[4]. رابطه بین هم نویسندگی و فعالیت انتشاراتی نویسنده، اطلاعاتی درباره نقش بالقوه هم نویسنده‌ها در شکل‌دهی واژگان ثابت یا ایجاد پیوندهای موقعیتی را آشکار می‌سازد. پرایس و گورسی[5] (۱۹۷۶) طرحی ماهرانه از آنچه که آنها آن را "آمار احصایی[6] جوامع علمی " نامیدند فراهم کردند. بر اساس تعریف پرایس و گورسی، زودگذرها نویسندگانی هستند که در یک سال معین منتشر کرده‌اند اما نه قبل و نه بعد از آن؛ تازه واردها نویسندگانی هستند که در سالی معین و بعد از آن منتشر کرده‌اند اما قبلاً هرگز منتشر نکرده‌اند. پایان‌بخش‌ها قبلاً در یک سال معین منتشر کرده‌اند اما نه هرگز بعد از آن؛ و پایدارها قبل و بعد از سالی معین منتشر کرده‌اند (به نقل از گلنزل و شوبرت، ۲۰۰۹).

مطالعه شبکه‌های هم نویسندگی ذاتاً به دلایل متعددی جذاب است: اول این‌که مطالعه این شبکه‌ها به ما فرصت می‌دهد تا نخبه‌های جامعه‌سنجانه[7] درون حوزه را که حداقل نویسندگانی که در میانه چندین گروه با علائق و موضوعات گوناگون و نه با دانشمندان برجسته آن حوزه مرتبط هستند، شناسایی کنیم و بنابراین به طیف وسیعی از موضوعات در این حوزه بپردازیم. دوم این‌که از آنجایی که بسیاری از تلاش‌های سرمایه-گذاری در جهت ایجاد شبکه‌هایی از دانشمندان مرتبط با یکدیگر است، تحلیل شبکه اجتماعی به ما فرصت می‌دهد که تأثیر واقعی آنها را در جامعه علمی بسنجیم. در نهایت یک نقشه جامع از شبکه اجتماعی در حال تکامل می‌تواند اثرات کاربردی بر کارهای مرتبط به داوری همایش‌ها و مجله‌های علمی داشته باشد، بدین صورت که در انتخاب و بررسی منتقدان مقالات بر اساس درجه جدایشان مورد استفاده قرار می‌گیرند (کوتا و مرلو[8]، ۲۰۰۷). هم نویسندگی برای سنجش همکاری مزایایی دارد که عبارت‌اند از: نتایج بررسی هم نویسندگی تا حدود زیادی غیرقابل تغییر و معتبر بوده و روش عملی و ارزانی برای کمی سازی همکاری است. هم نویسندگی می‌تواند نمونه‌های خیلی بزرگ را تطبیق دهد تا نتایج آماری خیلی موثق را فراهم نماید (کتز[9]، ۱۹۹۲؛ به نقل از رودریگوز، گومز و فلیکس[10]، ۲۰۱۰). هم نویسندگی مقالات در مجله‌های علمی دریچه‌ای را به الگوهای همکاری درون جوامع دانشگاهی باز می‌کند. هم نویسندگی یک مقاله، مستند کردن همکاری بین دو یا چند نویسنده است و این همکاری یک شبکه هم نویسندگی را شکل می‌دهد که در این شبکه گره‌ها نویسنده‌ها را نمایش می‌دهند و دو نویسنده توسط خطی به هم متصل می‌گردند، مشروط بر اینکه

1 Continuants
2 Transients
3 Newcomers
4 Terminators
5 Price & Gürsey
6 Actuarial
7 Sociometric Stars
8 Cotta And Merelo
9 Katz
10 Rodrı´Guez, Go´Mez & Fe´Lix

آن‌ها یک یا بیش از یک مقاله نوشته باشند. ساختار این گونه شبکه‌ها، در نهایت شاخصه‌های جالب زیادی از جوامع دانشگاهی را آشکار می‌سازند (نیومن، ۲۰۰۴).

احتمال هم نویسنده شدن از یک رشته به رشته دیگر و در طی زمان متفاوت است. هم نویسندگی در علوم طبیعی نسبت به علوم اجتماعی خیلی رایج‌تر است، اما به‌طور یکنواختی در کل رشته‌ها افزایش پیدا کرده است (مودی، ۲۰۰۴). به‌طور کلی هم نویسندگی پیوند یا گرهی را بین دو پژوهشگر به وجود می‌آورد. این پیوندها می‌توانند به عنوان یک شبکه اجتماعی مورد بررسی قرار گیرند و الگوهایی که در شبکه اجتماعی افراد و هم نویسنده‌هایشان نمایش داده می‌شوند، می‌تواند این مطلب را به گونه‌ای قابل ملاحظه روشن نماید که یک نویسنده چگونه کار می‌کند و چگونه با همتایانش تعامل می‌نماید.

شبکه هم نویسندگی

یک شبکه هم نویسندگی، نگاشتی از گره‌های مشترک یا ارتباطات بین هم نویسنده‌های درون یک جامعه پژوهشی است. دو نویسنده همکار با هم پیوند دارند و اگر قبلاً مقاله‌ای را با هم نوشته‌اند؛ می‌توان گفت آنان با هم ارتباط علمی دارند. مطالعه چنین شبکه‌هایی نگرشی را از درون ساختار اجتماعی جوامع پژوهشی فراهم می‌نماید؛ به عبارت دیگر، این نوع شبکه آشکار می‌سازد کدام یک از نویسندگان همکار در فرآیند ارتباطات در شبکه دارای نقش مرکزی هستند. اولین مطالعه تجربی بر روی شبکه‌های اجتماعی در اثر میلگرام[1] (۱۹۶۷) ثبت گردیده است. اولین مطالعه ثبت شده درباره شبکه‌های هم نویسندگی را می‌توان به جوامع ریاضی نسبت داد، زیرا در سال ۱۹۶۹ مفهوم عدد اردوش، یعنی فاصله همکاری را به ریاضیدان مشهور پائول اردوش[2] نسبت داده‌اند (فت، یوجیم و راتناولو[3]، ۲۰۱۰).

شبکه‌های هم نویسندگی یک رشته ابزارهای مفید جهت بررسی الگوهای پژوهشی و مخصوصاً پژوهش‌های مشترک فراهم می‌کند. بررسی روابط بین هم نویسندگی و بهره‌وری، بررسی هم نویسندگی درون سازمانی بین بخش‌های دانشگاه یا با سایر بخش‌ها (همانند انجمن‌های علمی، تسهیلات پژوهشی ملی و صنعتی) از این جمله است (دورباچ، نایدو و موتون[4] ۲۰۰۸). تجزیه و تحلیل انتشارات هم نویسنده، روشی استاندارد برای سنجش میزان همکاری در پژوهش است. از طرفی ارزیابی شبکه‌های هم نویسندگی، به منظور سنجش پیوند میان مؤسسات و سازمان‌ها، از روش‌های بررسی کم و کیف همکاری علمی به شمار می‌رود

1 Milgram
2 Paul Erdos
3 Fatt, Ujum & Ratnavelu
4 Durbacha, Naidooa & Mouton

(لاندبرگ[1] و دیگران، ۲۰۰۶). شبکه‌ها می‌توانند اطلاعات مفیدی درباره پیوستگی، وابستگی متقابل و پیوند میان پژوهشگران در کشورهای مختلف را نشان دهند (واگنر[2]، ۲۰۰۵، نقل در حریری و نیکزاد، ۱۳۹۰).

شبکه‌های هم نویسندگی پیشینه مستند شده‌یِ مفصل و دقیقی از شبکه‌های اجتماعی و حرفه‌ای دانشمندان را فراهم می‌کنند. داده‌های هم نویسندگی، منبعی با ارزش برای پیگیری سؤالاتی همانند چگونگی ساختار انجمن‌ها یا دانشگاه‌های نامرئی و چگونگی الگوهای همکاری در طی زمان را ارائه می‌نمایند (نیومن، ۲۰۰۴). شبکه هم نویسندگی دانشمندان، الگویی از شبکه‌های پیچیده و در حال توسعه را نمایش می‌دهد. به علاوه یکی از گسترده‌ترین پایگاه‌های اطلاعاتی در شبکه‌های اجتماعی را عرضه می‌نماید (بارباسی، ۲۰۰۲). شبکه‌های هم نویسندگی طبقه‌ی مهمی از شبکه‌های اجتماعی هستند و به‌طور گسترده می‌توانند برای تعیین ساختار همکاری‌های علمی و وضعیت پژوهشگران به صورت انفرادی به کار برده شوند (لیو و دیگران، ۲۰۰۵).

Section ۱,۰۱ ساختار شبکه‌های هم نویسندگی

در طی سده‌های گذشته انتشار مقاله‌های علمی بیشتر به صورت انفرادی صورت می‌گرفت تا اینکه تلاش مشترک دو، سه و یا چند نویسنده که منابع، هوش و استعدادشان در تولید آثار را به اشتراک بگذارند. از لحاظ تاریخی این پدیده بیشتر در حوزه‌های علوم و علوم اجتماعی مشاهده گردیده است.

متخصصان علم‌سنجی، جامعه شناسان و روانشناسان مقاله‌های نسبتاً زیادی درباره جنبه‌های مختلف ساختار همکاری علمی به نگارش درآورده‌اند. متخصصان علم‌سنجی تلاش کرده‌اند تا الگوهای ریاضی را ایجاد نمایند که انواع متنوعی از ساختارهای عمومی و خاص را آشکار و توصیف می‌کنند؛ در حالی که جامعه‌شناسان و روانشناسان همیشه می‌خواهند، تفاسیر تاریخی، اجتماعی و روان‌شناختی را برای این ساختارها بیابند (لیانگ و دیگران، ۲۰۰۱). به‌طور همزمان دانشمندانی از حیطه‌های علم‌سنجی، جامعه‌شناسی علم و روانشناسی همچنین توجه خویش را بر سابقه شغلی متمرکز کرده و روابط بین بهره‌وری، تأثیرگذاری دانشگاهی دانشمندان و شناخت توسط نظام علم را مورد مطالعه قرار داده‌اند (سوین، ۱۹۹۰، فاکس، ۱۹۸۳، اور، ۱۹۸۹، به نقل از لیانگ و دیگران، ۲۰۰۱). حتی علاوه بر مسائل ذکر شده، رابطه بین تغییر مکان مرکز جهانی علم و سن اجتماعی دانشمندان مورد بررسی قرار گرفته است (ژائو و جیانگ، ۱۹۸۵).

ساختار شبکه، بر اطلاعات قابل دسترسی برای افراد و فرصت‌های همکاری تأثیر می‌گذارد. ساختار شبکه همچنین بر جریان کلی اطلاعات و طبیعت جوامع علمی تأثیرگذار است (لی- چون و دیگران، ۲۰۰۶).

1 Lundberg

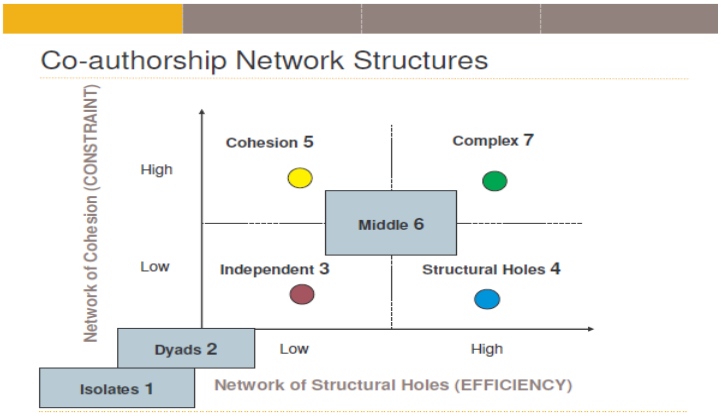
2 Wagner

به‌طور خلاصه یک مدرک علمی اگر بیش از یک نویسنده داشته باشد آن مدرک را هم نویسنده می‌نامند. هم نویسندگی رابطه‌ای را بین کل پژوهشگران دانشگاهی ایجاد می‌کند که مجموعه این روابط، شبکه‌ای فرد محور را برای یک شبکه بزرگ و یا یک شبکه دانشگاهی بین پژوهشگران به وجود می‌آورد. بر اساس نظریه شبکه، ساختار شبکه‌های فرد محور انعکاس دهنده دو ساختار هسته‌ی منسجم[1] و گسست‌های ساختاری می‌باشند.

با وجود این دو ساختار هسته در ساختار شبکه‌های خودمحور، پنج راهبرد هم نویسندگی شناسایی گردیده است که عبارت‌اند از: راهبردهای هم نویسندگی مستقل[2]، گسست‌های ساختاری، منسجم، میانه[3] و پیچیده[4] (رامزی وایرپو، ۲۰۰۶). اگر چه ممکن است پژوهشگرانی وجود داشته باشند که در هیچ کدام از این ساختارها قرار نگیرند، به عنوان مثال پژوهشگری که ساختار منزوی[5]- مقالاتی با یک نویسنده- را نشان می‌دهد، یا پژوهشگرانی که به صورتی جفتی[6]- نویسندگانی که تنها با یک نفر هم نویسنده‌اند- می‌باشند. این ساختارها نمی‌توانند گسست‌های ساختاری را پر کنند و یا اینکه انسجام را به وجود بیاورند، بنابراین در ساختار اصلی شبکه هم نویسندگی خودمحور قرار نمی‌گیرند (شکل ۳-۱). پژوهشگرانی که منطقه ۳ (مستقل) قرار می‌گیرند نویسندگانی هستند که کارآمدی و محدودیت پایینی دارند، نویسندگان منطقه ۴ (گسست‌های ساختاری) نویسندگانی هستند که کارآمدی بالا و محدودیت پایینی دارند، نویسندگان منطقه ۵ (منسجم) نویسندگانی هستند که محدودیت بالا و کارآمدی پایینی دارند و نویسندگانی که در منطقه ۷ قرار می‌گیرند نویسندگانی هستند که هم کارآمدی و هم محدودیت بالایی دارند. این نویسندگان معمولاً پر تولیدترین و تأثیرگذارترین نویسندگان در هر حوزه پژوهشی هستند.

1 Cohesive

2 Independent
3 Middle
4 Complex
5 Isolate
6 Dyadic



شکل ۳-۱: ساختار شبکه‌های هم نویسندگی (رامزی وایرپو، ۲۰۰۶)

Section ۱,۰۲ تحلیل شبکه‌های هم نویسندگی

موضوعی‌های اصلی که در تحلیل شبکه‌های هم نویسندگی و تعریف شبکه وجود دارد، عبارت‌اند از:

۱) جمع‌آوری داده‌ها، ۲) تعیین مرزهای شبکه؛ ۳) تعریف ماتریس دادهای هم نویسندگی و ۴) تحلیل داده‌های شبکه و تفسیر نتایج (استفانو، گیوردانو و ویتیل[1]،۲۰۱۱). در اینجا سعی می‌شود به توصیف مختصر آنها پرداخته شود.

۱- جمع‌آوری داده‌ها: حتی اگر همکاری میان پژوهشگران با استفاده از هم نویسندگی تعریف شده باشد، احتمال استفاده از منابع داده‌ای بی‌شمار و ناهمگن در جمع‌آوری داده‌های کتابشناختی وجود دارد. معمولاً این داده‌ها به جای مصاحبه با نویسنده مقاله یا پرسشنامه که مستقیماً به نویسنده مقاله داده می‌شود، با استفاده از اسناد آرشیوی جمع‌آوری می‌گردند (از طریق پایگاه‌های اطلاعاتی محلی یا بین‌المللی). اگر علاقه مطالعاتی جمع‌آوری داده‌هایی در یک جمعیت هدف یا یک جامعه علمی خاص است، نوع آرشیوی که مورد استفاده قرار می‌گیرد هم می‌تواند بر تعریف گره‌ها در شبکه هم نویسندگی و هم بر جامعه تحت پوشش تأثیر بگذارد.

۲- تعیین مرزهای شبکه: ارائه تعریفی از مرزهای شبکه به خاطر حضور مرزهای ناملموس[2] یا مبهم، اغلب مشکل است. در مطالعات شبکه، طرح شبکه فرد محور و جمع محور شبکه باید مورد استفاده قرار بگیرد (مارسدن[3]، ۲۰۰۵). اگر طرح کل شبکه دنبال می‌گردد، مجموعه‌ای از نقش‌آفرینان وابسته به هم به عنوان یک جمعیت اجتماعی دارای مرزبندی مشخص در نظر گرفته می‌شود. وقتی که از یک طرح فرد محور استفاده می‌گردد، مطالعه بر روی تعدادی از نقش‌آفرینان مرکزی و روابط آن‌ها در محدوده‌ی آن طرح متمرکز است.

1 Stefano, Giordano &Vitale
2 Soft
3 Marsden

وقتی قرار است الگوی هم نویسندگی میان پژوهشگران در یک جامعه علمی توصیف شود، رویکرد کل شبکه می‌تواند مورد توجه قرار گیرد و باید تصمیم گرفت که شبکه کدام نویسندگان را در برگیرد. لامان[1] و دیگران در سال ۱۹۸۹ سه رویکرد جامع تعیین مرزهای شبکه را تعریف نمودند: ۱) رویکرد موقعیتی[2] که بر ویژگی‌های نقش‌آفرینان یا معیارهای عضویت رسمی (به عنوان مثال استخدام توسط یک سازمان) تأکید دارد؛ ۲) رویکرد مبتنی بر حوادث[3] که نقش‌آفرینان با استفاده از حوادث هم‌حضوری[4] به هم مرتبط می‌شوند؛ و ۳) رویکرد رابطه‌ای که توسط پیوندهای اجتماعی بین نقش‌آفرینان هدایت می‌شود. رویکرد موقعیتی در هر جایی که به الگوهای بین سازمانی یا بین‌رشته‌ای روابط هم نویسندگی توجه شود، مورد استفاده قرار می‌گیرد. نمونه-هایی از این رویکرد را می‌توان در کارهای یوسفی نورایی[5] و دیگران (۲۰۰۸)؛ گوزارت و اوزمان[6] (۲۰۰۹)؛ و فرلیجو و کرونیجر[7]، (۲۰۰۹) یافت (نقل در استفانو، گیوردانو و ویتیل، ۲۰۱۱).

همچنین برت (۱۹۸۰) اعتقاد دارد که تحلیل‌گر شبکه از دو رویکرد تحلیلی به منظور توضیح اثرات شبکه‌ها در کنش‌های اجتماعی استفاده می‌نماید. رویکرد رابطه‌ای که بر ارتباط‌های مستقیم و غیرمستقیم نقش‌آفرینان با سایرین، یعنی چگونگی پیوند دادن، متمرکز است. روابط اجتماعی در این رویکرد به عنوان کانال‌هایی که از طریق آن‌ها منابع ویژه‌ای همانند اطلاعات، دوستی، کالاها یا جریان پول بسته به نوع جاسازی آن‌ها جابجا می‌گردد، درک می‌شوند (گرانوتر[8]، ۱۹۸۵). رویکرد موقعیتی که بر الگوی روابط نقش‌آفرینان بر اساس گره‌های مشابه با دیگران متمرکز است (برت،۱۹۸۰). این راهبرد بر اساس اندیشه هم‌ارزی ساختاری[9] نقش‌آفرینان استوار است (لوران و وایت[10]، ۱۹۷۱). از لحاظ ساختاری نقش‌آفرینان هم‌ارز نیازمند گره‌های مستقیم بین خودشان یا حتی متعلق به دسته یکسانی نیستند.

نقطه شروع تحلیل، تهیه سیاهه‌ای از پژوهشگران است که به مؤسسه‌ی پژوهشی یا رشته‌ای مشخص وابستگی دارند و علاقه آنان به سمت روابط خویش، بدون توجه به گره‌های ممکنی که با نویسندگان بیرونی برقرار کرده‌اند، است.

رویکرد مبتنی بر حوادث جایی مورد استفاده قرار می‌گیرد که مطالعه مبتنی بر محتوای پایگاه اطلاعاتی پژوهشی معینی باشد و سیاهه‌ای از نقش‌آفرینان موجود در شبکه هم نویسندگی در دست نباشد. در این مورد

1 Laumann
2 Positional
3 Event-Based
4 Co-Attendance
5 Yousefi-Nooraie
6 Gossart And Ozman
7 Ferligoj And Kronegger
8 Granovetter
9 Structural Equivalence
10 Lorraine And White

به‌طور ضمنی فرض می‌شود که هم نویسندگی یک مقاله رویداد رابطه‌ای خاصی را نمایش می‌دهد؛ به عبارت دیگر، نویسندگان موجود در این شبکه‌ها کل نویسندگانی هستند که در آن پایگاه وجود دارند.

این نوع رویکرد مخصوصاً به هنگام بازیابی داده‌ها از پایگاه‌های اطلاعاتی کتابشناختی رشته‌ای، مورد استفاده قرار می‌گیرد. رویکرد رابطه‌ای در اثر دوریان و وودوارد[1] در سال ۱۹۹۲ به عنوان رویه انتخاب فزاینده[2] طرح گردید. زمانی که هدف توصیف الگوی همکاری در گروهی از نویسندگان مشغول در یک رشته رشته و یا وابسته به مؤسسه‌ای خاص باشد، این رویکرد راهبردی مناسب عرضه می‌کند. در واقع این رویکرد امکان تعیین مرزهای شبکه را به وجود می‌آورد و از سیاهه‌ی موقتی[3] از نقش‌آفرینان ثابت به ظاهر موجود در شبکه شروع می‌شود و سپس دیگر نقش‌آفرینان را به وسیله ارتباط‌های مشاهده شده از این هسته اولیه اضافه می‌نماید.

۳- به‌طور کلی تحلیل شبکه‌های اجتماعی بر متغیرهای ساختاری که بر اساس زوج‌هایی از نقش‌آفرینان (داده‌های ارتباطی) سنجیده می‌شوند، مبتنی می‌باشند. داده‌های هم نویسندگی از مجموعه‌ای از مقالات و نویسندگان استخراج می‌گردد که می‌تواند در یک ماتریس وابستگی مرتب گردند؛ بنابراین $A\ (n \times p)$ ماتریس وابستگی است که در آن عناصر کلی[4] aij $(i = 1,..., n; j = 1,..., p)$ برابر با یک است در صورتی که نویسنده i ام در مقاله J ام حضورداشته باشد و در غیر این صورت صفر است. ما می‌توانیم ماتریس مجاورت Gw به اندازه $(n \times n)$ از ماتریس محصول A و برگردان $\grave{A}$ آن استخراج نماییم.

$$Gw = A\grave{A}$$

Gw یک ماتریس مجاورت وزن دهی شده غیر مستقیم را نشان می‌دهد که اگر دو نویسنده هرگز با هم همکاری علمی نداشته‌اند، مدخل‌های آن برابر با صفر است، در جایی دیگر آنها تعدادی مقالات حاصل هم نویسندگی با زوج‌هایی از نویسنده‌ها دارند. در تحلیل شبکه‌های اجتماعی ماتریس Gw معمولاً بعد از حذف مدخل‌های مورب (که تعداد کل مقالات هر نویسنده را نشان می‌دهد) تحلیل می‌گردد و تمام مدخل‌های آن بزرگتر از صفر تا یک می‌باشند و ما سپس می‌توانیم یک ماتریس مجاورت دودویی غیر مستقیم Gb به دست آوریم، جایی که تنها حضور گره‌ها به حساب می‌آید. اگر چه این روش در تحلیل داده‌های هم نویسندگی بسیار رایج است، ولی با در نظر گرفتن ماتریسGb اطلاعات مرتبطی از دست می‌رود. یک نظام وزن دهی جایگزین می‌تواند در نظر گرفته شود. به عنوان مثال وقتی که به هم نویسندگی به عنوان ابزار انتقال دانش علاقه‌مندیم، ما می‌توانیم از وزنی که توسط نیومن در سال ۲۰۰۱ پیشنهاد گردیده است استفاده نماییم. در این

1 Doreian And Woodward
2 Expanding Selection Procedure
3 Provisional
4 Generic

نظام وزن دهی، روابط هم نویسندگی بین دو پژوهشگر در صورتی قوی‌تر است که آنان تنها نویسندگان یک مقاله باشند.

۴- بعد از تعیین ماتریس‌های مجاورت دودویی و وزن دهی‌شده، فنون تحلیل شبکه‌های اجتماعی برای داده‌های هم نویسندگی به کار گرفته می‌شود. این تحلیل این امکان را به وجود می‌آورد تا جریان دانش را در کل یک حوزه علمی توصیف نمود و نقش و موقعیت پژوهشگران درون شبکه را برجسته نمود. به‌طور کلی می‌توانیم نتایج شبکه‌ای که برای سطح جهانی به دست می‌آیند را از سنجش‌های سطح عامل یا برای ماتریس مجاورت دودویی در مقابل وزن دهی‌شده متمایز کنیم. تحلیل شبکه هم نویسندگی می‌تواند بر موارد زیر متمرکز باشد. الف: ویژگی‌های ساختاری شبکه؛ ب) پیکربندی کل شبکه که به دنبال ریخت‌شناسی ویژه‌ای از شبکه می‌گردد و ج) شاخص‌های سطح عامل[1] (شاخص‌های مرکزیت، ضریب خوشه‌بندی و مانند آن). ویژگی‌های ساختاری شبکه را می‌توان با استفاده از تعدادی از شاخص‌های سطح جهانی (تراکم، ارتباط و مانند آن) توصیف نمود که می‌توانند برای ارزیابی ارتباط عامل‌ها به عنوان یک کل و همچنین برای مقایسه شبکه‌های مختلف خیلی مفید باشند.

یافته‌های تجربی پژوهشگران مشخص نموده است که پیکربندی جهان کوچک[2] (واتس و استر گوتز[3]، ۱۹۹۸) و بدون مقیاس[4] (باربا‌سی و آلبرت[5]، ۱۹۹۹) می‌توانند برای توصیف ریخت‌شناسی شبکه‌های هم نویسندگی به کار برده شوند. پیکربندی جهان کوچک، حضور همزمان خوشه‌بندی محلی متراکم[6] با فواصل کوتاه شبکه که می‌توانند جریان دانش درون شبکه را تسهیل نمایند را توصیف می‌کند. این بدین معناست که در یک شبکه هم نویسندگی، گروه‌های منسجم کوچکی از پژوهشگران با ارتباط‌های اندکی بین آنها، وجود داد. با نگاهی به رتبه توزیعی عامل‌ها - یعنی فراوانی توزیع تعداد هم نویسنده‌ها به نویسنده- اگر توزیع قانون توان[7] مشاهده شود، بعد از آن یک ساختار بدون مقیاس در شبکه پدیدار می‌گردد. این روش به وجود سازوکار شکل‌گیری یک گره ویژه که وابستگی امتیازی[8] نامیده می‌شود، اشاره دارد که به‌طور قراردادی برای تمایل به تعامل با بهترین نویسندگان متصل به کار می‌رود. شاخص مرکزیت مبتنی بر عامل، به جایگاه هر نویسنده در شبکه بر اساس تعاریف مختلف مرکزیت (رتبه، نزدیکی، بردار ویژه و غیره) اشاره دارد. استفاده از داده‌های دودویی یا وزن‌دهی شده بر ارزش شاخص‌های سطح عامل در شبکه تأثیر می‌گذارد. ماتریس

1 Actor-Level
2 Small World
3 Watts And Strogatz
4 Scale Free
5 Barabasi And Albert
6 Dense Local Clustering
7 Power Law Distribution
8 Preferential Attachment

مجاورتی وزن‌دهی شده مناسب، می‌تواند برخی اطلاعات اضافی درباره روابط هم نویسندگی را آشکار سازد (استفانو، گیوردانو و ویتیل،۲۰۱۱).

هم واژگانی و هم رخدادی

بسامد وقوع واژه‌ها مقیاسی مهم در تحلیل محتوا است. این سنجه برای تعیین مهم‌ترین موضوع‌های پژوهشی در یک حوزه با تمرکز بر روی واژه‌های پر بسامد مورد استفاده قرار می‌گیرد؛ یعنی فراوانی یک واژه به عنوان شاخصی از اهمیت، توجه، یا تأکید بر آن واژه یا اندیشه در نظر گرفته می‌شود، یا مفهوم به آن مرتبط است. دو مشکل اساسی در مورد این نوع تحلیل عبارت است از: کمبود متن و امکان ناچیز شماری بسامد، به سبب استفاده از مترادف‌ها. هرچند راه‌حلی که به‌طور گسترده برای این مشکلات در دسترس باشد، مخصوصاً در حجم بالای داده‌ها وجود ندارد. مشکل اول با به کار بردن فنون پیچیده‌تری که محتوا را در نظر بگیرد قابل حل است. درحالی که مشکل مترادف‌ها، سخت است که به‌طور کامل حل گردد.

در هر حوزه‌ای از علم و فناوری مجموعه‌ای از مفاهیم وجود دارد که ساختار دانش آن حوزه را می‌سازند. این مفاهیم با الفاظی که برای دلالت بر آنها وضع می‌شود نام‌گذاری می‌شوند. مجموعه این الفاظ اصطلاحات و مفاهیم حوزه‌های علم و فناوری را تشکیل می‌دهند (رفیعی خضری، ۱۳۸۷). مفاهیم از داده‌ها و از درک نظری قبلی پژوهشگران در یک حوزه مورد مطالعه به‌دست می‌آیند که به اولی استنتاجی و به دومی قیاسی می‌گویند. یک موضوع قیاسی از این موارد حاصل می‌شود: خواص عناصر مورد مطالعه، آنچه که در تعاریف حرفه‌ای در پیشینه ادبیات مربوط در آن توافق شده است، برساخت‌های محلی و عوام، یافته‌های پژوهشگران ، مباحث نظری و تجربه‌های شخصی. کوربین و اشتراوس[1]، به آن حساسیت نظری می‌گویند. تصمیم پژوهشگران درباره اینکه به چه موضوعاتی بپردازند و چگونه مباحث اطلاعاتی را درمورد آن موضوعات بیان دارند، یک منبع غنی از مفاهیم قیاسی است. در حقیقت، اولین مرحله از تعمیم مفاهیم غالباً از سؤالاتی برمی‌آید که از یک پروتکل مصاحبه‌ای حاصل شده است. برخلاف پیشینه پژوهشگران این مفاهیم تاحدودی تجربی هستند (رایان و برنارد، ۲۰۰۳).

1. Strauss & Corbin

بسیاری از مفاهیم از داده‌های استنتاجی حاصل می‌شوند که محصول تحلیل متون، تصاویر، اصوات و حتی سؤالات ثابت است. کسی نمی‌تواند همه مفاهیم را قبل از تحلیل داده‌ها پیش‌بینی کند. عمل کشف مفاهیم آن چیزی است که در روش نظریه دانش بنیاد، به آن کدگذاری باز و همچنین تحلیل‌گران محتوای کلاسیک مانند برلسون به آن تحلیل کیفی یا کدگذاری نهفته می‌گویند (رایان و برنارد، ۲۰۰۳).

کشف مفاهیم و روابط میان آنها از طریق ارتباط واژگانی در اسناد و مدارک، زمینه ایجاد نقشه علمی را در حوزه‌های علمی فراهم می‌کند (احمدی؛ سلیمی و زنگیشه، ۱۳۹۲). زمانی که نقشه‌دانش در مورد یک رشته علمی ترسیم می‌شود، یک نقشه علم حاصل می‌شود. نقشه علم زیرحوزه‌های هر زمینه علمی و میزان دانش موجود در هر زیرحوزه و نیز ارتباط و تعامل زیر حوزه‌های مختلف با یکدیگر را مشخص می کند (ناصری جزه و همکاران، ۱۳۹۱).

ترسیم نقشه‌های علم، از زیرشاخه‌های حوزه علم‌سنجی است. این، حوزه، از طریق پردازش، استخراج و مرتب‌سازی اطلاعات به ترسیم نقشه‌دانش می‌پردازد و امکان تحلیل، مسیریابی و نمایش دانش را فراهم می‌آورد؛ علاوه بر آن، این حوزه در جهت سهولت بخشیدن دسترسی به اطلاعات، آشکارسازی ساختار دانش و کمک به جستجوگران دانش برای رسیدن به نتایج موفقیت‌آمیز حرکت می‌کند (نوروزی چاکلی، ۱۳۹۰). استفاده از روش‌های علم سنجی برای کشف روابط میان مفاهیم ، تاکنون در بسیاری زمینه‌ها به منظور ترسیم شبکه مفهومی آن زمینه استفاده شده است که در قسمت پیشینه به مرور برخی از این منابع پرداخته می‌شود.

مفهوم ترسیم نقشه‌های مبتنی بر مدارک و اسناد، با مطالعات دانشمندان هلندی به خصوص با مطالعات نیونز و ران[1] توسعه یافت. نیونز و ران، متدولوژی جدیدی برای ترسیم این نقشه‌ها به وجود آوردند. فرض اولیه آن‌ها بر این مبنا بود که هر زمینه تحقیقاتی با مجموعه‌ای از کلیدواژه‌ها شناخته می‌شود، هر مدرک منتشر شده در آن حوزه نیز، با زیر مجموعه‌ای از این کلیدواژه‌های اولیه شناخته می‌شود. این زیر مجموعه‌ها شبیه اثر انگشتهای DNA در مورد آن مقاله خاص است با مقایسه اثر انگشتهای DNA در مورد دو مدرک منتشر شده می‌توانیم به شباهت‌های مدارک منتشره پی ببریم. هرچه دو مدرک دارای کلیدواژه‌های مشترک بیشتری باشند، شباهت بیشتری به هم دارند، و در این صورت، به احتمال بیشتری، از یک حوزهٔ تحقیقاتی ناشی شده‌اند. با ادامه دادن استعاره DNA می‌توان تصور کرد در صورتی که شباهت‌ها از یک سطح بیشتر باشد دو مدرک منتشر شده

[1] . Noyons & Raan

متعلق به یک گونه از تحقیقات هستند. آنها این نقشه را توسعه دادند تا بتوانند در حوزه‌های خاص میزان تأثیر زیر گروه‌های موضوعی خاص را بسنجند، و بتوانند به این سؤال اساسی پاسخ دهند که در هر زمینه موضوعی خاص، زیر موضوع‌ها در کدام قسمت قرار گرفته‌اند (نقل در عابد جعفری؛ ابویی اردکان و آقازاده دهده، ۱۳۸۹).

نقشه دانش یک رشته را می‌توان با روش "تحلیل هم استنادی" (از روش‌های مطالعات کتابسنجی) و یا روش "تحلیل هم‌رخدادی کلمات" (از روش‌های مطالعات علم سنجی) ترسیم کرد که امروزه پرکاربردترین روش، تحلیل هم‌رخدادی کلمات است (ناصری‌جزه و همکاران، ۱۳۹۱).

ایده «تحلیل هم‌رخدادی کلیدواژه‌ها» در سال ۱۹۸۳ توسط کالون[1] مطرح شده است. ایده وی، این بود که آمدن کلمات با هم در یک مدرک، نشان‌دهنده محتوای آن مدرک است. لذا اگر میزان این هم‌رخدادی را اندازه‌گیری کنیم، می‌توانیم شبکه مفاهیم یک زمینه علمی را ترسیم کنیم (ناصری جزه و همکاران، ۱۳۹۱؛ الهی و همکاران، ۱۳۹۱؛ مهدی‌زاده مرقی و همکاران، ۱۳۹۲). هرچه فراوانی هم‌رخدادی دو کلیدواژه بالا نیز باشد، نشان‌دهنده این است که ارتباط بین آن دو محکمتر و نزدیکتر است. به‌عبارتی، هر چه ارتباط بین دو کلیدواژه نزدیکتر باشد، ارتباط نزدیکتری بین مفاهیمی که به آن اشاره دارند، وجود دارد (وانگ و اینابا، ۲۰۰۹). در دو دهه گذشته، این فن توسط گروههای تحقیقاتی مختلف اجرا شده و به عنوان ابزاری قدرتمند برای کشف‌دانش اثبات شده است (بیون[2]، ۱۹۸۶) و یکی از فنون تحلیل محتوا است که از الگوهای هم‌رخدادی (به عنوان مثال واژه یا عبارت اسمی)در یک مجموعه از متون بهره می‌گیرد تا ارتباط میان اندیشه‌ها در حوزه موضوعات را شناسایی کند (هی[3]، ۱۹۹۹). بنابر توضیحات بالا، تجزیه و تحلیل هم‌واژگانی، برای کشف ارتباط میان موضوعات در حوزه‌های تحقیقی و در نهایت ردیابی توسعه علم از طریق فراوانی هم‌رخدادی دو واژه یا عبارت است. این روش، از روشهای کمی کشف ساختار دانش می‌باشد (بیون، ۱۹۸۶) و بر این فرض استوار است که کلیدواژه‌های یک مدرک وسیله مناسبی برای توصیف محتوای آن هستند. به عبارتی، ارتباط دو واژه کلیدی که درون یک مدرک قرار می‌گیرند، از ارتباط بین موضوعاتی نشان دارد که آن مقاله به آن‌ها می‌پردازد. این فن، توانایی آن را دارد که از طریق اندازه‌گیری میزان ارتباط اصطلاحات درون مدارک منتشره در یک حوزه

[1] . Callon
[2] . Bauin
[3] . He

خاص، الگوها و رویکردهایی را در آن حوزه نشان دهد (وانگ و اینابا، ۲۰۰۹) . کیم[1] و همکاران (۲۰۰۹)، معتقد هستند استخراج عبارت‌های کلیدی نقش اساسی در کشف پیشرفت‌ها و گرایشهای تکنولوژی دارد.

بنا به گفته وانگ و اینابا (۲۰۰۹)، مراحل مشترک در تجزیه و تحلیل هم‌واژگانی دارای چهار بخش است: گام‌اول، جمع‌آوری داده‌هاست. کلمات، مهمترین عناصر تجزیه و تحلیل هم‌واژگانی هستند. دو راه برای استخراج کلمات کلیدی از مقالات مجلات، مقالات کنفرانس‌ها و یا گزارش‌های فنی وجود دارد. یک روش عناوین یا لیست کلیدواژه‌ها و روش دیگر استفاده از متن کامل است. درهرصورت، تنها کلمات و عباراتی که فراوانی مناسبی دارند، و بر موضوعات اصلی در حوزه علمی دلالت دارند، به منظور تحلیل واژگانی انتخاب می‌شوند.

گام‌دوم، استانداردسازی داده‌هاست. مفاهیم مشابه زیادی با عبارات یا کلمات مختلف (مترادفات) در یک مجموعه وجود دارد . برای استانداردسازی این کلمات، محققان باید مترادفات، متضادها، ایهام‌ها، اصطلاحات عام و اصطلاحات خاص را در نظر بگیرند. مانند دانش، نظریه‌ها، تاثیر، پروژه‌ها، توسعه، کاربردها، تولید، پیاده‌سازی، تعریف و غیره.

گام‌سوم، ساخت ماتریس است. هنگامی که موضوعات تحقیق انتخاب شدند، یک ماتریس بر اساس هم‌رخدادی کلمات ایجاد می‌شود. هر چقدر که فراوانی رخداد دو واژه بیشتر باشد، به همان اندازه ارتباط میان آنها نزدیک‌تر است.

گام‌چهارم، تجزیه و تحلیل داده‌ها و ترسیم نقشه است. معروف‌ترین روش، روش پیمایش چند بعدی[2] است. در این روش، نقاطی که یک فضای دارای ابعاد بسیار زیاد را تشکیل می‌دهند، در یک فضای دو یا سه بعدی به نمایش در می‌آیند. این‌کار، از طریق سنجش فاصله هر نقطه با نقاط دیگر(سنجش جفتی) صورت می‌گیرد و در حد امکان فاصله آن نقاطی که در اصل دارای ابعاد بسیار زیاد هستند ، به صورت دقیق محاسبه می‌شود.

بعد از جنگ جهانی دوم، دامنه و حجم پژوهش‌های علمی رشد چشمگیری یافت. پرایس (۱۹۶۳) تخمین زده بود که متون علمی هر ۱۰ سال دو برابر می‌شوند. سه دهه بعد به لطف پیشرفت‌های فناوری اطلاعات، به خصوص در حوزه ذخیره داده‌ها، حجم اطلاعات هر ۲۰ ماه دو برابر شده بود (فراولی و

1 . Kim
2 . Multidimensional scaling(MDS)

دیگران[1]، ۱۹۹۱). در چنین شرایطی ردیابی حوزه‌های موضوعی و ارتباط میان حوزه‌ها برای دانشمندان دشوار بود. سیاست‌گذاران علم برای ترسیم حرکت علم و برنامه‌ریزی‌های آتی با مشکل مواجه بودند. روش سنتی برای ترسیم روابط میان مفاهیم، اندیشه‌ها و مسائل علمی، جستجوی نظرات جمع کوچکی از متخصصان یک حوزه است و در برخی مقاصد لازم‌الاجرا است؛ ولی نقاط ضعفی نیز بر آن وارد است. این روش‌ها بسیار گران، زمان‌بر و در عین حال محدودند. از سوی دیگر آیا می‌توان گروه کوچکی را نماینده یک جامعه بزرگ دانست (لاو[2] و وایتاکر، ۱۹۹۲). به همین دلیل فنون کمی ترسیم ساختار علم مورد توجه قرار گرفت.

تحلیل‌های هم واژگانی[3]، یک فن تحلیل محتوا است که الگوی هم رخداده جفت واژه‌ها یا عبارت‌های درون مجموعه‌ای از متون را جستجو می‌کند تا به روابط میان ایده‌ها در حوزه‌های موضوعی دست یابد (هیی[4]، ۱۹۹۹). تحلیل هم واژگانی بر این فرض استوار است که حوزه‌های پژوهشی را می‌توان بر اساس الگوهای به‌کارگیری واژگان در انتشارات توصیف کرد (نف و کورلی[5]، ۲۰۰۹). این فن ابزار قدرتمندی در کشف دانش و ترسیم نقشه کتابشناختی است. تحلیل‌های هم واژگانی، روشی پرکاربرد برای به دست آوردن ساختارهای سطح بالا از الگوهای رخداد- واژه در متن است. این نوع تحلیل ابزاری برای توضیح ساختارهای عقاید، مشکلات و غیره است که در مجموعه‌های مناسب از اسناد ارائه شده است. تحلیل هم واژگانی بر پایه تحلیل فراوانی هم‌رخداده‌ای‌های[6] کلیدواژه‌های استخراج شده از عنوان‌ها، چکیده‌ها یا متن‌ها به صورت کلی هستند. این روش در دهه ۱۹۸۰ توسط کالون، کورتیال و ترنر[7] توسعه پیدا کرد. این روش اغلب به عنوان جایگزینی برای رویکردهای استنادی و هم‌استنادی برای ترسیم ساختار علم مورد استفاده قرار می‌گیرد.

در این روش برای اندازه‌گیری میزان ارتباط بین موجودیت‌ها نمایه‌هایی بر اساس بسامد هم رخدادی ساخته می‌شود. بر اساس این نمایه‌ها، موجودیت‌ها (واژه‌ها- عبارت‌ها) در گروه‌هایی خوشه‌بندی شده و به شکل شبکه نمایش داده می‌شوند. با مقایسه نقشه‌های حاصل در دوره‌های زمانی مختلف، پویایی علم ردیابی می‌شود (هیی، ۱۹۹۹). این فن با تحلیل مجموعه‌ای از مدارک به ارزیابی میزان ارتباط آنها می‌پردازد. در این تحلیل واحد اندازه‌گیری تعداد واژه‌ها یا عبارت مهم و مشترک در مجموعه است (ریپ و کورتایل[8]، ۱۹۸۴). در تحلیل هم واژگانی، هم رخدادی کلیدواژه‌ها در عنوان، چکیده یا متن مقاله‌ها بررسی می‌شود. هم رخدادی کلیدواژه‌ها میزان ارتباط شناختی میان یک مجموعه مدارک را نشان می‌دهد. تفاوت مهمی که میان هم واژگانی

1 Frawley
2 Whittaker & Law
3 Co-Word Analysis
4 He
5 Neff & Corley
6 Co-Occurrence
7 Callon,Courtial And Turner
8 Rip

و هم استنادی وجود دارد این است که در یک دوره پژوهشی مشخص، تحلیل هم استنادی به منابع استناد دهنده (مقاله استناد دهنده، مؤلفان استناد دهنده) و مأخذ استناد شده (مؤلف استناد شده، مدرک استناد شده) نیازمند است؛ اما تحلیل هم واژگانی فقط نیازمند مجموعه‌ای از مقاله‌های مجله‌ها در یک حوزه موضوعی خاص است. داده‌های حاصل از این فن توسط ماتریس بسامد سنجیده می‌شود و می‌توان نتایج حاصل را در خوشه‌های سلسله مراتبی با مقیاس چند بعدی به نمایش گذاشت.

تحلیل بسامد واژگان عناوین مقالات، ابتدا توسط اسمال و همکارانش به کار گرفته شد تا موضوعیت یا به‌طور خاص اجماع مفهومی مدرک استناد شده را مشخص کنند (گریفیث و اسمال، ۱۹۷۴). البته انتظار نمی-رفت این پروفایل واژگانی محتوای مدارک درون یک خوشه هم استنادی را تعیین کند؛ اما حضور مکرر یک عبارت و یا یک واژه در عناوین مقالات استناد دهنده به مقاله استناد شده را باید معنادار دانست (اسمال و کرین، ۱۹۷۹). واژگان نمایه‌سازی، کدهای سازمان‌دهی، واژگان عنوان و چکیده مقالات نشانه‌های ساختاری یک مقاله علمی هستند که از دو نظر با هم مشابه هستند: اول اینکه، آن عبارت و واژه‌ها مرتبط با محتوای مقاله علمی هستند و آنها ابزار خلاصه‌سازی، چکیده‌نویسی و یا طبقه‌بندی موضوع‌های مقاله‌اند. دوم این واژه-ها نشانگرهای شناسایی و متمایز کردن هر مقاله و نه مؤلف آن است؛ بنابراین هر یک از این کلمات می‌توانند برای توصیف عناوین پژوهشی مناسب باشند و مشابهت میان مجموعه مقالات در خوشه‌های هم استنادی را نمایان سازد.

برام و همکاران (۱۹۹۸) تحلیل واژه را برای تشخیص موضوعات خاص پژوهشی مفید می‌دانند و معتقدند تحلیل واژگان مقالاتی که در خوشه‌های هم استنادی قرار دارند کمک می‌کند تا قابلیت تحلیل استنادی جامعیت این تکنیک خوشه‌بندی آزموده شود. زیربنای به کارگیری این راهکار ترکیبی این نکته اصلی است که یک رشته اختصاصی علمی از ارتباط منطقی مجموعه‌ای از مسائل پژوهشی، سؤالات و مفاهیم حاصل می‌شود و عده‌ای از پژوهشگران صرف نظر از موقعیت اجتماعی یا فکری‌شان درگیر مسائل آن رشته می‌شوند. اگر پژوهشگران مختلف بر روی موضوعات پژوهشی و مفاهیم یکسانی تمرکز و مطالعه کنند، این انتظار به وجود می‌آید که در طیفی گسترده واژگان مشابهی را برای مفاهیم مهم و یا مسائل حوزه تخصصی خود به کار برند؛ و همچنین همین تعداد پژوهشگر، هم زمان مجموعه محدود و کوچکی از مقالات پیشین خود را مورد استناد قرار دهند، چرا که مقالات حامل، هم زمان واژه‌ها و مأخذ مشابه یا همگرایی بین نوشته های قبلی را نشان می‌دهد (برام و دیگران، ۱۹۹۱).

هم واژگانی به عنوان شاخصی در علم‌سنجی در دهه ۸۰ معرفی شد (کالون و دیگران، ۱۹۸۲). در حالی که نقشه‌های هم استنادی نمایانگر درک و آگاهی استناد دهندگان بود، تحلیل هم واژگانی شبکه‌ای معنایی میان

استناد دهندگان را آشکار ساخت. پژوهشگران حوزه علم‌سنجی دریافته‌اند که استنادها موجودیت‌های ویژه‌ای هستند، اما با شاخص‌های متنی اشتراکاتی دارند. هر حوزه پژوهشی با مجموعه‌ای از کلیدواژه‌های مهم یا ترکیب آن واژه‌ها قابل شناسایی است و مفاهیم آن حوزه را می‌توان در قالب چندین کلیدواژه سیاهه کرد.

فن هم واژگانی برای اولین بار در دهه ۱۹۸۰ در فرانسه در مرکز جامعه‌شناسی خلاقیت کول[1] به‌طور جدی به کار گرفته شد. آنها سیستم طراحی شده بر اساس این فن را LEXIMAPPE نامیدند. نگارش کتاب "ترسیم پویایی علم و فناوری[2]" نقطه عطف تحلیل هم واژگانی بود. کالون، لاو و ریپ (۱۹۸۶) با انتشار این کتاب تحلیل هم واژگانی را از فرانسه به سایر کشورها بردند. در این کتاب کالون بر نقش قدرتمند متون تأکید می‌کند. او معتقد است که بهترین راه برای درک پویایی علم بررسی "شبکه نقش‌آفرین[3]" است. آنها در آزمایشگاه‌ها دانش نوین و پیچیده را خلق و سپس از طریق متون آن را منتقل می‌کنند (لاتور[4]، ۱۹۸۷). متون علمی اگر تنها راه نباشد، ولی مهم‌ترین راه انتشار اندیشه‌ها و ارتباط با سایر دانشمندان است؛ بنابراین مطالعه متون می‌تواند فرآیند انتقال اندیشه‌ها از آزمایشگاه‌ها را به جامعه نشان دهد (هی،۱۹۹۹)؛ بنابراین به جای ردگیری نقش‌آفرینان تغییر جهانی علم می‌تواند متون مکتوب آنها را ردگیری نماید. تحلیل هم واژگانی پویایی علم را نتیجه تغییر راهبرد نقش‌آفرین‌ها می‌داند؛ زیرا تغییرات یک حوزه موضوعی تابع راهبردهای افراد فعال در آن حوزه است (کالون و دیگران، ۱۹۹۱).

گام اول در تحلیل هم واژگانی استخراج کلیدواژه‌ها از متون است. پس از استخراج کلیدواژه‌ها از مدارک یک ماتریس هم رخدادی ساخته می‌شود. مهم‌ترین بخش تحلیل هم واژگانی، تحلیل شاخص‌های این ماتریس است. بسته به نوع سؤالی که درباره شبکه علم مطرح باشد، ماتریس هم رخدادی عملکردهای متفاوتی دارد. در اغلب موارد دو درخواست مورد نظر است: ردیابی سلسله مراتبی در میان حوزه‌های یک مسئله پژوهشی و ردیابی حوزه‌های کوچک، اما مستعد رشد. برای پاسخ به این دو پرسش دو شاخص معرفی شده است. شاخص شمول[5] و شاخص مجاورت[6] (کالون و دیگران، ۱۹۸۶). کالون برای نشان دادن سلسله مراتب موجود در حوزه‌های موضوعی از شاخص شمول استفاده کرد که با استفاده از فرمول زیر محاسبه می گردد.

$$I_{ij=} C_{ij/} \min (C_i, C_j) \quad (۱)$$

هرگاه:

1 Center De Sociologie De La Innovation Of The Ecole
2 Mapping The Dynamic Of Science And Technology
3 Actor Network
4 Latour
5 Inclusion Index
6 Proximity Index

C_{ij}= تعداد مدارکی که کلیدواژه‌های M_i و M_j با هم در آنها وجود داشته باشند

C_j= بسامد رخداد کلیدواژه M_j در مجموعه‌ای از مقالات

C_i= بسامد رخداد کلیدواژه M_i در مجموعه‌ای از مقالات

$\min(C_i, C_j)$= حداقل دو بسامد C_j و C_i

ارزش عدد I_{ij} بین صفر و یک است و احتمال مشروط را نشان می‌دهد. وقتی $C_i > C_j$ باشد، M_i کلی‌تر از M_j و گاه در برگیرنده M_j است. I_{ij} احتمال یافتن M_i در مقاله‌ای است که M_j در آن حضور دارد. در نتیجه وقتی $I_{ij}=1$ است، M_j تماماً شامل M_i است یعنی دو واژه در یک مقاله هم رخداد هستند.

در صورتی که I_{ij} ارزش پایینی را نشان دهد، معرف آن است که واژه‌های میانجی وجود دارند که هم رخداد بالایی ندارند اما از نظر مفهومی دارای ارتباط هستند. برای بیرون کشیدن چنین الگوهایی از شاخص مجاورت استفاده می‌شود که با استفاده از فرمول زیر محاسبه می‌شود.

$$P_{ij=} (C_{ij/} C_iC_j).N$$

C_iو C_j و C_{ij} همان معنای فرمول اول را دارند و N معرف تعداد مقالات مجموعه است. واژه‌های میانجی که شاخص مجاورت را نشان می‌دهند، معرف حوزه‌های کوچک‌تر اما مستعد رشد هستند.

در مطالعات بعدی شاخص دیگری معرفی شد که همبستگی ارزش میان دو جفت واژه را محاسبه می-کرد. کالون و همکارانش (۱۹۹۱) این شاخص را شاخص هم ارز[1] و کولتر و همکارانش (۱۹۹۸) آن را شاخص استحکام[2] نامیدند. با توجه به فرمول ۱ فرمول شاخص جدید را به این صورت محاسبه می‌شد:

$$E_{ij=} (C_{ij} / C_i). (C_{ij} / C_j)_=(C_{ij})^2/(C_i.C_j)$$

ارزش Eij بین صفر و یک است. به مثابه فرمول اول Eij احتمال حضور همزمان i در مجموعه مدارکی که با واژه j نمایه شده‌اند را نشان می‌دهد و برعکس. به همین دلیل ترنر و همکارانش (۱۹۹۹) Eij را "عامل مشترک شمول دوطرفه " نامیدند. پس از محاسبه شاخص شمول و مجاورت، نقشه‌های شمول و مجاورت رسم می‌شود. نقشه‌های شمول موضوعات مرکزی یک حوزه و روابط آنها را با کلیدواژه‌های که بسامد رخداد کمتری دارند را نشان می‌دهد. نقشه مجاورت ارتباط میان اندیشه‌های کوچک‌تر یا پنهان در اطراف نقاط مرکزی را نمایش می‌دهد. برای ترسیم نقشه شمول، یالی که بالاترین ارزش را دارد اول انتخاب می‌شد و سپس یال‌های کم‌ارزش‌تر رسم می‌شوند تا به آستانه I برسیم.

1 Equivalence Index
2 Strength Index

کلیدواژه‌هایی که در بالای نقشه قرار می‌گیرند قطب‌های مرکزی یک حوزه پژوهشی را نشان می‌دهند. کلیدواژه‌هایی که هم در موقعیت قطب مرکزی هستند و هم واژه‌های وابسته دارند، واژه‌های میانجی نامیده می‌شوند (کالون و دیگران، ۱۹۸۶).

نحوه ترسیم نقشه مجاورت همانند روند قبلی است. در نقشه‌های مجاورت هر چه آستانه P پایین‌تر باشد، مجاورت بیشتری میان دو کلیدواژه در نقشه ظاهر خواهد شد؛ بنابراین قطب‌های مرکزی و میانجی با هم نمایش داده می‌شوند. در نتیجه رابطه موضوعات مرکزی و پیرامونی قابل مطالعه است. این دو نقطه دو نوع مطالعه تحلیلی را ممکن می‌سازد. نوع اول مطالعه مبسوط یک موضوع خاص و نوع دوم ارتباط میان موضوعات مختلف.

کالون و دیگران (۱۹۸۶) امتیاز تحلیل هم واژگانی نسبت به روش‌های کیفی این گونه عنوان می‌کنند: مسئله تشخیص موفقیت یا عدم موفقیت تحلیل کیفی در تفسیر نتایج را می‌توان با راهکار کمی سنجید. روش‌های کیفی اغلب از تحلیل جزئیات مباحث علمی به سرعت می‌گذرند و به توصیفات عمومی می‌پردازند. برعکس تحلیل هم واژگانی، با خلاصه‌سازی مقالات در واژه‌های قدرتمند و محاسبه رخداد و هم رخدادی تشخیص دقیق‌تری نسبت به حوزه موضوعی ارائه می‌دهد.

تحلیل هم واژگانی شبکه پژوهشی را از طریق نگاشت ارائه می‌دهد. این نگاشت‌ها می‌توانند در ساده‌ترین صورت ساختار شبکه ارائه دهند. از سوی دیگر فرد می‌تواند بر نواحی خاص تمرکز کرده و الگوهای هم واژگانی را با جزئیات دلخواه ردگیری کند.

کالون (۱۹۸۶) با به کار گیری تحلیل هم واژگانی در مطالعه مجموعه‌ای از پروانه ثبت اختراعات به قابلیت انعطاف این روش پی برد. او از دو فن برای تحلیل نقشه استفاده کرد. فن اول ساده‌سازی نقشه‌ها بود. برخی از واژه‌ها در سراسر حوزه حضور می‌یابند بدون آنکه اطلاعات جدیدی اضافه کنند. با حذف آنها، ساختار نقشه‌ها ساده‌تر می‌شود اما تغییر نمی‌کند. فن دوم تمرکز روی یک قطب است. مثلاً چرا واژه‌ای که در سال ۱۹۸۰ یک قطب مرکزی بوده است (آنزیم) سال بعد به‌طور کلی از نقشه شمول محو می‌شود. این فن تمرکز نشان داد که این تغییر ناگهانی میان دو دوره زمانی چیزی بیش از یک نوسان ساده است. در واقع ظهور تعداد محدودی پروانه ثبت اختراع مرکز توجه جدیدی را ایجاد کرده و در نتیجه روابط جدیدی شکل گرفته است. روش دیگری هم که به انعطاف‌پذیری کمک کرده سطح‌بندی خوشه‌هاست که امکان می‌دهد نقشه نواحی پژوهشی در سطوح مختلف قابل ارائه باشد.

کیفیت نتایج حاصل از تحلیل هم واژگانی بستگی به عوامل زیادی دارد. کیفیت کلیدواژه‌ها و واژه‌های نمایه‌ای، دامنه پایگاه اطلاعاتی، قابلیت روش‌های آماری و ارائه یافته‌ها مهم‌ترین مسئله در این روش انتخاب

کلیدواژه است که از آن به "تأثیر نمایه‌ساز " یاد شده است و اغلب پژوهشگران بر آن تأکید کرده‌اند. نمایه‌سازی مداخله تحلیلگر با متن است و اعتبار نقشه بستگی به ماهیت نمایه‌سازی دارد.

وایتاکر[1] (۱۹۸۸) معتقد است که نتایج تحلیل هم واژگانی بستگی به نحوه انتخاب کلیدواژه‌ها توسط نمایه‌ساز دارد تا از طریق آنها حوزه‌های علمی را مفهوم‌سازی کند؛ اما نتایج نمایه‌سازی خود بستگی به درک نمایه‌ساز دارد تا خود دانشمندان که آثارشان نمایه می‌شود.

علاوه بر تأثیر نمایه‌ساز، انتقادات دیگری نیز بر این تحلیل وارد شده است:

- برخی از بازنمایی‌های حاصل از هم واژگانی قابل خواندن نیست (وایتکر، ۱۹۸۹)؛
- پوشش پایگاه داده‌ها اغلب کامل نیست. انواع خاصی از متون مثلاً پروانه ثبت اختراعات در اغلب پایگاه‌ها نمایه نمی‌شوند؛ بنابراین نتایج حاصل از این تحلیل را نمی‌توان تصویر کاملی از کل رشته پژوهشی دانست؛
- فاصله زمانی میان نگارش مقاله و زمان نمایه‌سازی و ورود به پایگاه داده، امکان ردگیری موضوع در مراحل اولیه را از بین می‌برد.

تحلیل هم واژگانی بر اساس چند پیش فرض استوار است (ویتاکر، ۱۹۸۹).

- مؤلفان مقالات علمی واژه‌های فنی خود را آگاهانه و با دقت انتخاب می‌کنند.
- وقتی واژه‌های مختلف در یک مقاله واحد به کار گرفته می‌شود به این دلیل است که مؤلف برخی روابط مهم میان آنها را درک کرده و یا مورد قیاس قرار داده است.
- در صورتی که تعداد زیادی از مؤلفان نیز روابط مشابهی را (که در مورد قبل آمد) درک کنند، بنابراین این روابط معرف معنای ویژه‌ای در یک حوزه موضوعی است (همان)
- کلیدواژه‌های انتخابی توسط نمایه‌سازان به عنوان توصیفگر مفاد مقالات، در حقیقت نماینده قابل اعتمادی از مفاهیم علمی مقالات است و می‌توان از آنها به عنوان کلیدواژه‌ی تحلیل هم واژگانی استفاده کرد (هی، ۱۹۹۹).

اگر پیش فرض‌ها را منطقی بدانیم آنگاه می‌توان از بسامد جفت واژه‌های موجود در یک مجموعه مقاله برای ترسیم ساختار مفاهیم موجود در مقالات استفاده کرد.

همان‌طوری که اشاره شد تأثیر نمایه‌ساز یکی از مشکلات جدی تحلیل هم واژگانی است و انتقاد اصلی بر این تحلیل است. یکی از نتایج تأثیر نمایه‌ساز این است که کلیدواژه‌هایی که نمایه‌ساز برای مقالات انتخاب می‌کند روزآمد نیستند، این امر به سه دلیل پیش می‌آید:

1 Whittaker

- اصطلاح‌نامه‌ای که مورد استفاده قرار می‌گیرد روزآمد نیست.
- نمایه‌ساز ممکن است از ترکیب کلیدواژه‌هایی استفاده کند که مد نظر وی است.
- فاصله زمانی انتشار مقاله تا حضور در پایگاه داده اجتناب ناپذیر است (وایتاکر، ۱۹۸۹).

در سال ۱۹۹۲، لاو و وایتاکر طی مصاحبه‌ای با برخی از متخصصان، کلیدواژه‌های منتخب نمایه‌سازی پایگاه PSCL را مورد بررسی قرار دادند. اگر چه اغلب نظرات درباره کیفیت انتخاب مثبت بود، اما سه مورد اعتراض نیز وجود داشت. ۱: برخی کلیدواژه‌های نمایه‌سازان بسیار کلی است. ۲: یک یا دو واژه تخصصی از لیست حذف شده و ۳: در سطح‌بندی تخصصی واژه‌ها چند اشتباه رخ داده بود. برای فائق آمدن بر تأثیر نمایه‌ساز چند سازوکار برای نمایه‌سازی خودکار به کار گرفته شده است.

کؤئست پلاس[1]: این آزمون یک مقام بازیابی اطلاعات متنی است که در فرانسه توسط TELESYSTEMES به کار گرفته شده است. کالون و همکارانش نیز در این طرح مشارکت داشتند. آنها فنون مختلفی را با کؤئست پلاس ترکیب کردند و همزمان LEXIMAPPE را نیز به کار گرفتند. حاصل این همکاری، فرآیند زنجیره‌ای از متن کامل به نقشه‌های خودکار شمول یا مجاورت بود. در مقایسه با فایل نمایه‌سازی دستی، نتایج حاصل از این فرآیند جنبه‌های مثبتی در پی داشت. ابتدا، واژه‌های کلی و زائد- که نقشه را پیچیده کرده اما اطلاعات جدیدی ارائه نمی‌کرد- کاهش یافتند. سپس، تعداد بیشتری از موضوعات خاص و پیرامونی در نقشه‌های شمول ظاهر شدند. سوم ساختار نقشه‌های مجاورتی غنی‌تر و جزءنگرتر شد.

LEXINET: ترنر و دیگران (۱۹۸۸) آزمونی را برای فائق آمدن بر نمایه‌سازی دستی از طریق رایانه انجام دادند که LEXINET نام داشت. هدف سیستم LEXINET، کمک به متخصص برای ساخت نمایه لغات مناسب برای یک حوزه خاص بود. این کار از طریق ارزیابی تعاملی میان ماشین و متخصص انجام می‌شد. مطالعات آنها نشان داد فرآیند نمایه‌سازی توسط LEXINET شتاب گرفت. در حقیقت این سیستم فاصله زمانی انتشار تا زمان نمایه شده مقالات را به‌طور قابل ملاحظه‌ای کاهش داد. در عین حال کیفیت اطلاعات موجود با کنترل تأثیر نمایه‌ساز ارتقاء یافت.

نقشه‌های حاصل از تحلیل هم واژگانی اغلب پیچیده‌اند؛ بنابراین تفسیر آنها نیاز به دقت فراوان دارد. کالون (۱۹۸۶) معتقد است نقشه‌های شمول و مجاورت باید با هم تحلیل شوند. در برخی موارد باید از نظرات متخصصان برای تحلیل سود جست. این نقشه‌ها تنها یک عکس از دانش یک حوزه نیست و نباید با آن برخورد آماری کرد. بلکه باید به شکل پویا به روابط درونی شبکه و روابط میان شبکه‌ها نگریست.

1 Questel-Plus

در سال ۱۹۸۷، لیدسدورف به تأثیر نمایه‌ساز در تحلیل‌های هم واژگانی انتقاد کرد و پیشنهاد داد که برای رفع این اشکال باید از واژه‌های عنوان برای تحلیل هم واژگانی استفاده کرد. این روش دسترسی مستقیم به نظرات مؤلفان را امکان‌پذیر کرده و توصیفگرها از واژه‌های نمایه‌ساز قابل اعتمادتر هستند. وایتاکر (۱۹۸۹) بر خلاف لیدسدورف، معتقد بود که مؤلفان واژه‌ها را برای تأثیر بر مخاطب انتخاب می‌کنند. از سوی دیگر همه عناوین استاندارد نیستند؛ مثلاً عناوینی که در علم و فن بیان به کار می‌رود. وایتاکر برای کشف اینکه آیا واژه‌های عنوان به کلیدواژه‌های تحلیل هم واژگانی ارجحیت دارد یا نه مطالعه‌ای تطبیقی انجام داد. او دریافت که هر دو روش تصویر مشابهی را ایجاد می‌کند؛ اما تصویر کلیدواژه‌های متنی مفصل‌تر است. در عین حال هیچ کدام از روش‌ها برتری ویژه‌ای را نشان نداد.

بررسی متون نشان می‌دهد که واژه‌های مورد استفاده در تحلیل هم واژگانی، از کلیدواژه‌های یک اصطلاح‌نامه تا واژه‌های متن کامل گسترش می‌یابد. اولین مطالعات تحلیل هم واژگانی بر اساس کلیدواژه‌های اصطلاح‌نامه‌ای انجام شده است (باینف، ۱۹۸۶). بعد از آن متون بر اساس عنوان، خلاصه و یا تعداد مشخصی کلیدواژه‌های منحصر (توصیفگرها) یک اصطلاح‌نامه مورد مطالعه قرار گرفتند (کالون و دیگران، ۱۹۹۱). روتو و مورگان (۱۹۹۷) پیشنهاد دادند که تحلیل هم واژگانی را می‌توان در سطح چکیده‌ها و با واژه‌های پیشنهادی متخصصان اجرا کرد. کتسف[1] معتقد بود یکی از امتیازات تحلیل تمام متن توانایی بازیابی عبارات مهم و در عین حال کم بسامد است که در سایر تحلیل‌ها نادیده گرفته می‌شود (کتسف و دیگران، ۱۹۹۷).

با توجه به توضیحات ارائه شده، تحلیل هم استنادی و تحلیل هم واژگانی دو روش معمول برای ساخت نقشه‌های راهبردی و موضوعی یک حوزه‌اند، پرسشی که ممکن است مطرح شود این است که کدام روش باید برگزیده شود (هی، ۱۹۹۹). پژوهشگران رویکردهای مختلفی نسبت به این دو گزینه داشته‌اند. باین و همکاران (۱۹۹۱) به دو دلیل تحلیل هم واژگانی را برگزیدند. دلیل اول این بود که آنها به دنبال مطالعه ساختار دانش یک حوزه بودند و نه روابط میان پژوهشگران. تحلیل هم واژگانی بر مفاهیم علمی انتشار تأکید دارد. دلیل دوم، دلیل روش‌شناختی بود. آنها به دنبال آزمون سودمندی این روش در فرآیند طراحی راهبردی بودند تا ارزش آن را به عنوان ابزاری در مدیریت علم دریابند. کالون و دیگران هم واژگانی را ابزار مناسب‌تری برای مطالعه روابط درونی میان پژوهش‌های دانشگاهی و فناورانه دانستند؛ زیرا شاخص‌های تحلیل استنادی تنها وجود ارتباط را نشان داده و اطلاعاتی از موضوع مورد پرسش ارائه نمی‌دهند. برای درک این مسئله که آیا عامل و بنیان اصلی اختراع یا ابداع و فناوری، پژوهش‌های علمی است، باید به خود متون

1 Kostoff

مراجعه کرد و محتوای مقالات و اختراعات را بررسی کرد. بررسی مطالعات مختلف در زمینه تحلیل هم واژگانی نشان می‌دهد که این روش در موارد زیر کاربردی است:

- ترسیم حرکت و پویایی علم؛
- ترسیم ساختار پژوهش علمی؛
- ترسیم روابط میان پژوهش‌های بنیانی و پژوهش‌های فناورانه؛
- ارزیابی درون داد/ برونداد روابط در یک شبکه پژوهشی؛ و
- طبقه‌بندی مدارک بر اساس موضوعات.

به‌طور کلی رویکرد تحلیل هم واژگانی روی سه فرضیه استوار است: (۱) کلمات استفاده شده در متون علمی به دقت توسط نویسندگان انتخاب می‌شوند (۲) استفاده از کلمات مختلف در متنی یکسان، لازمه وجود برخی روابط غیر جزئی[1] بین آن کلمات است (۳) تکرار هم‌رخدادی‌های واژه‌ها در متون توسط نویسندگان مختلف، بدین معنی است که روابط بین این واژه‌ها، در حوزه علومی که مورد مطالعه واقع می‌شوند اهمیت دارد (میلوجویک، ۲۰۰۹).

در پاسخ به توسعه نقشه‌های هم استنادی توسط اسمال در سال ۱۹۷۳، کالون و دیگران در سال ۱۹۸۳، توسعه نقشه‌های هم واژگانی را به عنوان جایگزینی برای مطالعه روابط معنایی در متون علمی و فناوری پیشنهاد نمودند. از آن زمان به بعد فناوری‌هایی برای "ترسیم هم واژگانی " بیشتر و بیشتر توسعه پیدا نمودند. این روش‌ها بر روی ماتریس واژه- مدرک[2] که در آن اسناد می‌توانند به عنوان سلول‌هایی (ردیف‌ها) که در آن واژه‌ها به عنوان فراهم کننده موردها (ستون‌ها) نسبت داده می‌شوند، در نظر گرفته می‌شوند (لیدسدورف و ولبرز[3]، ۲۰۱۱).

هنگام بررسی متون با استفاده از تحلیل هم واژگانی، رخداد واژگان در اسناد به وسیله ماتریس نمایش داده می‌شود. در همین خصوص اسناد به عنوان بخش تحلیل در نظر گرفته می‌شوند. این اسناد می‌توانند از نظر اندازه‌هایشان از اسناد بزرگ تا یک جمله متغیر باشند. اسناد شامل واژه‌هایی هستند که می‌توانند در درون جملات، پاراگراف‌ها و بخش‌ها سازماندهی شوند. ساختار معنایی در روابط میان واژگان می‌توانند در سطوح مختلف تراکم[4] متفاوت باشند (لیدسدورف، ۱۹۹۲، ۱۹۹۵)؛ بنابراین پژوهشگر، در ابتدا باید تصمیم بگیرد که چه چیزی به عنوان بخش مرتبط تحلیل، در نظر گرفته شود. دوماً کدام واژه‌ها باید در تحلیل گنجانده شوند؟

1 Non-Trivial

2 Word-Document Matrix
3 Leydesdorff & Welbers
4 Aggregation

گزینه آشکار برای این بخش فراوانی رخداد واژه‌ها (بعد از تصحیح برای واژگان سیاهه بازدارنده[1]) است. اگر چه سالتون و مک گیل (۱۹۸۳) پیشنهاد کردند که بیشترین و کمترین فراوانی‌های رخداد واژه‌ها نسبت به واژه‌هایی که فراوانی متوسطی دارند، از اهمیت کمتری برخوردارند. بدین منظور این نویسندگان سنجه‌ای را پیشنهاد نمودند که "فراوانی اصطلاح- برعکس فراوانی سند[2]" نامیده می‌شود؛ یعنی وزنی که با فراوانی اصطلاح i افزایش می‌یابد، اما هنگامی که اصطلاح در اسناد (k) بیشتری در مجموعه رخ می‌دهد (از nسند) کاهش می‌یابد. tf-idf می‌تواند به صورت زیر فرمول‌بندی شود:

به‌طور کلی پژوهشگر می‌تواند از چهار معیار برای انتخاب سیاهه‌ای از واژگانی که باید در تحلیل گنجانده شوند، استفاده نماید.

$$\text{Tf-Idf}_{ik} = \text{FREQ}_{ik} * [\log_2 (n / \text{DOCFREQ}_k)]$$

۱- فراوانی واژه

۲- ارزش tf-idf

۳- مشارکت ستون‌ها به ماتریس خی دو[3] و

۴- مجموع تفاضل مقدار مشاهده شده /مقدار مورد انتظار برای هر کلمه[4]

در مطالعاتی که توسط لیدسدورف و ولبرز صورت گرفت، متوجه شدند که سنجه‌ی آخری راحت‌تر است. اگر چه کل چهار سنجه قابل دسترس هستند (لیدسدورف و ولبرز، ۲۰۱۱).

تحلیل هم رخدادی[5]

متن‌کاوی کاربرد داده‌کاوی در متون زبان طبیعی است. مهم‌ترین نکته مرتبط در انتقال فنون عمومی داده-کاوی به حوزه بازنمون متون است که شامل مرحله قبل از پردازش و ویژگی‌های آماری خاص از داده‌های متنی است که زمینه تشکیل الگوریتم‌های خاص داده‌کاوی را باعث می‌شود (لئوپارد و دیگران، ۲۰۰۴).

منظور از هم رخداده‌ای واژگان این است که دو اصطلاح با هم در یک مدرک به کار برده شوند و هر چه قدر هم بیشتر با همدیگر تکرار شده باشند، این دو واژه ارتباط بیشتری با هم دارند. هم رخدادی دو

1 Stop Words
2 Term Frequency-Inverse Document Frequency
3 The Contribution Of The Column To The Chi-Square Of The Matrix
4 The Margin Totals Of Observed/Expected For Each Word
5 Co-Occurrences

اصطلاح یا دو واژه برای کشف پیوند و رابطه میان دو موضوع در یک حوزه پژوهشی استفاده می‌شود و از این طریق می‌توان توسعه و پیشرفت آن حوزه از علم را پیگیری نمود.

متن‌کاوی و تجزیه و تحلیل هم رخداده‌ای یکی از روش‌های مهمی است که می‌توان از طریق آن روند شکل‌گیری حوزه‌های علمی را مورد بررسی قرار داد (بوهم و دیگران، ۲۰۰۲). تحلیل هم رخداده‌ایی و تحلیل هم واژگانی گهگاه به جای هم نیز بکار برده می‌شوند.

سنجه های مورد استفاده برای تحلیل نقشه های علمی

داده‌هایی که برای پژوهش‌های مختلف جمع‌آوری می‌گردد با استفاده از روش‌ها و مدل‌های مختلفی مورد تحلیل قرار می‌گیرند. پژوهش‌های مرتبط با شبکه‌های اجتماعی نیز از این قاعده مستثنی نمی‌باشند. برای تحلیل داده‌های به دست آمده جهت تحلیل شبکه‌های اجتماعی از سنجه‌های مختلفی استفاده می‌گردد، تحلیل شبکه‌های اجتماعی از انواعی از روش‌ها که برای تحلیل سطوح و صفات مختلف یک شبکه اجتماعی طراحی شده‌اند، تشکیل شده است. روش‌های شبکه معمولاً برای مفاهیم در سطح خاصی از تحلیل مناسب می‌باشند در این فصل تلاش گردیده است تا پرکاربردترین‌های آنها در علم‌سنجی و مطالعات علم معرفی و تعریف شوند.

سنجه تراکم[1]

اگر دو جمعیت را مقایسه کنیم و متوجه شویم که نقش‌آفرینان زیادی در یکی از آنها وجود دارد که به دیگر نقش‌آفرینان اتصال پیدا نکرده‌اند (ایزوله هستند) و در دیگر جمعیت اکثر نقش‌آفرینان حداقل در یک اشتراک دوتایی جای دارند، نتیجه می‌گیریم که زندگی اجتماعی در این دو جمعیت خیلی متفاوت است؛ به عبارت دیگر در جمعیت اولی تعاملات بین نقش‌آفرینان بسیار کم و در جمعیت دوم تعاملات فراوانی وجود دارد؛ یعنی در اولی تراکم کم و در دومی تراکم زیاد است.

تراکم به عنوان تعداد آرک‌ها (روابط مستقیم بین نقش‌آفرینان) در یک شبکه تعریف شده و در بردارنده‌ی بالاترین سهم در میان روابط ممکن در شبکه است (نووی، مروار و باتاگلج[2]، ۲۰۰۵). تراکم عبارت است از نسبت تعداد آرک‌ها که با حرف "L" مشخص شده است، بر تعداد آرک‌های ممکن. چون یک آرک یک زوج مرتب شده[3] از رئوس است، کل آرک‌های ممکن برابر با n (n _ 1) خواهد بود که n تعداد کل

1 Density

2 Nooy, Mrvar & Batagelj

3 Ordered Pair

رئوس در شبکه است. تراکم یک شبکه حاصل کسری است که در محدوده‌ای از حداقل صفر تا حداکثر یک، وقتی که تمام آرک‌ها در شبکه حضور دارند قرار دارد (راچرلا و هو ۲۰۰۸، ص ۱۲). هرگاه تراکم شبکه را با علامت Δ نشان داده شود، مقدار آن از فرمول زیر به دست می‌آید.

$$\Delta = L / n (n \_ 1)$$

اندازه‌گیری تراکم یک شبکه شاخصی آماده از رتبه ارتباط‌های دوتایی در یک جمعیت را به ما می‌دهد. برای داده‌های دودویی، تراکم نسبتِ تعداد مجاورانی که موجودند، تقسیم بر تعداد زوج‌ها است- چه بخشی از تمام ارتباط‌های دوتایی ممکن، واقعاً موجود است. اگر ما گره‌های بین نقش‌آفرینان را با ارزش‌هایشان بسنجیم (قدرت، نزدیکی، احتمالات و غیره) تراکم معمولاً به عنوان مجموع ارزش‌های تمامی گره‌ها تقسیم بر تعداد گره‌های ممکن تعریف می‌شود؛ یعنی با داده‌های ارزش‌گذاری شده، تراکم معمولاً به عنوان میانگین قدرت گره‌ها در سراسر کل گره‌های ممکن تعریف می‌شود. جایی که داده‌ها متقارن یا غیر مستقیم هستند، تراکم در ارتباط با تعداد زوج‌های منحصربه‌فرد ((n*n-1)/2) محاسبه می‌شود؛ جایی که داده‌ها مستقیم هستند، تراکم در سراسر تعداد کل زوج‌ها محاسبه می‌گردد (هانمان و ریدل، ۲۰۰۵). به‌طور مثال سهیلی (۱۳۹۱) در پژوهش خود برای به دست آوردن تراکم شبکه هم نویسندگی مجله‌های علم اطلاعات از فرمول زیر استفاده نمود؛ که در آن L تعداد خطوط موجود و n تعداد گره‌های درون شبکه را نشان می‌دهد.

$$\frac{L}{-N(N \quad 1)/2}$$

نتایج وی نشان داد که "مجله انجمن انفورماتیک پزشکی آمریکا "، با تراکم ۰/۳ در رتبه اول قرار گفته است. در تصویر ۶-۱ و ۶-۲ تصویر دو شبکه سست و گسسته و یک شبکه نسبتاً متراکم نشان داده شده است. ترکم ۰/۳ به این معنا است که تنها ۳۰ درصد از پیوندهای ممکن در بین این شبکه‌ایجاد شده است.

به منظور درک بهتر تراکم شبکه به فهمیدن مفهوم قطر شبکه و فشردگی شبکه نیازاست. برای محاسبه قطر شبکه هم نویسندگی موجود در شبکه هم نویسندگی با استفاده از ماتریس‌های ایجاد شده از روابط بین نویسندگان، ابتدا شبکه موجود در بین این نویسندگان با استفاده از نرم‌افزارهای دیداری‌سازی ترسیم می‌گردد. سپس برای محاسبه قطر شبکه هم نویسندگی، مؤلفه اصلی هر شبکه استخراج می‌گردد و بر اساس داده‌های این مؤلفه، به اندازه‌گیری قطر شبکه اقدام می‌شود، هر چه قطر شبکه بیشتر باشد نشان‌دهنده آن است که تبادل

اطلاعات در آن شبکه نسبت به سایر شبکه‌ها با کندی صورت می‌گیرد. به‌طور کلی شبکه گسسته یا پراکنده شبکه‌ای است که اتصال بین پیوندها در یک نگاشت کم است، یا به عبارت دیگر تعداد خطوط یا پیوندهای متناظر کمتر از تعداد رئوس باشد و شبکه متراکم (پیوسته) شبکه‌ای است که تعداد خطوط یا پیوندها در یک نگاشت بیشتر از تعداد رئوس باشد. چنین شبکه‌ای شبکه پیوسته یا متراکم نامیده می‌شود (سهیلی، ۱۳۹۱).

همچنین فشردگی[1] از فاکتورهای بررسی انسجام شبکه است و در واقع به تفسیر قطر شبکه کمک می‌نماید. فشردگی از ۰ تا ۱ ارزش گذاری می‌شود. هر چه فشردگی به ۱ نزدیک‌تر باشد، نشان دهنده آن است که شبکه انسجام بیشتری دارد و هر چه به سمت صفر نزدیک باشد نشان دهنده انسجام پایین شبکه است. در واقع برای یک شبکه هر چه تراکم شبکه افزایش پیدا کند، قطر شبکه کوچک‌تر می‌شود.

فاصله شبکه

مفهوم دیگری که در تحلیل ساختار شبکه معرف است فاصله شبکه است. فاصله بین هر دو گره در شبکه، تعداد لبه‌هایی است که باید به منظور رسیدن به یک گره از دیگر گره‌ها طی شود و سنجش اینکه چگونه آن دو گره به دقت به هم پیوند خورده‌اند. نویسندگانی که یک مقاله را مشترک نوشته‌اند، فاصله یک دارند، دو نویسنده که هیچ مقاله‌ای با هم ننوشته‌اند اما هر کدام یک مقاله هم نویسنده با نویسنده سوم مشترک‌اند در فاصله ۲ قرار دارند و غیره. اندیشه فاصله در ورای عدد مشهور اردوش قرار دارد که نزدیکی هر نفر با این ریاضیدان پر تولید را نشان می‌دهد.

درجه ارتباط

درجه ارتباط یک شبکه یا شبکه فرعی با استفاده از سنجه تراکم معلوم می‌شود که عبارت از درصد تعداد ارتباط‌های واقعی در تعداد کل ارتباط‌های ممکن است (چئونگ و کوربیت، ۲۰۰۹). یکی دیگر از جنبه های جالب یک شبکه، شاخصی از تعداد دفعاتی برای یک ارتباط است که از طریق شبکه می‌گذرد. سنجه رایج مورد استفاده برای آن قطر شبکه[2] است؛ هرچه قطر شبکه کوتاه‌تر باشد، توزیع ارتباط سریع‌تر صورت می‌گیرد. قطر یک شبکه با استفاده از بلندترین فاصله ژئودیسک در شبکه اندازه‌گیری می‌شود، بدین گونه که با فاصله‌های ژئودیسک کوتاه‌ترین مسیر بر حسب تعداد پیوندهای موجود بین هر کدام از دو گره است.

1 Compactness
2 Diameter Of The Network

گسست‌های ساختاری

گسست‌های ساختاری به نام برت[1] در سال ۱۹۹۲ برای اشاره به برخی جنبه‌های مهم سود یا زیان موقعیتی نقش‌آفرینان در شبکه ثبت شده است. او برای توضیح اینکه چگونه و چرا مسیرهایی که نقش‌آفرینان به هم متصل می‌شوند بر محدودیت‌ها و فرصت‌ها و پس از آن بر رفتارشان تأثیر می‌گذارد تعدادی از سنجه‌ها را توسعه داد.

گسست‌های ساختاری بر حسب اندازه مؤثر شبکه اندازه‌گیری می‌شوند، یعنی تعداد ارتباط‌هایی که یک فرد دارد، منهای میانگین تعداد ارتباط‌هایی که هر فرد با سایر افراد دارد (چئونگ و کوربیت، ۲۰۰۹).

برت (۲۰۰۴) در پژوهش خود متوجه گردید در یک ساختار اجتماعی افرادی که نزدیک گسست‌ها قرار دارند احتمال زیادی دارد که ایده‌های خوبی داشته باشند. در این عقیده مزایای واسطه‌ای که از گسست‌های ساختاری ناشی می‌شوند، عمدتاً از طریق مزیت نگرشی، بر کارایی تأثیر می‌گذارد. او بیان کرد افرادی که شبکه‌های آنها، گسست‌های ساختاری بین گروه‌ها را پر می‌کند، زودتر از دیگران به انواع گسترده‌ای از اطلاعات دسترسی داشته و تجربه تفسیر اطلاعات در سراسر گره‌ها را دارند. آنان همچنین از مزیت کشف و توسعه فرصت‌های ارزشمند برخوردارند و به چگونگی مبادله مفید اطلاعات تسلط دارند. آنان قادرند تا زودتر و خیلی گسترده‌تر ببینند و اطلاعات را در سراسر گروه‌ها تفسیر کنند. واسطه در سراسر گسست‌های ساختاری بین گروه‌ها نگرشی از گزینه‌ها را فراهم می‌کند که در غیر این صورت نادیده گرفته می‌شوند. بارت همچنین ارزش اطلاعات را به عنوان یک منبع شبکه‌ای بیان کرد. ایده‌های نو اغلب شامل ترکیبی از بیت‌هایی از دانش در سراسر گروه‌ها می‌باشند. بارت همچنین خاطر نشان کرد که هر چه افراد متخصص‌تر باشند، ارزش اندیشه‌ها و اطلاعات مکمل بالاتر است؛ زیرا همگام شدن با توسعه سایر متخصصان، غیر ممکن است؛ بنابراین بازاری برای معامله با سود اطلاعاتِ کارآفرینان[2] شبکه وجود دارد (به نقل از جانسن و دیگران، ۲۰۱۰).

ارتباط ضعیف بین گروه‌ها در شبکه‌های اجتماعی، باعث ایجاد گسست‌هایی در ساختار شبکه می‌گردد. این گسست‌ها در ساختار اجتماعی- یا ساده‌تر گسست‌های ساختاری- مزایای رقابتی را برای فردی که روابطش این گسست‌ها را پر می‌کند ایجاد می‌کنند. گسست‌های ساختاری بین دو گروه به این معنا نیست که افراد در گروه‌ها از یکدیگر بی اطلاع هستند. بلکه بدین معناست که افراد آنچنان بر فعالیت‌های خودشان متمرکز هستند که آنها نمی‌توانند در فعالیت‌های سایر افراد در دیگر گروه‌ها حاضر شوند. گسست‌ها، شبیه یک عایق یا جداکننده در یک مدار الکتریکی جداکننده هستند. مردم در اطراف گسستی ساختاری در جریان‌های

1 Burt
2 Entrepreneurs

مختلف اطلاعات حرکت می‌کنند؛ بنابراین گسست‌های ساختاری فرصتی برای واسطه جریان اطلاعات بین مردم و کنترل بر پروژه‌هایی که افراد را از لبه‌های مخالف گسست کنار هم می‌آورد، می‌باشند (برت، ۲۰۰۲). گسست‌های ساختاری منابع غیرتکراری[۱] اطلاعات را جدا می‌کنند، منابعی که بیشتر فزاینده[۲] هستند تا همپوشان. دو شاخص برای غیرتکراری وجود دارد: انسجام[۳] و تعادل[۴]. ارتباط‌های منسجم (ارتباط‌هایی که به شدت به هم متصل هستند) احتمالاً اطلاعات مشابهی دارند و بنابراین اطلاعات غیر تکراری مفیدی را فراهم می‌کنند. ارتباط‌های ساختاری تعادلی (ارتباط‌هایی که مدیر را به همان شخص سوم پیوند می‌دهد[۵]) منابع اطلاعات مشابهی دارند و اطلاعات غیرتکراری مفیدی را به همراه دارد (برت، ۲۰۰۰).

مرکزیت[۶]

در طول تاریخ افراد، کشورها، سازمان‌ها و شرکت‌های برجسته، معروف و قدرتمند همواره مورد توجه بوده و دارای نقش‌های کلیدی و تأثیرگذار بوده‌اند، این افراد، شرکت‌ها و غیره با استفاده از ارتباط‌هایی که میان آنها با سایر همنوعانشان وجود دارد شبکه‌هایی را به وجود می‌آورند. این شبکه‌های اجتماعی ایجاد شده مورد تحلیل قرار می‌گیرند و اطلاعات ارزشمندی از آن جهت برنامه‌ریزی، مدیریت و پیش بینی اهداف بلند مدت و کوتاه مدت به دست می‌آید. این افراد و مؤسسات برجسته بر اساس مرکزیتشان مورد ارزیابی قرار می‌گیرند؛ یعنی میزان قدرت و تأثیرگذاری آنها در میان شبکه اجتماعی که عضوی از آن می‌باشند. هر چه شخص یا مؤسسه‌ای مرکزی‌تر باشد یعنی پرستیژ و اقتدار بیشتری دارد و افراد یا سازمان‌هایی که هم به آنها نزدیک هستند معمولاً نقش میانجی و واسطه در تبادل و جریان اطلاعات دارند و همچنین ممکن است نقش بینابینی ایفا نمایند. وقتی که یک شبکه اجتماعی نظام‌مند میان مجموعه‌ای از نقش‌آفرینان وجود دارد، مرکزیت هر نقش‌آفرین در شکل دهی شبکه اجتماعی، جهت یافتن مشخصه‌ها و ساختار شبکه اجتماعی اهمیت زیادی پیدا می‌کند؛ زیرا مرکزیت یک نقش‌آفرین، نشان دهنده اهمیت، شأن، قدرت و شهرت نقش‌آفرین برای شکل-دهی روابطش با دیگر نقش‌آفرینان در شبکه اجتماعی است (سهیلی و عصاره، ۱۳۹۲).

مرکزیت هر نقش‌آفرین در شکل دهی شبکه اجتماعی، جهت یافتن مشخصه‌ها و ساختار شبکه اجتماعی اهمیت زیادی پیدا می‌کند؛ زیرا مرکزیت یک نقش‌آفرین، نشان دهنده اهمیت، شأن، قدرت و شهرت نقش‌آفرین برای شکل‌دهی روابطش با دیگر نقش‌آفرین‌ها در شبکه اجتماعی است.

1 Nonredundant
2 Additive
3 Coheicive
4 Equivalence
5 Contacts Who Link A Manager To The Same Third Parties
6 Centrality

مرکزیت یک فرد در شبکه اجتماعی نشان دهنده پرستیژ و اقتدار فرد در شبکه است. افرادی که در مرکز شبکه قرار دارند از لحاظ علمی تأثیرگذاری بیشتری دارند. در یک شبکه مردم اغلب به شناسایی برجسته‌ترین نقش‌آفرین (ها) علاقه‌مندند. مرکزیت سنجه‌ای است که برتری یک نقش‌آفرینان فردی که در شبکه جاسازی شده است را کمی سازی می‌نماید. موقعیت یک نقش‌آفرین معمولاً بر حسب مرکزیتش بیان می‌گردد، یعنی سنجش چگونگی مرکزیت آن نقش‌آفرین در شبکه. نقش‌آفرین‌های مرکزی به خوبی با سایر نقش‌آفرین‌ها مرتبط هستند و سنجه مرکزیت سعی در اندازه‌گیری یک نقش‌آفرین (تعداد پیوندهای بیرونی و درونی) و فاصله نسبی با نقش‌آفرین‌های دیگر یا رتبه‌ای که کوتاه‌ترین مسیرهای بین هر جفت نقش‌آفرین را که از نقش‌آفرین مورد نظر عبور می‌کند اندازه‌گیری می‌کند، دارد (لیو و دیگران، ۲۰۰۵). ساده‌ترین مقیاس سنجه مرکزیت، شمار پیوندهایی است که عضو یک شبکه (فرد) با دیگر اعضای شبکه دارد.

اولین کاربردهای ایده مرکزیت افراد، از طریق تحلیل شبکه‌های اجتماعی صورت گرفت؛ و ریشه این ایده را می‌توان در مفهوم نخبه‌های جامعه سنجانه[1] پیدا کرد، یعنی مشهورترین فرد یا افراد در مرکز توجه (اسکات، ۲۰۰۷)؛ بنابراین نقش‌آفرین مرکزی، نقش‌آفرینی است که در مرکز تعداد زیادی ارتباط قرار می‌گیرد، یعنی نقش‌آفرینی با تعداد زیادی پیوند مستقیم با سایر نقش‌آفرینان.

میزان مرکزیت با استفاده از رتبه‌ی[2] گره‌های مختلف در شبکه اندازه‌گیری می‌شود. رتبه‌ای که تعداد گره‌های دیگری که یک گره با آنها همسایه است را نمایش می‌دهد، اندازه‌گیری می‌شود. این نوع سنجش مرکزیت به عنوان مرکزیت محلی[3] شناخته شده است، چون در آن ارتباط‌های غیر مستقیم به یک گره‌ی خاص نادیده گرفته شده است؛ بنابراین اندیشه مرکزیت به مرکزیت جهانی[4] گسترش می‌یابد (فریمن، ۱۹۷۹) تا ارتباط‌های دوردست را هم در بر بگیرد. این امر به وسیله نزدیکی[5] گره‌ها به سایر گره‌ها که برحسب فاصله بین گره‌های مختلف بیان شده، اندازه‌گیری می‌شود. بینابینی[6] (فریمن، ۱۹۷۹) یکی دیگر از سنجه‌های مرکزیت مرکزیت است که فضایی که در آن یک گره خاص بین دیگر گره‌های متعدد شبکه قرار گرفته است را اندازه-گیری می‌کند. یک گره نسبتاً درجه پایین ممکن است نقش یک گره میانجی مهم را بازی بکند (به عنوان مثال واسطه، دروازه‌بان و غیره) و از این‌رو یک گره مرکزی در شبکه به حساب بیاید.

مرکزیت به موقعیت یک گره درون یک شبکه مخصوص اشاره دارد؛ بنابراین دو سنجه مرکزیت باید در طی تحلیل در نظر گرفته شود. مرکزیت محلی و مرکزیت جهانی (هاتالا، ۲۰۰۶). مرکزیت محلی با تعداد

1 Sociometric Star
2 Degree
3 Local Centrality
4 Global Centrality
5 Closeness
6 Betweenness

گره‌های مستقیم با کل گره‌ها در شبکه سر و کار دارد. عددِ مرکزیت محلی بالا، نشان‌دهنده‌ی موقعیت مرکزی‌تر گره است. این گره‌ها می‌توانند به تسهیل جریان اطلاعات از یک گروه به گروه دیگر درون یک بافت سازمانی کمک کنند. بدون این گره‌ها، حفره‌های ساختاری به وجود می‌آیند. در نتیجه جریان اطلاعات به‌طور آزاد از یک گروه به گروه دیگر مشکل خواهد بود، مگر اینکه آن گره از طریق فردی که به گروه متصل است، گذر کند. به دلایل آشکار، افرادی که این فاصله را پر می‌کنند در موقعیت قدرت قرار دارند و می‌توانند کنترل کنند که چه اطلاعاتی برای آنان در شبکه جریان پیدا کند (بارت[1]، ۱۹۹۲، ۱۹۹۷).

مرکزیت جهانی به وسیله افزودن تمامی مسیرها از یک گره خاص به کل گره‌های دیگر در شبکه محاسبه می‌شود. اگر یک گره از طریق گره دیگر اتصال پیدا کرده باشد، دو مسیر به محاسبه کلی مرکزیت جهانی اضافه می‌گردد. محاسبه مرکزیت جهانی ممکن است برای گره‌هایی که خیلی ارتباط برقرار نکرده‌اند، اما پیوندهایی از مجموعه‌ای از گره‌ها به دیگر مجموعه‌ها فراهم کرده‌اند، مفیدتر باشد.

مرکزیت یکی از مهم‌ترین و پراستفاده‌ترین سنجه‌ها در تحلیل شبکه‌های اجتماعی است. مرکزیت ویژگی توصیفی برای نقش‌آفرینان یا گروهی از نقش‌آفرینان با مشخصه‌های ساختاری متعدد و پارامتری تعیین کننده برای درک و تحلیل نقش‌های نقش‌آفرینان در شبکه‌های اجتماعی است (نیومن، ۲۰۰۵). معمولاً از مرکزیت برای شناسایی نقش‌آفرینان قدرتمند و بانفوذ یا مهم استفاده می‌شود. به خاطر ادراک متفاوت از قدرت اجتماعی و کاربردهای متنوع تحلیل شبکه‌های اجتماعی، مرکزیت تعریف‌های گوناگونی دارد (کاررینگتون، ۲۰۰۵). گسترده‌ترین تعریف پذیرفته شده از مرکزیت در اواخر ۱۹۷۰ توسط فریمن ارائه گردید. در تعریف فریمن، سنجه مرکزیت عمدتاً بر اساس سه جنبه رتبه، نزدیکی و بینابینی سنجیده می‌شود. مرکزیت یکی از قدیمی‌ترین مفاهیم در تحلیل شبکه است و اکثر شبکه‌های اجتماعی شامل افراد یا سازمان‌هایی هستند که مرکزی می‌باشند. آنها به خاطر جایگاهشان دسترسی بهتری به اطلاعات و فرصت بهتری برای گسترش اطلاعات دارند. این موضوع به عنوان رویکرد خودمحور به مرکزیت شناخته می‌شود. شبکه از چشم‌انداز جمع‌محور هم متمرکز است. اندیشه مرکزیت به جایگاه رئوس افراد درون شبکه اشاره دارد، در حالی که تمرکز برای مشخص کردن کل شبکه به کار می‌رود. به شبکه خیلی متمرکز گفته می‌شود که مرز واضحی بین مرکز و پیرامون آن وجود داشته باشد. در شبکه خیلی متمرکز، اطلاعات به آسانی گسترش می‌یابد، اما وجود مرکز برای انتقال اطلاعات الزامی است (یاسمین[2] و دیگران، ۲۰۰۸).

مرکزیت مفهومی است که برای تحلیل شبکه‌ها به کار رفته و دارای انواع متفاوتی است که براساس تعریف مسئله و هدف پژوهش یک یا چند مرکزیت مورد استفاده قرار می‌گیرد؛ اما به‌طور کلی از مرکزیت‌ها

1 Burt
2 Yasmin

برای شناسایی و تعیین مهم‌ترین نقش‌آفرینان در شبکه استفاده می‌شود. در خصوص اینکه مرکزیت واقعاً چیست و یا در مورد بنیادهای مفهومی آن اتفاق نظر وجود ندارد؛ تاکنون تنها توافق اندکی در خصوص روش مناسب اندازه‌گیری آن حاصل شده است. به‌طور کلی مرکزیت بیشتر یک نقطه، سبب دارا بودن درجه بالاتر، داشتن ارتباطات بیشتر و کسب موقعیت مطلوب‌تر است که نهایتاً فرد را قدرتمندتر می‌سازد. به‌طور کلی مرکزیت شبکه‌های مختلف بر اساس شاخص‌های متعددی سنجیده می‌شود که مهم‌ترین آنها عبارت‌اند از: رتبه، نزدیکی، بینابینی، بردار ویژه، گستردگی مرکزیت، الگوریتم رتبه صفحه، الگوریتم اچ آی تی اس و غیره. در این بخش هر یک از این شاخص‌ها معرفی و به‌طور مختصر تعریف می‌شوند.

مرکزیت رتبه

تحلیل مرکزیت رتبه دو نمره را در بر دارد: رتبه بیرونی (تعداد ارتباط‌های ارسال شده به بیرون یعنی به عنوان نویسنده اصلی) و رتبه درونی (تعداد ارتباط‌های دریافت شده یعنی به عنوان هم نویسنده). افراد با نمره‌های بالای رتبه بیرونی، می‌توانند به عنوان افرادی که در شبکه نفوذ و تأثیر بالایی دارند در نظر گرفته شوند. درحالی که آنانی که نمره‌های رتبه درونی بالایی دارند به عنوان اشخاص با اعتبار یا مشهور در نظر گرفته شوند.

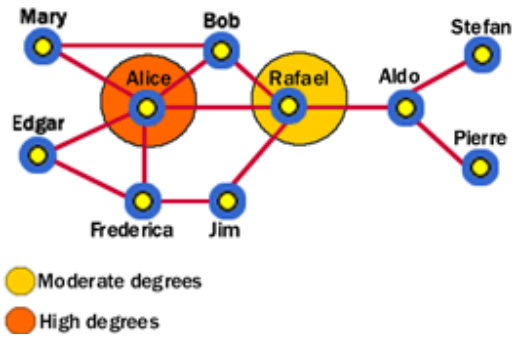
یکی از سنجه‌ها یا شاخص‌های شبکه‌ای که در تحلیل ساختارهای کل شبکه و موقعیت‌های افراد در شبکه مفید است مرکزیت رتبه است. مرکزیت رتبه به تعداد پیوندهای داده شده یا خارج شده از یک گره در یک شبکه اشاره دارد (فریمن، ۱۹۹۷) این سنجه به موقعیت افراد در یک شبکه مربوط است. شخصی مرکزی در شبکه اطلاعات (یعنی با نمره رتبه مرکزیت بالا) به حساب می‌آید که می‌تواند مهارت‌ها، تجربه‌ها و حافظه سازمانی برای دیگران ایجاد کند و از وی می‌توان به عنوان دارائی[1] سازمان نام برد. این شخص همچنین می‌تواند به عنوان یک مربی برای تازه واردها نقش ایفاء کند (پاریسی[2]، ۲۰۰۷). ضروری است این افراد را که می‌توانند به عنوان گلوگاهی برای جریان اطلاعات عمل کنند و نیز قادرند به‌طور بالقوه‌ای با درخواست‌های اطلاعاتی بیش از حد بار شوند شناسایی شوند (کروس و پروساک[3]، ۲۰۰۲). سنجش مرکزیت همچنین به مدیران فرصت می‌دهد تا افراد پیرامون شبکه را نیز شناسایی کنند (یعنی افراد با نمره پایین). شناخت افراد پیرامونی نیاز است، چون ممکن است در بر دارنده دانش با ارزشی باشند که چنانکه دارای موقعیت بهتری در شبکه بودند می‌شد آن را با دیگران به اشتراک گذاشت (پاریسی و دیگران، ۲۰۰۶).

1 Asset
2 Parise
3 Cross & Prusak

ساده‌ترین نوع مرکزیت است که ارزش مرکزیت هر نقطه تنها با شمارش تعداد همسایگانش به دست می‌آید؛ هر چه مرکزیت رتبه‌ی یک فرد بیشتر باشد، ارتباطات و شبکه‌ی بیشتری در اختیار داشته و تأثیرگذارتر است. گروه‌بندی نیز امکان درک چگونگی رفتار یک فرد یا سازمان درون یک گروه و رفتار کل شبکه را فراهم می‌سازد؛ در واقع به بررسی ساختار شبکه می‌پردازد (بورگتی، ۲۰۰۵).

مرکزیت رتبه عبارت است از تعداد روابط مستقیمی که یک موجودیت در شبکه دارد. موجودیتی با مرکزیت رتبه بالا دارای ویژگی‌های زیر است:

- به‌طور کلی نقش‌آفرینی فعال در شبکه است؛
- اغلب متصل‌کننده یا تقسیم کننده[1] در شبکه است؛
- ضرورتاً مرتبط‌ترین موجودیت در شبکه نیست (موجودیت ممکن است روابط زیادی داشته باشد که بیش‌تر آنها به موجودیت‌های سطح پایین باشد)؛
- ممکن است در موقعیت ممتازی در شبکه قرار داشته باشد؛
- ممکن است مسیرهای جایگزینی برای برآوردن نیازهای سازمان داشته باشد و در نتیجه احتمالاً خیلی کم به دیگر افراد وابسته باشد؛
- اغلب می‌تواند به عنوان شخص سوم یا واسطه شناخته شود (سنتینلویژوالیز[2]، ۲۰۱۰).



تصویر ۶-۳: نمودار شبکه رتبه مرکزیت نقش‌آفرینان در یک شبکه

در تصویر ۶-۳، آلیس بالاترین مرکزیت رتبه را دارد، این بدین معناست که او کاملاً در شبکه فعال است. اگر چه او ضرورتاً قدرتمندترین شخص نیست، زیرا او صرفاً به صورت مستقیم درون یک رتبه به سایر افراد در دسته‌اش مرتبط است، او باید از طریق رافائل به سایر دسته‌ها برود.

1 Hub
2 Sentinelvisualize

مرکزیت رتبه به‌طور ساده با شمارش تعداد ارتباط‌هایی که توسط هر نقش‌آفرین در شبکه نگهداری می‌شود، اندازه‌گیری می‌شود. در یک نگاشت، این کار با شمارش تعداد گره‌ها یا خطوط وارد یا خارج شده از یک گره خاص تحقق می‌یابد. یک نقش‌آفرین با بیشترین خطوط، بالاترین رتبه و بنابراین مرکزی‌ترین گره است (چنگ، ۲۰۰۶). سنجه مرکزیت رتبه معمولاً انعکاس دهنده شهرت و فعالیت رابطه‌ای یک نقش‌آفرین است و به‌طور کلی مرکزیت رتبه محاسبه میزان پیوندهایی است که فرد با دیگر افراد در شبکه دارد.

مرکزیت رتبه به عنوان سنجه‌ای به بررسی میزان خروجی و ورودی دانش یا اطلاعات در یک گره می‌پردازد و گره‌هایی که دارای بیشترین ارتباط با دیگر گره‌ها هستند را به عنوان گره¬هایی با مرکزیت رتبه بالا معرفی می¬کنند. اگر تعداد پیوند، زیاد باشد به آن گره برجسته یا دارای جایگاه بالا گفته می¬شود (هانمان و ریدل، ۲۰۰۵). وقتی یک گره دارای مرکزیت رتبه بالایی باشد، توانایی بالایی در نفوذ بر سایر گره‌های شبکه را دارد.

به‌طور مثال در تصویر ۶-۴، گلنزل بالاترین مرکزیت رتبه را داد و در نتیجه جایگاه و قدرت تأثیرگذاری بیشتری در این شبکه خواهد داشت. به‌طور کلی قدرت مفهومی است که بر پایه منزلت، جایگاه و ارتباطات هر فرد یا سازمان درون شبکه و به دلیل محدودیت‌ها یا فرصت‌های به وجود آمده برای وی، افزایش یا کاهش می‌یابد. هرچه میزان محدودیت فرد یا سازمان کمتر باشد، فرصت‌های او بیشتر شده و در نتیجه به جایگاه مطلوب‌تری دست می‌یابد، پس تبادلات بیشتری با دیگران برقرار کرده و تأثیر بیشتری بر آنها می‌گذارد یعنی توانمندتر می‌شود. افراد یا سازمان‌های مرجع، در دسترس‌تر و یا مرکزی‌تر دارای موقعیت مطلوب‌تری بوده و توانمندتر می‌باشند (براندز و الرباچ،۲۰۰۵). به‌طور کلی نقش‌آفرینانی که نمره مرکزیت بالاتری دارند، از فرصت‌ها و جایگزین‌های بیشتری نسبت به سایر نقش‌آفرینان برخوردار هستند. این نقش‌آفرینان، گره‌های بیشتر و فرصت‌های بیشتری را نیز دارند چون انتخاب‌های بیشتری دارند. این استقلال آنها را مستقل و به نقش‌آفرین خاص وابسته نمی‌کند. این افراد همچنین موقعیت‌های ممتازی دارند، چون که گره‌های زیادی دارند و راه‌هایی جایگزینی برای ارضای نیازهای خود داشته و از این رو کمتر به افراد دیگر وابسته هستند. آنها به بیشتر منابع شبکه به‌طور کل دسترسی داشته و قادرند بیشتر منابع درون شبکه را فرا بخوانند.

مرکزیت نزدیکی

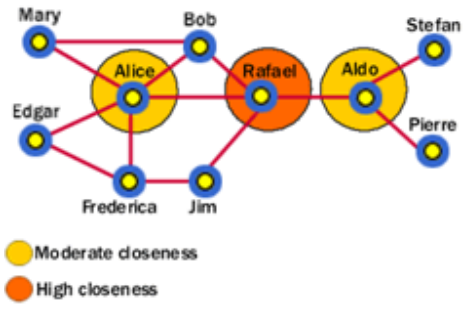
مرکزیت نزدیکی، فاصله یک فرد با کلیه افراد دیگر در شبکه را می‌سنجد، هر چه یک فرد به دیگران نزدیک‌تر باشد، آن فرد برگزیده‌تر و مشهورتر است. افرادی با نمرات نزدیکی بالا، احتمالاً اطلاعات را خیلی سریع‌تر از دیگران دریافت می‌کنند، به خاطر این‌که میانجی‌های کمتری بین آنها وجود دارد.

در سنجش مرکزیت نزدیکی، ارزیابی از طریق قضاوت کردن درباره نزدیکی یک نقش‌آفرین به نقش‌آفرینانی دیگر صورت می‌گیرد. در این نوع مرکزیت از طریق طول مسیرها یا گام‌هایی که برای یک نقش‌آفرین مورد نیاز است تا به دیگر نقش‌آفرینان در شبکه برسد، اندازه‌گیری صورت می‌گیرد. نقش‌آفرینانی که قادرند به دیگر نقش‌آفرینان با طول مسیر کوتاه‌تری برسند یا آن‌هایی که با طول مسیرهای کوتاه‌تر توسط دیگر نقش‌آفرینان دسترس‌پذیرترند، در موقعیت ممتازی قرار دارند و به‌طور کلی قدرت و نفوذ بیشتری در درون شبکه دارند (چنگ، ۲۰۰۶).

اینکه یک موجودیت در شبکه چقدر سریع می‌تواند به موجودیت‌های بیشتری در آن شبکه دسترسی پیدا کند سنجه‌ای است که مرکزیت نزدیکی را سنجش می‌کند. موجودیتی با مرکزیت نزدیکی بالا به‌طور کلی دارای ویژگی‌های زیر است:

- دسترسی سریعی به سایر موجودیت‌ها در شبکه دارد؛
- مسیر کوتاهی به سایر موجودیت‌ها دارد؛
- به سایر موجودیت‌ها نزدیک است؛
- رؤیت پذیری بالائی درباره آنچه در شبکه در حال اتفاق افتادن است دارد (سنتینلویژوالیز، ۲۰۱۰).

در تصویر ۶-۵، رافائل بالاترین مرکزیت نزدیکی را دارد، زیرا او می‌تواند از طریق مسیرهای کوتاه به موجودیت‌های بیشتری برسد. همین‌طور مکان رافائل به او اجازه می‌دهد تا با موجودیت‌های دسته‌ی خودش و با موجودیت‌هایی که در سایر دسته‌ها گسترش یافته‌اند ارتباط برقرار کند.



تصویر ۶-۵: نمودار شبکه نزدیک مرکزیت نقش‌آفرینان در یک شبکه

نکته قابل توجه این است که اگر شبکه دارای موجودیتی باشد که هیچ پیوندی دریافت نکرده است (یعنی به هیچ موجودیت دیگر پیوند نداده باشد)، مقدار نزدیکی برای کل موجودیت در شبکه صفر خواهد بود. این امر به دلیل فرمول‌ها و الگوریتم‌های ایجاد شده در تحلیل شبکه‌های اجتماعی است.

سنجه مرکزیت نزدیکی بر اساس کوتاه‌ترین فاصله[1] محاسبه می‌شود. این سنجه مقدار فاصله یک گره از سایر گره‌ها را اندازه‌گیری می‌کند. این سنجه نشان دهنده‌ی دسترس‌پذیری، سلامت و امنیت نقش‌آفرینان است (فرانک[2]،۲۰۰۲).

خیلی از پژوهشگران اجتماعی اظهار می‌دارند که برای شبکه‌های بزرگ سنجه‌ی مرکزیت نزدیکی جذاب نیست؛ زیرا در یک شبکه اجتماعی بزرگ، معمولاً یک نقش‌آفرینان تنها به مجموعه کوچکی از نقش‌آفرینان نزدیک است. معمولاً سنجه مرکزیت بینابینی، برای بیشتر نقش‌آفرینان در شبکه‌های اجتماعی بزرگ خیلی کوچک است. دلیل این مشکل این است که حاصل جمع تمام فاصله‌های ژئودیسک اطلاعات زیادی را از بین می‌برد؛ زیرا توزیع فاصله ژئودیسک از گره منبع، به تمامی گره‌ها اطلاعات مهم[3] را در بر دارد. برای مثال وقتی که گسترش فاصله‌ها تحلیل می‌شوند، لازم است که از این فاصله‌ها، به منظور برآورد میزان گسترش یک ایده در یک شبکه استفاده شود؛ بنابراین در تحلیل شبکه‌های اجتماعی بزرگ، مرکزیت نزدیکی برای یک گره توسط دو نوع پارامتر ارائه می‌گردد. یکی از آنها سنجه نزدیکی بر اساس فاصله ژئودیسک است که در بالا تعریف گردید. دیگری یک بردار فاصله است که مسافت‌های ژئودیسک از این گره شاخص را تا همه گره‌های دیگر ذخیره می‌کند (پن، ۲۰۰۷). اگر بخواهیم از تعاریف فوق جمع‌بندی نماییم مرکزیت نزدیکی عبارت است از تنوع مجموعه کوتاه‌ترین مسیرها بین هر فرد و دیگر افراد در شبکه؛ و مرکزیت نزدیکی، نقطه‌ای است که به‌طور متوسط به کلیه نقطه‌ها نزدیک است. نقطه‌ای دارای بیشترین مرکزیت نزدیکی است که به‌طور میانگین به کلیه نقطه‌ها نزدیک باشد. هرچه نقطه‌ای به مرکز نزدیک‌تر باشد، توانمندتر است.

سنجه مرکزیت نزدیکی، امکان محاسبه دوری و نزدیکی[4] هر کدام از گره‌ها را با سایر گره‌ها شبکه فراهم می‌آورد. دوری جمع فاصله هر کدام از گره‌ها با دیگران در یک شبکه است. محاسبه دوری یا نزدیکی

1 Geodesic
2 Frank
3 Non-Trivial
4 Farness And Closeness

یک گره در شبکه این امکان را به وجود می‌آورد که عدم یکنواختی، تفاوت توزیع و فاصله میان گره‌ها را بررسی کرد. انواع مختلفی از رویکردها برای محاسبه دوری یا نزدیکی در سنجه مرکزیت نزدیکی وجود دارد که عمومی‌ترین و معروف‌ترین آن "کوتاه‌ترین فاصله مسیر[1]" نامیده می‌شود که در واقع مجموع طول کوتاه‌ترین فاصله یک گره با سایرین است (هانمان و ریدل، ۲۰۰۵).

در تصویر ۶-۶، گلنزل با مرکزیت نزدیکی برابر با ۱/۴۵۷ بالاترین میزان نزدیکی به سایر گره‌های موجود در شبکه را دارد.

به منظور درک ساده‌تر از نتیجه بررسی مرکزیت نزدیکی در یک شبکه پیچیده و بزرگ، از روش مرکزیت دسترسی استفاده می‌کنند. این روش یکی از روش‌هایی است که خیلی ساده و روشن به سنجش نزدیکی پرداخته و چگونگی و سطح دسترسی به یک گره در میان سایر گره‌ها را بیان می‌کند. به عبارتی این سنجه به این پرسش جواب می‌دهد که چه اندازه و یا چند درصد از گره‌های موجود در یک شبکه می‌توانند در قدم‌های (پله) اول، دوم، سوم و غیره به یک گره در شبکه دسترسی پیدا بکنند. در بحث جریان دانش، به این مفهوم است که دانش تولید شده توسط یک گره در شبکه، چگونه و با چه سهولتی قابل دسترسی است.

جدول ۶-۲ نتیجه بررسی مرکزیت دسترسی را نشان می‌دهد. در این جدول فاصله دسترسی هر کدام از گره‌ها نسبت به کل گره‌ها آمده است، به عبارتی پاسخ به این پرسش است که چقدر باید مسیر طی شود تا یک گره به سایر گره‌ها برسد. حداکثر عدد موجود در این جدول برابر با تعداد کل گره‌ها است و زمانی به دست می‌آید که امکان دسترسی به یک گره در قدم اول برای تمامی گره‌ها شبکه باشد. داده‌ها نشان داد که کشورهای آمریکا، ژاپن و آلمان، فرانسه، بریتانیا، سوئیس، کانادا، کره جنوبی، تایوان و ایتالیا بیشترین فرصت را دارند که در قدم اول، سایر کشورها به آنها دسترسی داشته باشند و به عبارتی از دانش تولید شده آنها استفاده کنند. ضمن اینکه تمامی این کشورها در قدم دوم امکان ارتباط با تمامی دیگر کشورها را دارند. همچنین داده‌ها نشان داد که در قدم سوم تمامی کشورها قابلیت دسترسی را دارند. به عبارتی حداکثر سه مرحله یا سه فاصله بین کشورها وجود دارد تا ارتباط میان آنها برقرار گردد.

1 Geodesic Path Distance

جدول ۶-۲: شانس در دسترس قرار گرفتن کشورها در قدم‌های مختلف (منصوری، ۱۳۹۱).

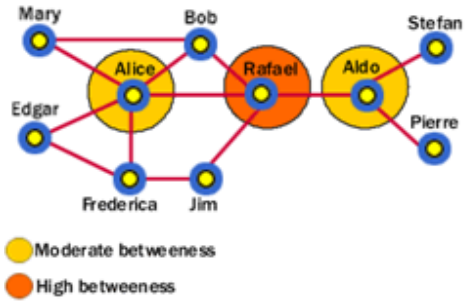
کشور	قدم اول (درصد)	قدم دوم (درصد)	قدم سوم
آمریکا	۹۴	۱۰۰	۱۰۰
ژاپن	۹۰	۱۰۰	۱۰۰
آلمان	۶۹	۱۰۰	۱۰۰
فرانسه	۶۰	۱۰۰	۱۰۰
بریتانیا	۵۸	۱۰۰	۱۰۰
سوئیس	۵۵	۱۰۰	۱۰۰
کانادا	۵۴	۱۰۰	۱۰۰
کره جنوبی	۵۲	۱۰۰	۱۰۰
تایوان	۵۱	۱۰۰	۱۰۰
ایتالیا	۴۶	۱۰۰	۱۰۰

مرکزیت بینابینی

سنجه مرکزیت بینابینی، شاخصی است که مسیر دقیق‌تری جهت اندازه‌گیری مرکزیت یک نقش‌آفرین را عرضه می‌نماید. این شاخص، مرکزیت را با بررسی وسعتی که در آن یک نقش‌آفرین خاص بین دیگر نقش‌آفرینان متنوع در شبکه، قرار می‌گیرد را اندازه‌گیری می‌نماید (چنگ، ۲۰۰۶).

سنجه مرکزیت بینابینی، موقعیت یک موجودیت را درون یک شبکه بر حسب توانایش جهت ایجاد ارتباط با سایر زوج‌ها یا گروه‌ها در شبکه، شناسایی می‌کند. موجودیتی با بالاترین مرکزیت بینابینی به‌طور کلی دارای ویژگی‌های زیر است:

- موقعیت مطلوب و قدرتمندی در شبکه به دست آورده است؛
- نقطه‌ی مجزایی از گسیختی را به نمایش می‌گذارد؛
- تأثیر خیلی زیادی بر آنچه که در شبکه اتفاق می‌افتد دارد (سنتینلویژوالیز، ۲۰۱۰).



تصویر ۶-۷: نمودار شبکه بینابینی مرکزیت نقش‌آفرینان در یک شبکه

در تصویر بالا رافائل بالاترین بینابینی را دارد زیرا او بین آلیس و آلدو که میان سایر موجودیت‌ها هستند قرار دارد. آلیس و آلدو بینابینی اندکی دارند؛ زیرا آنها اساساً تنها بین دسته‌های خود هستند، بنابراین اگرچه آلیس بالاترین رتبه مرکزیت را دارد، رافائل در بعضی از جنبه‌های خاص اهمیت بیشتری در شبکه دارد.

در تصویر ۶-۷، رافائل بالاترین بینابینی را دارد زیرا او بین آلیس و آلدو که میان سایر موجودیت‌ها هستند قرار دارد. آلیس و آلدو بینابینی اندکی دارند؛ زیرا آنها اساساً تنها بین دسته‌های خود هستند، بنابراین اگرچه آلیس بالاترین رتبه مرکزیت را دارد، رافائل در بعضی از جنبه‌های خاص اهمیت بیشتری در شبکه دارد.

روش دیگر سنجش مرکزیت مشخص کردن بینابینی گره‌ها است. این روش اشاره به گره مخصوصی دارد که در بین دیگر گره‌ها در شبکه واقع شده است. یک گره با رتبه نسبتاً پایینِ بینابینی ممکن است نقش میانجی مهمی را بازی بکند و برای شبکه خیلی مرکزی باشد (اسکات، ۲۰۰۰). به عنوان مثال بخشی درون یک سازمان که بینابینی بالایی دارد، نسبت به اختلال جریان اطلاعات آسیب پذیر است. در صورتی که فردی قصد ترک سازمان را داشته باشد نسبت به اختلال جریان اطلاعات آسیب پذیر است؛ بنابراین، مهم است تا این نقش‌آفرینان را به منظور هدایت مداخله‌های مناسب[1] شناسایی کنیم. مداخله امکان‌پذیر، می‌تواند ایجاد جلسات ماهیانه‌ای را که به تمامی اعضا هر دو بخش اجازه می‌دهد تا اطلاعاتشان را به اشتراک بگذارند، در بر بگیرد. این فرآیند رسمی تضمین خواهد کرد که اطلاعات بین اعضاء به اشتراک گذاشته شده و بین بخش‌ها در حال جریان پیدا کردن است (هاتالا، ۲۰۰۶).

سنجه مرکزیت بینابینی به بررسی کوتاه‌ترین مسیری که یک گره میان دیگر جفت‌های گره‌ها در یک شبکه می‌تواند قرار بگیرد، می‌پردازد. سنجه مرکزیت بینابینی یکی از مهم‌ترین سنجه‌ها برای بررسی و کنترل جریان دانش میان شبکه‌ها است (ناکی و یانگ، ۲۰۰۸ و بورگتی و اورت[2]، ۲۰۰۶). به‌طور مثال وقتی در یک شبکه جریان دانش، گره "الف " از طریق گره "ب " به گره "ج " متصل می‌شود، گره "ب " دارای مسئولیت کنترل انتقال دانش از گره "الف " به "ج " است. در اغلب موارد گره "ب " در بین کوتاه‌ترین مسیر میان گره‌های بی‌شماری قرار گرفته است، بنابراین این قابلیت و ویژگی را در گره "ب " به وجود می‌آورد که توانایی کنترل جریان دانش میان گره‌های بسیاری را داشته باشد. سنجه مرکزیت بینابینی به عبارتی به بررسی میزان قدرت و تأثیرگذاری یک گره در شبکه می‌پردازد. در شبکه به منظور اعمال تأثیر، نیاز به رابطه‌هایی

1 Administer The Appropriate Intervention
2 Borgatti And Everett

هست که شرایط را برای اعمال قدرت و تأثیر یک گره فراهم می‌آورد. با ذکر مثالی دیگر مفهوم قدرت در سنجه مرکزیت بینابینی مشخص می‌شود. قرار است یک عضو هیأت علمی در محیط دانشگاهی یک دستگاه رایانه تهیه نماید. ایشان برای تهیه رایانه، نیاز به نامه‌نگاری و طی سلسه‌مراتب اداری دارد. در این میان افرادی هستند که ممکن است به درخواست ایشان پاسخ مثبت ندهند؛ بنابراین ممکن است که کانال‌های دیگری برای پیگیری درخواست خود داشته باشد. هر چه تعداد کانال‌های امکان خرید رایانه برای متقاضی بیشتر باشد، میزان قدرت متقاضی در شبکه بیشتر است و میزان وابستگی به یک شخص کمتر.

به‌طور کلی مرکزیت بینابینی، نقطه‌ای است که بینابین بسیاری از جفت نقاط دیگر باشد؛ در واقع نقاطی واسطه‌ای هستند که راه‌های ارتباطی نقاط دیگر از آنها می‌گذرد. این نقاط دارای قدرت ایزوله کردن یا افزایش ارتباطات می‌باشند. مرکزیت بینابینی به‌طور خلاصه عبارت است از: تعداد افرادی در شبکه که یک شخص به‌طور غیرمستقیم از طریق خطوط مستقیم آنها متصل شده است. سنجه مرکزیت بینابینی، توانایی نقش‌آفرینان برای تأثیرگذاری یا کنترل تعامل‌های بین نقش‌آفرینان را نشان می‌دهد. بینابینی به عنوان سنجش تأثیری است که افراد بر روی جریان اطلاعات بین دیگران دارند. افرادی که به عنوان واسطه برای جریان اطلاعات عمل می‌کنند نمرات بینابینی بالائی خواهند داشت. در تصویر ۶-۸ مرکزیت بینابینی در مجله ساینتومتریکس نمایش داده شده است همان‌طوری که مشاهده می‌گردد گلنزل دارای بالاترین مرکزیت بینابینی است و با دایره‌ای بزرگ در وسط تصور مشخص است.

هر چند که قدرت به وسیله مذاکرات و مبادلات مستقیم نشان داده می‌شود، اما قدرت همچنین از طریق عمل کردن به عنوان یک "نقطه مرجع " که سایر نقش‌آفرینان خودشان را به وسیله آن مورد قضاوت قرار می‌دهند و به واسطه مرکز توجه قرار گرفتن توسط افرادی که دیدگاه‌هایشان توسط تعداد زیادی از نقش‌آفرینان شنیده می‌شود، رخ می‌دهد. نقش‌آفرینانی که قادرند در کوتاه‌ترین مسیر به دیگر نقش‌آفرینان دسترسی داشته باشند، یا فردی که توسط دیگر نقش‌آفرینان در کوتاه‌ترین مسیر در دسترس است، موقعیت‌های مطلوبی دارند. این مزیت ساختاری می‌تواند به قدرت تعبیر شود و کسانی که چنین موقعیتی در شبکه دارند قدرتمندتر هستند.

مرکزیت بردار ویژه

مرکزیت بردار ویژه یکی دیگر از سنجه‌های مرکزیت است و بر اساس این ایده پیشنهاد شده است که مرکزیت یک گره خاص نمی‌تواند مجزا از مرکزیت دیگر گره‌هایی که با آن متصل شده است تخمین زده شود. نمرات مرکزیت، به گره‌ها بر اساس این اصل که ارتباط به گره‌های با نمره بالا در نمرات یک گره خاص

نسبت به ارتباط (اتصال) به گره‌های با نمره پایین مشارکت بیشتری دارد، اختصاص داده می‌شود (بوناسیچ[1]، ۱۹۷۲).

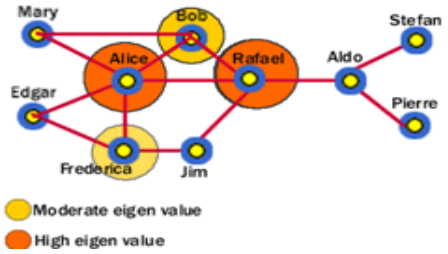
مرکزیت بردار ویژه؛ نقطه‌ای دارای بیشترین مرکزیت بردار ویژه است که دارای همسایگان مرکزی بسیاری باشد، در واقع مرکزیت بردار ویژه بیشتر سبب قدرت بیشتر می‌شود (براندز و الرباچ[2]، ۲۰۰۵) توصیه می‌شود که در زمینه ماتریس‌های ارزش‌دار از مرکزیت‌های عادی (نرمال) شده بهره‌گیری شود که از تقسیم مرکزیت محاسبه شده به بیشینه مرکزیت ممکن به دست می‌آید. در این وضعیت امکان مقایسه کامل‌تر مرکزیت‌ها و نیز آگاهی از نسبت هر اندازه به بیشینه ممکن میسر می‌شود (بورگتی و دیگران، ۲۰۰۲)

بوناسیچ در سال ۱۹۷۲ روشی برای استخراج مرکزیت یک نقش‌آفرین با استفاده از بردار ویژگی (آیگن)، ماتریسی از روابط دوستی یا گزینه‌های دوستی میان مجموعه‌ای از نقش‌آفرینان معرفی نمود. ماتریس روابط دوستی که با این روش سر و کار دارد تصور می‌شود که نظام‌مند است. روش بوناسیچ مبتنی بر بردار ویژگی مطابق با بالاترین ارزش ویژگی است. هر عنصر از بردار ویژگی، مرکزیت هر نقش‌آفرین را نمایش می‌دهد. این روش یک مشخصه خوب دارد که مرکزیت هر نقش‌آفرین به‌طور بازگشتی[3] با استفاده از مجموع وزن دهی شده مرکزیت‌های تمام نقش‌آفرینان تعریف می‌شود، جایی که وزن، قدرت روابط دوستی بین نقش‌آفرین و دیگر نقش‌آفرینان نیست. این روش گسترش داده شد تا با هر ماتریس نظام‌مندی از روابط دوستی سر و کار داشته باشد. جایی که (a) روابط از نقش‌آفرین jبه kبا روابط از نقش‌آفرین k به j یا (b) روابط بین مجموعه‌ای از نقش‌آفرینان و سایر مجموعه نقش‌آفرینان مشابه نیست. مورد اول (a) بدین معناست که داده‌ها یک وجهیِ دو راهه و مورد دوم (b) بدین معناست که داده‌ها دو وجهی و دوراهه هستند. این روش‌ها بردار ویژگی که مطابق با بزرگترین ارزش ویژگی است را مورد استفاده قرار می‌دهد (اوکادا، ۲۰۰۸).

مرکزیت بردار ویژه، مقدار نزدیکی یک موجودیت به موجودیت‌های بسیار نزدیک دیگر در شبکه را اندازه‌گیری می‌کند؛ به عبارت دیگر بردار ویژه مرکزی‌ترین موجودیت‌ها را بر اساس ساختار جهانی و یا کلی شبکه مشخص می‌کند. موجودیتی با بالاترین مرکزیت بردار ویژه به‌طور کلی دارای ویژگی‌های زیر است:

- نقش‌آفرینی را مشخص می‌نماید که به الگوی اصلی فواصل بین تمام موجودیت‌ها، مرکزی‌تر است.
- سنجه‌ای منطقی از یک جنبه از مرکزیت بر حسب مزیت موقعیتی است.

1 Bonacich
2 Brandes & Erlebach
3 Recursively



تصویر ۶-۹.: نمودار شبکه مرکزیت بردار ویژه نقش‌آفرینان در یک شبکه

در تصویر ۶-۹، می‌توانیم مشاهده کنیم که آلیس و رافائل به موجودیت‌های خیلی نزدیک در شبکه، نزدیک‌تر هستند. همچنین باب و فردریکا در مرکز دو گروه خیلی نزدیک هستند، اما این دو در مقایسه با آلیس و رافائل از ارزش کمتری برخوردارند.

در تعریف دیگر آمده است، مرکزیت بردار ویژه نقطه‌ای است که دارای همسایگان مرکزی بسیاری باشد؛ در واقع این مفهوم جایگاه نقاط را در دو بعد نشان می‌دهد: بعد کلی و بعد محلی که البته تأکید بیشتر آن بر اساس ساختار کلی است (محمدی کنگرانی، ۱۳۹۰).

دو سنجه‌ی مرکزیت نزدیکی که در بالا معرفی شد، بر پایه مجموع کوتاه‌ترین فاصله میان یک گره با سایر گره‌ها سنجیده می‌شود. در شبکه‌های بزرگ و پیچیده این احتمال وجود دارد که مخاطب بعضی از اطلاعات را به دلیل حجم زیاد اطلاعات از دست بدهد و یا اینکه قدرت تشخیص گره‌های با مرکزیت بالا را نداشته باشد. به عنوان مثال اگر دو گره "الف " و "ب " در درون یک شبکه بزرگ باشند. در درون شبکه، شرایط گره "الف " به گونه‌ای است که نزدیکی بیشتری با سایر گره‌ها دارد، اما مجموعه‌ای که گره "الف " در آن قرار گرفته، از نظر مرکزیت رتبه، بسیار ضعیف و تعداد گره‌ها موجود در شبکه نیز کم است. گره "ب " دارای فاصله‌ای متوسط از سایر گره‌ها، ولی در یک مجموعه بزرگتر و تعداد گره‌هایی با مرکزیت درجه بالا قرار دارد. سنجش دوری و نزدیکی گره‌ها "الف " و "ب " می‌تواند تقریباً شبیه باشد. ولی در عمل، گره "ب " مرکزیت بیشتری نسبت به گره " الف " دارد. چرا که امکان در دسترس قرار گرفتن بیشتر گره "ب " در شبکه بواسطه ارتباط با گره‌های با مرکزیت رتبه بالا و به عبارتی قدرتمندتر، وجود دارد.

به منظور یافتن گره‌های با مرکزیت بالا، سنجه مرکزیت بردار ویژه تلاش می‌کند، گره‌هایی را مشخص کند که به واسطه ارتباط با گره‌های قدرتمند در شبکه، دارای قدرت می‌شوند. این گره‌ها هرچند در ظاهر ارتباطات کمی دارند، ولی به واسطه ارتباطی که با گره‌های قدرتمند و دارای رتبه بالا برقرار می‌کنند، به عنوان گره‌های قدرتمند محسوب می‌شوند. در تصویر زیر نیکولاس داری بالاترین نمره مرکزیت بردار ویژه را دارد چون به گره‌های قدرتمندی در شبکه نزدیک است. رویکرد مرکزیت بردار ویژه بر اساس اندیشه نزدیکی/

دوری ساخته شده است. تصویر ۶-۱۰ نشان می‌دهد که نیکولاس به سایر گره‌های درون شبکه نزدیک‌تر است و توانایی تأثیرگذاری و نفوذ بیشتری بر سایر افراد درون شبکه دارد.

الگوریتم رتبه صفحه[1]

الگوریتم رتبه صفحه بر اساس فرآیند مسیریابی در وب است که در آن مسیریاب به صورت تصادفی پیوندها را انتخاب می‌کند و نمره رتبه صفحه برای یک صفحه برابر با احتمالی است که موجب انتخاب آن صفحه گردیده است (براین و پیج[2]، ۱۹۹۸).

الگوریتم اچ آی تی اس

الگوریتم اچ‌آی‌تی‌اس صفحات وب را به هاب‌ها و آوتوریتی[3] تقسیم می‌کند. هر صفحه‌ای نمرات هاب و آوتوریتی دارد. نمره آوتوریتی یک صفحه بستگی به نمره‌های هاب مجاور و نمره هاب بستگی به نمره‌های آوتوریتی مجاور دارد، می‌توان این گونه گفت که نمرات مرکزیت وابسته به یکدیگرند (کلینبرگ[4]، ۱۹۹۹، به نقل از لیو و فنگ، ۲۰۰۹).

گستردگی مرکزیت[5]

جوامع مختلف دانش، منابع، روش‌شناسی و اسلوب فکر کردن متفاوتی دارند. پژوهشگری که با افراد دیگری از جوامع مختلف همکاری می‌کند یا با دیگران درون و بیرون از جامعه خودش همکاری می‌کند، می‌تواند دانش و منابع متفاوتشان را به اشتراک بگذارند. برای کل شبکه همکاری همچنین می‌تواند جریان دانش بین آن جوامع در آن حوزه را بهبود بخشد. یک پژوهشگر باید بیشترین تلاشش را به کار گیرد تا همانند این الگو همکاری کند؛ بنابراین گستردگی روابط همکاری یک نویسنده، زمان تحلیل مرکزیت نویسنده باید به حساب آورده شود. لیو و فنگ در پژوهش خود نوع جدیدی از سنجش برای مرکزیت تحت عنوان گستردگی مرکزیت پیشنهاد نمودند که در آن توزیع روابط همکاری یک نویسنده به حساب می‌آید. گستردگی مرکزیت

1 Pagerank Algorithms

2 Brin & Page

3 Authorities

4 Kleinberg

5 Extensity Centrality

بر سهم همکاری نویسندگان در میان جوامع یا گستردگی روابط همکاری تأکید دارد در حالی که رتبه تعداد همکاری‌های یک نویسنده را نمایش می‌دهد (لیو و فنگ، ۲۰۰۹).

مرکزیت اطلاعات

علاوه بر سنجه‌های کلاسیک مرکزیت، نوعی سنجه مرکزیت که از نظریه اطلاعات کلود شانون مشتق می‌شود نیز پیشنهاد شده است. بر اساس این نوع سنجه، مرکزیت یک فرد به توزیع احتمالاتِ جریانی که از افراد شروع و در هر کدام از گره‌ها در شبکه متوقف می‌گردد، مربوط است (توتزائور[1]، ۲۰۰۷).

مرکزیت اطلاعات به‌وسیله استفن سون و زلن[2] (۱۹۸۹) به عنوان سنجش مرکزیت گره‌ها در شبکه‌های اجتماعی مطرح شد که برگرفته از نظریه انتقال آماری اطلاعات شانون و ویور است. مرکزیت بینابینی بر اساس کوتاه‌ترین فاصله میان دو گره محاسبه می‌شود، در حالی که سنجه مرکزیت اطلاعات به بررسی و مطالعه این امر می‌پردازد که ممکن است اطلاعات و دانش از طریق مسیرهای متفاوتی انتقال یابد. این نوع سنجه بر اساس شدت و قدرت گره‌ها و فاصله میان آنها محاسبه می‌شود. به عبارتی این سنجه میزان انتقال اطلاعاتی که می‌تواند بین دو نقطه در شبکه انتقال یابد را مورد بررسی قرار می‌دهد.

مرکزیت رتبه نویسنده

مرکزیت رتبه نویسنده[3]، این سنجه در حوزه کتابخانه‌های دیجیتالی توسط لیو و دیگران پیشنهاد شده است. بر اساس ایده الگوریتم اچ آی تی اس، شخص می‌تواند بین اولین نویسنده (رهبر) و دیگر نویسندگان (پیروان) تفاوت قائل گردد و سپس اهمیت نویسندگان به عنوان دو نقش را به ترتیب تحلیل نماید (یوشیکانه[4] و دیگران، ۲۰۰۶).

عدد اردوش

یکی از پر تولیدترین ریاضیدانان تمام دوره‌ها پائول اردوش[5] ریاضی‌دان مجارستانی است که بیش از ۱۴۰۰ مقاله با بیش از ۵۰۰ نفر هم نویسنده، نوشته است. این بهره‌وری بی‌نظیر مفهوم عدد اردوش را القاء کرد که وی به خود نمره صفر داده است، هر کسی که با او هم نویسنده شده است عدد یک گرفته و کسی که با نویسنده همکار او و نه خود او مقاله نوشته است نمره ۳ گرفته است تا آخر.

سنجه سالتون

1 Tutzauer
2 Stephenson And Zelen
3 Author Rank Centrality
4 Yoshikane
5 Paul Erd˝os

در سنجش استحکام همکاری بین نواحی، سنجه سالتون[1] یکی از مشهورترین روش‌ها است (سالتون و مک گیل[2]،۱۹۸۳). این سنجه بر اساس تعداد مقالات cij که توسط دو ناحیه iو j هم نویسنده شده‌اند، است. تعداد مقالاتci ناحیه i و تعداد مقالات cj ناحیه j است. استحکام همکاری دو ناحیه عبارت است از:

$$scj = cij / \sqrt{ci \times cj}$$

لیو و فنگ برای سنجش استحکام گره‌های همکاری بین نویسندگان سنجه سالتون منطقی[3] را بکار بردند. استحکام rijبه صورت زیر سنجیده می‌شود:

$$rij = hij / \sqrt{hi \times hj}$$

جایی که hij تعداد مقالات هم نویسنده به‌وسیله نویسنده i و نویسنده j وhj و hj تعداد مقالات آنها به ترتیب است (لیو و فنگ، ۲۰۰۹).

کاربرد خوشه بندی در بازیابی اطلاعات

مقدمه

فن خوشه بندی داده‌ها

فن را می‌توان ترفندی نامید که در عمل قابل پیاده سازی است. با این نگاه، فراگیری فن خوشه‌بندی در عمل منجر به فراگیری یک ترفند برای دسته‌بندی اشیاء مشابه می‌شود.

بسیاری از امور را می‌توان با ترفندهایی از طریق رایانه به صورت خودکار انجام داد که دسته‌بندی اشیاء نمونه‌ای از آن است. مثلاً، وقتی می‌خواهیم به روش خودکار به یک مدرک موضوع بدهیم، تعداد تکرار واژه‌ها را در متن آن تعیین می‌کنیم و محاسبات لازم را روی فراوانی واژه‌ها انجام می‌دهیم تا نهایتاً موضوعات مدرک را استخراج کنیم. در برابر، در روش غیر خود کار(دستی)، متن را می‌خوانیم و موضوعاتی که به ذهنمان متبادر می‌شود، به عنوان کلید واژه برمی‌گزینیم.

از مثال بالا نتیجه می‌گیریم که روش دستی با تحلیل و تشخیص ذهنی صورت می‌گیرد ولی روش خودکار با محاسبات ریاضی و آماری در ارتباط است. بنابراین، در روش خودکار ترفندها بیشتر در این جهت است که

1 Salton
2 Mcgill
3 Reasonable Salton's Measure

چگونه تعدادی داده قابل محاسبه به دست آوریم و چه محاسبات آماری و ریاضی روی آن‌ها انجام دهیم تا به نتیجه مورد دلخواه برسیم. خوشبختانه، امروزه برای استفاده از این ترفندها لازم نیست که متخصص ریاضی و آمار باشیم چون با گسترش فناوری اطلاعات و ارتباطات، سخت افزارها و نرم افزارهای فراوانی به کمک آمده-اند. نرم افزارهای آماری مانند SPSS نمونه‌ای از آن هستند.

برای درک و فراگیری فن خوشه بندی لازم است حداقل با چند موضوع آشنا شد که از مباحث بنیادی در این زمینه به شما می‌آیند:

- داده کاوی[1]
- مفهوم شیئ[2] و داده[3]
- تشکیل ماتریس متقارن[4]
- تعیین فاصله بردارها[5] (
- دسته بندی اشیاء[6]

مقدمه ای بر داده کاوی

پیش درآمد

امروزه حجم داده‌ها و اطلاعاتی که توسط افراد و سازمان‌های مختلف تولید می‌شود افزایش فوق العاده‌ای یافته است. به تعبیری انفجار اطلاعات[7] رخ داده است. طبق برخی از گزارش‌ها به عنوان مثال هر ساله ۱,۵ بیلیون گیگابایت اطلاعات در مخازن مختلف اطلاعاتی از قبیل کتابخانه‌ها، اینترانت سازمان‌ها، آژانس‌های خبری، اینترنت و رایانه‌های شخصی جمع آوری می‌گردد (دارفی، ۲۰۰۸)[8]. این مسئله اگر چه به سر ریز اطلاعات[9] در در حوزه‌های مختلف منجر شده، اما واقعیت این است که بسیاری از این اطلاعات غیر معتبر، تکراری یا فاقد

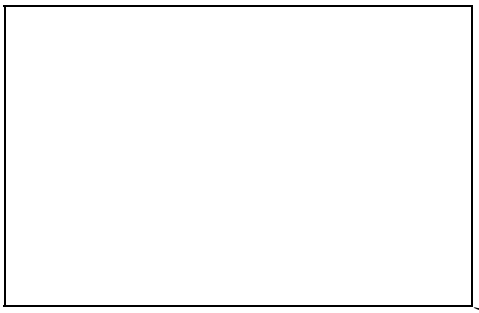
[1] datamining
[2] object
[3] data
[4] proximity matrix
[5] vector spaces
[6] object segmentation
[7] Information explosion
[8] Durfee, 2008
[9] Information overload

ارزش اطلاعاتی است که از این پدیده به عنوان آلودگی اطلاعات[1] یاد می‌شود. انفجار، سرریز و آلودگی اطلاعات بر مسائل مربوط به اطلاعات و اطلاع رسانی دامن زده و بازیابی اطلاعات مناسب، در زمان مناسب برای استفاده‌کنندگان را بیش از پیش دشوار نموده است. از اینرو، در عصر اطلاعات نیز می‌بینیم که خیلی از تصمیم گیری‌ها همچنان در فقر اطلاعاتی اتخاذ می‌شود. در برخورد با این مشکلات در چند دهه‌ی گذشته تلاش‌های زیادی در جهت توسعه روش‌ها و ابزارهای نوین ذخیره و بازیابی اطلاعات صورت گرفته و بانک-های اطلاعاتی قدرتمندی با قابلیت‌های جستجوی فوق العاده تولید و روانه بازار شده است. با این همه، در حالی‌که پیشرفت فناوری امکان جمع آوری، ذخیره و انتقال داده‌ها در حجم زیاد را برای ما فراهم ساخته است، توانایی ما در پردازش و استنباط داده‌ها به آن حد نرسیده که بتوانیم از داده‌های موجود به شکل مطلوب استفاده کنیم. در سال‌های اخیر، برای اینکه بتوان از قابلیت‌های اطلاعات موجود به خوبی بهره برداری نمود، علاوه بر مسئله ذخیره و بازیابی اطلاعات، مسئله تبدیل اطلاعات به دانش با استفاده از الگوها و روش‌های مختلف کشف و استخراج دانش نیز، مورد توجه قرار گرفته است. در این راستا، متخصصان اطلاع رسانی به تدریج بیشتر بر فناوری‌های نوین در دستیابی، تحلیل، خلاصه کردن و تفسیر هوشمندانه اطلاعات متکی شده‌اند و مطالعات داده کاوی اهمیت بسیاری پیدا کرده است.

در فرآیند داده کاوی، خروجی مورد نظر دانش است. از این‌رو برخی از صاحب نظران اصطلاح دانش کاوی[2] را را صحیح‌تر از داده‌کاوی خوانده‌اند و تصریح می‌دارند، همان‌گونه که در استخراج طلا، آنچه که کاوش می‌شود، طلاست، نه سنگ طلا، فرآیند داده کاوی نیز که به کاوش دانش می‌پردازد، باید دانش کاوی نامیده شود (غضنفری و همکاران،۱۳۸۷). در هر حال دانش کاوی یا داده کاوی، ابزاری مفید در کشف و استخراج دانش و کمکی مؤثر در سیر تکامل داده به اطلاعات و دانش است. در بررسی هرم دانایی، حداقل سه جزء اساسی قابل ملاحظه است (شکل ۱). در روشن ساختن مفهوم هر یک از این عناصر می‌توان گفت که داده همان حقایق، مشاهدات و برداشت‌های پردازش نشده است که به صورت اعداد خام یا اظهارات بدون هدف وجود دارد. مثلاً سن و جنس تماشاچیان یک مسابقه می‌تواند به عنوان داده محسوب شود. اما، داده زمانی که در متن قرار گیرد و معنی پیدا کند تبدیل به اطلاعات می‌شود. به بیان دیگر اطلاعات همان داده‌ی پردازش شده است که به فرم قابل استفاده درآمده است. مثلاً زمانی که داده‌ی خام مربوط به فروش یک محصول و داده‌ی مربوط به کد پستی

[1] Information pollution
[2]

خریداران در ارتباط با یکدیگر بررسی شود، اطلاعاتی از وضعیت فروش جغرافیایی محصول به دست می‌آید. دانش از ترکیب تجربه و دانسته‌های قبلی فرد با اطلاعات موجود به دست می‌آید. به بیان دقیق‌تر دانش از تفسیر اطلاعات و کاربرد آن در یک بافت جدید حاصل می‌شود. بنابراین اطلاعات مفید زمانی می‌تواند به عنوان دانش محسوب شود که بتواند در حوزه عمل منشاء اثر قرار گیرد. به گفته کوچن[1] (۱۹۷۴) دانش، تجربه‌ای است که فرد در اثر تعامل با محیط به دست می‌آورد و عامل برانگیختن رفتار فرد است. با این تفاسیر، کشف دانش به این معنی است که فرد بتواند اطلاعات مرتبط را شناسایی کند، بداند چه طور این اطلاعات را با دانش قبلی خود تلفیق کند و بتواند دانش خود را به گونه‌ای تغییر دهد که از آن در حل مشکل کمک بگیرد. در مثال قبلی در صورتی دانش خلق می‌شود که اطلاعات به دست آمده از وضعیت فروش جغرافیایی یک محصول با اطلاعات دیگر از آن منطقه جغرافیایی از قبیل تغییرات جمعیتی و غیره بررسی شود و اطلاعات جدید به دست آمده به همراه دانش قبلی فرد بتواند در تصمیم گیری‌های مربوط به توزیع محصول منشاء اثر قرار گیرد.



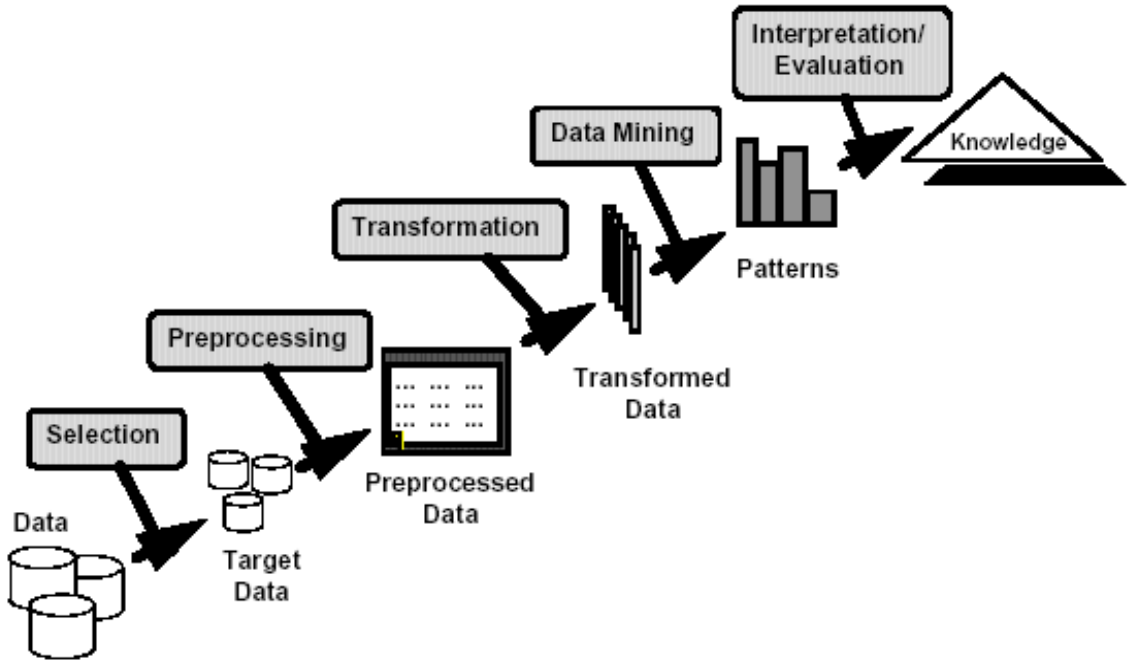
اطلاعات= داده پردازش شده دانش= اطلاعات+تفکر

شکل ۱: عناصر اساسی هرم دانایی

در سیر تکاملی داده به اطلاعات و دانش، ملاحظه می شود که چنانچه اطلاعات تلفیق شود و روابط میان آنها شناخته شود، درک یا تصویر ذهنی جدیدی از موضوع شکل می گیرد. در این رابطه داده کاوی به عنوان ابزاری مفید در فرایند کشف دانش از داده [2] مورد استفاده قرار می گیرد و به استخراج اطلاعات بالقوه مفیدی که در بین حجم انبوهی از داده ها نهان شده است کمک می کند. شکل زیر جایگاه داده کاوی در فرایند تبدیل داده به دانش را نشان می دهد (شکل ۲). همانگونه که در شکل نشان داده شده است، مجموعه های مورد نظر داده، پس از طی یکسری مراحل مقدماتی که به منظور پردازش و آماده سازی داده ها صورت می گیرد، وارد فرایند داده کاوی می شوند. در این مرحله با بکارگیری روشهای مناسب داده کاوی و انجام تجزیه و تحلیلهای لازم

[1] Kochen, 1974
[2] Knowledge Discovery from Data (KDD)

الگوها و مدلهای مفیدی که در دل داده ها مستتر شده اند کشف شده و مورد تفسیر و ارزیابی قرار می گیرند تا به این وسیله به ایجاد دانش منجر شوند.



شکل ۲: جایگاه داده کاوی در فرایند تبدیل داده به دانش

دانش بدست آمده از طریق این فرایند به افراد و سازمان ها کمک می کند تا بتوانند تصمیم گیری های حیاتی را سریع تر و با میزان اطمینان بیشتر داشته باشند.

تعریف داده کاوی: داده کاوی کاوش در حجم عظیم داده ها به منظور یافتن الگوها[1] و مدلهای موجود در بین داده هاست. داده کاوی به عبارتی دقیق تر، جستجو برای اطلاعات ارزشمند در حجم زیاد داده هاست که می تواند روابط پنهان[2]، الگوها، و وابستگی متقابل[3] داده ها را کشف کند و قواعد تعمیم پذیری در پیش بینی همبستگی ها[4] بدست دهد (هند، مانیلا و اسمیت، ۲۰۰۱؛ گرگنو و راجد،۱۹۹۹)[5] . منظور از مدل در این تعاریف" یک تصویر کلی[6] از ساختاری است که روابط سیستماتیک میان داده ها را بیان می کند. و منظور از

[1] pattern
[2] Hidden relationships
[3] Interdependencies
[4] Correlations
[5] Hand, D. J., Mannila, H., & Smyth, P. , 2001; **Gargano, M. L., & Ragged, B. G., 1999**
[6] Global representation

الگو ساختاری محلی است که فقط به چند متغیر محدود و تعدادی مشاهده مرتبط است" (غضنفری و همکاران،۱۳۸۷،ص.۹).

داده کاوی در فرم های مختلف اطلاعات:

امروزه، اطلاعات به شکل های مختلف شامل اطلاعات متنی، تصویری، صوتی، و ویدئوئی و در قالب رسانه های متعدد – اعم از منابع چاپی و الکترونیکی به بازار عرضه می شود. این تنوع در شکل و محمل های اطلاعاتی به پیدایش مفاهیم مختلفی از جمله متن کاوی، وب کاوی، و کاوش در چند رسانه ای ها منتهی شده است.

متن کاوی به علوم مختلف از جمله آمار، زبانشناسی، پردازش زبان طبیعی، تحلیل متن، یادگیری ماشینی، بازیابی اطلاعات، و کتابداری و اطلاع رسانی مربوط می شود. متن متداول ترین و راحت ترین وسیله برای انتقال مطلب از نویسنده به خواننده است. تحقیقات نشان داده است که بخش عظیمی از اطلاعات سازمان ها (۸۰٪) به شکل متنی (تن، ۱۹۹۹)[1] و در قالب مدارک و رسانه های مختلف از جمله صفحات وب سازمان ها، مدارک داخل و خارج سازمانی، گزارشات فنی و مالی، بازخورد مشتریان، مکاتبات و پستهای الکترونیکی، آگهی ها، مقالات ، کتاب ها و کتابخانه های الکترونیکی و غیره عرضه می شوند. **متن کاوی**[2] که اغلب به عنوان داده کاوی و گاه به عنوان بخشی از آن یاد می شود، به کاوش دراطلاعات متنی و فاقد ساختار اطلاق می شود و به دنبال کشف دانش و الگوهای مفید از بین این داده هاست. ابزارهای متن کاوی در پی تحلیل ودرک مفاهیم نهفته در دل اطلاعات و استخراج دانش و الگوهای مفید از آن می باشند. متن کاوی به بیان دقیق تر، به تحلیل، تفسیر و استنتاج معنی از اطلاعات متنی می پردازد و ابزار مهمی در کشف الگوهای مستتر در دل پایگاههای اطلاعات متنی است که برای مقاصد مختلف می تواند مفید باشد.

در حالیکه، کشف الگوهای ناشناخته از داده های عددی کار دشواری نیست، آنچه که در متن کاوی مشکل آفرین است این است که اطلاعات متنی ساختار تلویحی و پیچیده ای دارد و به دلیل ماهیت خاص خود می

[1] Tan, 1999
[2] Text Mining

تواند معانی و تفاسیر متعددی داشته باشد. پیچیدگی ساختار اطلاعات متنی به گونه ای است که گاه ممکن است حتی نویسنده ی یک متن از برداشت ها و تفاسیری که می تواند از متن صورت گیرد، اطلاعی نداشته باشد. تفسیر تازه از اطلاعات موجود به این دلیل حاصل می شود که خوانندگان مختلف، درک متفاوتی از یک متن بر اساس پیش زمینه خود خواهند داشت.

البته، متن کاوی همانگونه که ویتن و همکارانش (۱۹۹۸)[1] تصریح داشتند، این پتانسیل را دارد که فرد را در استخراج اطلاعات مفید از متن بدون نیاز به فهم کل آن کمک نماید. در واقع، ابزارهای متن کاوی با استخراج اطلاعات کلیدی از متن و با شناسایی روابط میان مدارک به کشف دانش و الگوهای مفید از اطلاعات متنی کمک می کنند و در خلاصه برداری ، اولویت بندی، و درک مفاهیم مدارک بدون نیاز به مطالعه کامل آن ها کمک می کنند.

شناسایی مشخصه های مدارک اولین گام در فرایند متن کاوی است. این مشخصه ها ممکن است داخلی (مربوط به محتوای مدرک) باشند، یا ممکن است خارجی (مثلا در مورد نویسنده، تاریخ نشر، فرمت و غیره) باشند. دربکارگیری خوشه بندی، که در ادامه به عنوان یکی از روش های داده کاوی معرفی خواهد شد، عموما مشخصه های داخلی مدارک، با هدف کشف دانش، مد نظر قرار می گیرند. به این منظور، در روش های سنتی و در اکثر مطالعات متن کاوی از کلمات یا کلید واژه هایی که به طور اتوماتیک از متن استخراج می شدند استفاده شده است. این روش استخراج کلیدواژه ها، بعدها توسط برخی از صاحب نظران مورد انتقاد قرار گرفت. در این رابطه به عنوان مثال، استینباچ و همکاران(۲۰۰۰)[2] به نا کارآمدی این شیوه به دلیل اینکه مبتنی بر شباهت سنجی مدارک بر اساس کلمات مشترکشان است اشاره می کنند و معتقدهستند که کلماتی که از مدرک بیرون کشیده شده اند تنها بخشی از مدرک هستند. بعلاوه در استفاده از کلید واژه های گرفته شده از متن عده زیادی نیز به مسئله بروز اشتباهات مفهومی به دلیل مسائلی که از اصول ساختاری زبان نشات می گیرد اشاره کرده اند (دارفی،۲۰۰۸؛ فرناس و همکاران،۱۹۸۷، چن،۱۹۹۶،۱۹۹۴)[3]. در نوشتن یک متن اصول ساختاری زبان در انتخاب کلمات[4]، ساخت جملات دستوری[1]، و بیان مفاهیم[2] تاثیر گذارند. از اینرو، سه جزء

[1] Witten et al.,1998
[2] Steinbach, Karypis and Kumar, 2000
[3] Durfee, 2008; Furnas, et al. 1987; Chen, 1994,1996
[4] Morphology of language

متن، یعنی کاربرد کلمات[۳]، ساختار دستوری [۴]و محتوا[۵] در زبان های مختلف با هم متفاوتند. نویسندگان و خوانندگان متن ممکن است از کلمات مختلف برای بیان مفاهیم یکسان استفاده کنند (مترادفها)[۶] یا از کلمات یکسانی که معانی متعدد دارند برای بیان مفاهیم مختلف استفاده کنند[۷]. مثلا کلمه شیر می تواند در یک جا به معنی شیر خوراکی، در جای دیگر به معنی شیر جنگل و در متنی دیگر به معنی شیر آب باشد. این ویژگیهای زبان طبیعی، که در واقع واحد اصلی سازنده متن محسوب می شود، نه تنها در ۸۰ تا۹۰ درصد موارد به شکست در برقراری ارتباط منتهی می شود (فرناس و همکاران،۱۹۸۷)[۸]، بلکه فن آوریهای متن کاوی را نیز با مشکلاتی مواجه می سازد.

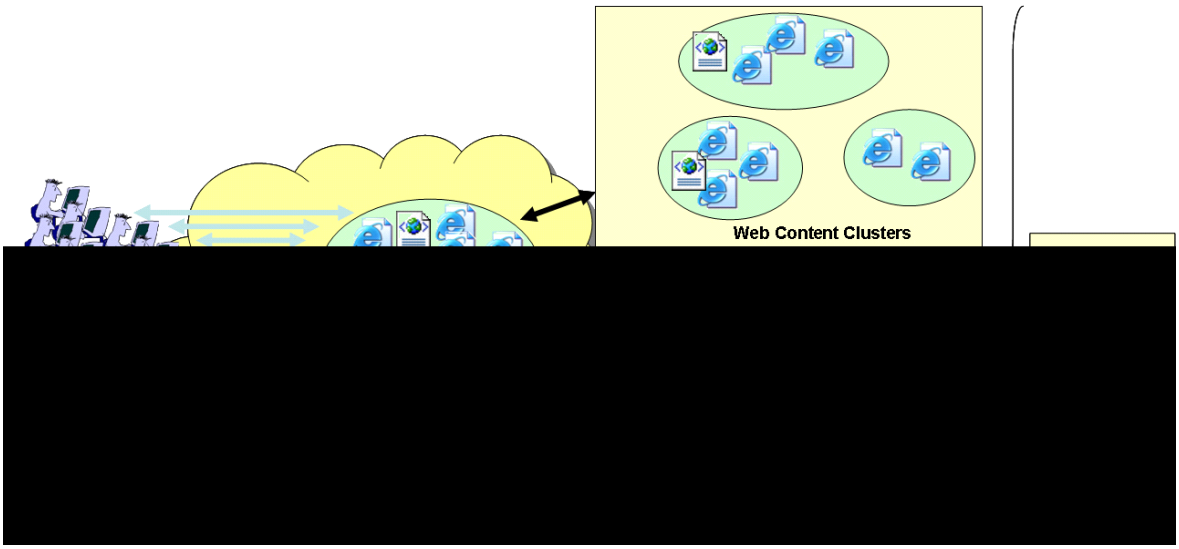
در برخورد با این مسئله استفاده از مفاهیم برای نشان دادن محتوای مدارک به جای کلید واژه ها پیشنهاد شده است (وایوز و همکاران،۲۰۰۸ ، ص. ۲۳۱-۲۳۰)[۹]. امروزه نیز متن کاوی بر پایه مفاهیم استخراج شده از متن و نیز سایر مشخصه های متن نظیر نویسنده، تاریخ نشر و ویرایش، و.. استوار است.

وب کاوی: وب کاوی همان داده کاوی بر روی گروه های استفاده کنندگان و انواع اطلاعات موجود در وب است. دسته بندی گروه های مختلف مراجعان به درک بهتر رفتار اطلاع یابی جستجوگران وب و در نتیجه به بهبود خدمات رسانی و شخصی سازی محتوای وب کمک می کند. کتابخانه های دیجیتالی، به عنوان مثال در یکی دو دهه ی گذشته از وب کاوی در خوشه بندی استفاده کنندگان (سالیس و همکاران،۱۹۹۹)[۱۰] و شخصی سازی خدمات مرجع الکترونیکی (چو،۲۰۰۰)[۱۱] سود جسته اند. از سوی دیگر، در حالیکه موتورهای جستجوی قدیمی قادر به ارائه کمک محدودی در بازیابی اطلاعات به استفاده کنندگان بودند، تکنیکهای داده کاوی توانست به دسته بندی مدارک مرتبط موجود در وب و بهبود فرایند بازیابی اطلاعات کمک کند و موتورهای جستجو را با ویژگی های پیشرفته تری همراه سازد. از کاربردهای داده کاوی در این حوزه به عنوان مثال می

---

[1] Syntax
[2] semantics
[3] Word usage
[4] Grammatical construction
[5] Content
[6] Synonymy
[7] Polysemy
[8] Furnas et al., 1987
[9] **Wives**, 2008
[10] Sallis, et al., 1999
[11] Chau,2000

توان به رتبه بندی صفحات وب[1] وافزایش ضریب دقت در بازیابی اطلاعات[2] ، و تحلیل های هم استنادی[3] اشاره کرد.



شکل ۳- خوشه بندی اطلاعات روی وب

رتبه بندی صفحات وب و افزایش ضریب دقت در بازیابی اطلاعات: موتورهای جستجوی وب برای بازیابی اطلاعات مرتبط با درخواست اطلاعاتی جستجوگر اساسا به رتبه بندی صفحات وب با استفاده از فنون مختلف می پردازند. به عنوان مثال یکی از روش های رایج در این زمینه بررسی مکان- بسامد[4] کلید واژه های مورد جستجو است. بدین معنی که بر اساس مکان قرار گرفتن کلید واژه/های مورد جستجو در صفحات وب و با توجه به میزان تکرار کلید واژه/ها در هر صفحه نتایج حاصل از بازیابی را رتبه بندی و مرتب می کنند. بدین ترتیب صفحاتی که حاوی کلید واژه/های مورد جستجو در عنوان یا در بالای صفحه یا در آغاز پاراگراف ها هستند و صفحاتی که کلیدواژه/های مورد جستجو در آن ها بیشتر تکرار شده اند، رتبه بالاتری دریافت می کنند. در روش دیگر، با تحلیل اتصال درونی[5] یکسری صفحات مرتبط، می توان صفحات وب را رتبه بندی نموده و از این رتبه بندی در بازیابی صفحات مربوط به موضوع مورد جستجو استفاده کرد. در رتبه بندی صفحات به این روش، اهمیت یک صفحه بر اساس تعداد صفحاتی که به آن پیوند می دهند محاسبه می شود.

[1] Ranking of pages
[2] Precision
[3] Co-Citation
[4] Location-Frequency
[5] Interconnection

از این شیوه به عنوان مثال در موتور جستجوی گوگل استفاده می شود. در رتبه بندی صفحات وب با این تکنیک به محتوای متن توجهی نمی شود و در نتیجه ضریب دقت در بازیابی اطلاعات پایین می آید. برای حل این مسئله توصیه می شود که مجموعه صفحات بزرگ به واحدهای کوچک تر شکسته شود و سپس میزان مرتبط بودن آن ها محاسبه شود. در این رابطه زنگ و دانگ (۲۰۰۲)[1] یک موتور جستجوی تصویر به نام ای-فایند[2] را با استفاده از تحلیل ماتریس در لاگ موتورهای جستجو [3] طراحی کردند که در آن از ایده کوری-لاگ[4] برای پیدا کردن روابط بین استفاده کنندگان، درخواستهای اطلاعاتی و نتایجی که استفاده کنندگان بر آن کلیک می کنند استفاده شده است.

تحلیل هم استنادی: مطالعات علم سنجی به بررسی قواعد حاکم بر روابط نشریات، نویسندگان و مقالات می پردازد. تحلیل استنادی تکنیکی است که در این رابطه بسیار کاربرد دارد. از تکنیک های تحلیل استنادی می توان برای شناسایی مدارک مشابه و مرتبط در موضوعات گوناگون و نویسندگانی که در یک حوزه یا حوزه های مشابه کار می کنند، استفاده نمود. به طور کلی تحلیل استنادی شامل تحلیل استنادی عمودی و تحلیل استنادی افقی می باشد. در تحلیل استنادی عمودی، قواعد حاکم بر روابط میان یک مدرک و استناد های آن مورد بررسی قرار می گیرد. تحلیل استنادی افقی بر این فرض استوار است که میان مدارکی که استنادهای مشترک دارند، به نوعی ارتباط موضوعی وجود دارد. تکنیک های مورد استفاده در این قبیل مطالعات می توانند به دو دسته تقسیم شوند:

۱- اشتراک در مأخذ یا زوج های کتابشناختی[5]: این اندیشه که نخستین بار توسط کسلر[6] (۱۹۶۳ مطرح) شد، بر این فرض استوار است که مقاله هایی که مآخذ مشترک بیشتری دارند، به احتمال زیاد شباهت موضوعی بیشتری نسبت به مقاله هایی که اشتراک مأخذ ندارند، دارا هستند.

۲- اشتراک در متن یا هم استنادی[7] : این ایده که نخستین بار از سوی هنری اسمال[8] (۱۹۷۳) عنوان شد، بر بر این فرض است که اگر دو یا چند مدرک هم استناد باشند، یعنی در تعدادی از مقاله های منتشر شده ی پس از خود با هم به آنها استناد شده باشد، می توان گفت که آن ها با هم اشتراک در متن یا به

[1] Zhang & Dong , 2002
[2] eefind
[3] Matrix Analysis on Search Engine Log (MASEL)
[4] query log
[5] Co-references or Bibliographic coupling
[6] M.M. Kessler
[7] Co-citation
[8] Small, 1973

عبارتی ارتباط موضوعی دارند. تحلیل هم استنادی به عنوان روشی در تحلیل ساختار فکری مطالعات علمی در وب به وفور مورد استفاده قرار گرفته است. این شیوه می تواند برای شناسایی نویسندگان حوزه های تحقیقاتی یکسان یا مشابه بکار گرفته شود (چن و لیو، ۲۰۰۹)[۱]. به عنوان نمونه چن و پل (۲۰۰۳)[۲] از مصور سازی عینی و طرح مجازی سه بعدی[۳] برای انجام این قبیل مطالعات تحلیلی استفاده استفاده کردند تا ساختار هم استنادی نویسندگان را نشان دهند. بر اساس یافته های آنان نویسندگانی که حوزه تاثیر وسیع تری دارند در اطراف مرکز ساختار فکری قرار می گیرند و نویسندگانی که در یک حوزه تخصصی خاص فعالیت دارند در اطراف این ساختار واقع می شوند.

کاوش در چند رسانه ای ها: فایلهای چندرسانه ای شامل شکل های مختلف اطلاعات از جمله متن، تصاویر ثابت و متحرک، صدا، ویدئو و غیره است. در حالیکه تحقیقات اولیه در زمینه داده کاوی بیشتر به کاوش متن پرداخته اند، کاوش در سایر شکلهای اطلاعات در آینده بسیار جای کار دارد. کاوش در چند رسانه ای ها بویژه در تحلیل اطلاعات ساختار نیافته ای که به صورت پیوسته در وب در دسترس هستند، مفید است و می تواند کمک بزرگی نیز در کاوش نسخ خطی باشد.

کاربرد داده کاوی در حوزه های مختلف علمی: از داده کاوی در حوزه های مختلف از جمله در پزشکی، علوم زیستی، ورزش، تجارت ، کتابداری و اطلاع رسانی و غیره استفاده می شود. تقریبا در تمام رشته ها و انواع سازمان ها، به دلیل وجود اطلاعات، می توان داده کاوی را مورد استفاده قرار داد. از کاربردهای معمول تجاری به عنوان مثال می توان به شناسایی تخلف[۴] از جمله تشخیص سوء استفاده از کارتهای اعتباری اشاره کرد. در این زمینه با برچسب زدن بر معاملات گذشته و شناسایی یک مدل خاص برای یک دسته از معاملات و با استفاده از اطلاعات معاملات کارتهای اعتباری و اطلاعات دارنده کارت می توان موارد کلاهبر داری را شناسایی کرد. کاربرد داده کاوی در هدایت یا تقسیم بازار[۵] مثال دیگری در این زمینه است. که در آن با جمع آوری مشخصات مختلف مصرف کنندگان بر پایه ی موقعیت جغرافیایی آنان و با شناسایی گروه مصرف کنندگان مشابه و اندازه گیری کیفیت گروه ها با توجه به الگوهای خرید مصرف کنندگان هر گروه می توان به تقسیم بازار به زیر مجموعه های مستقل مبادرت نمود. تحلیل اطلاعات مشتریان یک سازمان و شناسائی

[1] Chen and Liu, 2009
[2] Chen and Paul, 2003
[3] 3D virtual landscape
[4] Fraud Detection
[5] Market Segmentation

طبقات و گروههای اصلی مشتریان، ، تحلیل سبد خرید، و تعیین میزان تاثیر عوامل مختلف نظیر تبلیغات و تخفیف بر میزان و الگوهای فروش از دیگر کاربردهای داده کاوی در حوزه تجارت است. استفاده از داده کاوی در تحلیل داده های مربوط به کتابخانه ها، کتاب کاوی[1] خوانده می شود. کاوش در اطلاعات کتابشناختی منابع موجود در کتابخانه و منابع در دست سفارش، اطلاعات مربوط به کارگزاران، ناشران و معاملات قبلی ، اطلاعات مربوط به کارکنان و استفاده کنندگان، امانت و استفاده از منابع در محل، امانت بین کتابخانه ای، مذاکرات مرجع، جستجو و تورق در منابع، و نظایر آن می تواند کتابخانه ها را در مدیریت کارآمد مجموعه و تامین شایسته نیازهای اطلاعاتی کاربران و ارائه خدمات موثر یاری رساند. یافته های حاصل از کتاب کاوی می تواند پشتوانه ای برای تصمیم گیری های مدیریتی باشد و مدیر کتابخانه را در کلیه امور مدیریتی شامل مدیریت امور مالی؛ از جمله در دفاع از طرح بودجه و صرفه جویی در هزینه ها، در گسترش مجموعه و ایجاد یک سیستم خبره جهت استفاده در فراهم آوری منابع و انتخاب کارگزاران مناسب ، در مدیریت استراتژیک و پشتیبانی از خط مشی های اتخاذ شده ، در مدیریت نیروی انسانی و کشف الگوهای رفتاری کارکنان، و نیز در مدیریت مجموعه و کشف الگوهای رفتاری استفاده کنندگان، و الگوهای مربوط به استفاده از منابع توانمند سازد. به عنوان مثال کشف الگوهای بدست آمده از پرسش و پاسخهای مرجع، می تواند کتابخانه را در فراهم نمودن پاسخی فوری به پرسشهای بعدی یاری رساند و ارتباط مناسب بین پرسش های آتی و متخصصان برقرار نماید. دسته بندی استفاده کنندگان بر اساس منابعی که به امانت گرفته اند و افزودن اطلاعات دموگرافیک به هر دسته مدلهایی را در اختیار قرار می دهد که به کشف علائق و دانش جامعه استفاده کننده کمک می نماید و بررسی داده ها ی مربوط به استفاده های ناموفق می تواند نشانگر موارد ضعف مجموعه باشد (نیکلسون و ستنتن،۲۰۰۵). در بررسی روشهای مختلف داده کاوی، در ادامه مثال های بیشتری از کاربرد داده کاوی در سایر رشته ها آورده شده است.

عملکرد/روش های داده کاوی:

تکنیکهای متعددی برای داده کاوی وجود دارد که با موفقیت در زمینه های مختلف به کار گرفته شده است. این تکنیکها را می توان به طور کلی به دو گروه کلی توصیفی و پیشبینانه [2]تقسیم کرد. روشهای توصیفی به یافتن

[1] Bibliomining
[2] Descriptive and Predictive

الگوهایی که برای انسان قابل تفسیر باشد می پردازد و روشهای پیشبینانه برای پیش بینی موارد ناشناخته بکار گرفته می شود. خوشه بندی و قوانین تلازمی نمونه هایی از تکنیک توصیفی و دسته بندی نمونه ای از روش پیشبینانه است که در ادامه به آن پرداخته خواهد شد. ناگفته پیداست که، نوع دانش مورد انتظار، تکنیک داده کاوی مورد استفاده را مشخص خواهد کرد.

یادگیری ماشینی[1]:

از داده کاوی می توان برای انجام وظایف مختلف از جمله "یادگیری ماشینی" استفاده کرد. یادگیری ماشینی به مطالعه و طراحی گروهی از برنامه های کامپیوتری گفته می شود که می توانند موارد(الگوها) ، و قواعد یا مقررات را از مشاهدات گذشته یاد بگیرند و بر اساس نوع دانشی که کشف می شود می توانند به دو حوزه وسیع " یادگیری با ناظر انسانی [2]و "یادگیری بدون ناظر انسانی[3] تقسیم شوند. در تکنیک با ناظر، داده ها باید از پیش دسته بندی شده باشند. در این شیوه هرآیتم[4] با یک برچسب منحصر به خودش همراه است که دسته ای که آن آیتم به آن تعلق دارد را نشان می دهد. بعنوان مثال دسته بندی خبر ها به دسته های مختلف از پیش تعیین شده از قبیل خبرهای ورزشی، خبرهای فرهنگی و غیره و سپس قرار دادن خبرهای جدید در دسته یا دسته های مربوطه توسط عوامل انسانی، نمونه ای از حوزه یادگیری با ناظر است. به بیان صریح تر، در این شیوه ی یادگیری مجموعه ای از موارد برچسب گذاری شده (از پیش دسته بندی شده) در اختیار ما قرار می گیرد که با کمک آن ها می توان توصیف دسته ها و قوانین دسته بندی را یاد گرفت. در دسته بندی خبرها به عنوان مثال ابتدا کلید واژه هایی که درهر دسته از خبری های ورزشی بیشتر تکرار شده اند مشخص می شود و سپس متن های بعدی که کلید واژه های مربوط به یک دسته خاص را در بر دارند، درآن دسته خبری قرارداده می شوند.

اما، در یادگیری بدون ناظر نیازی به دسته بندی قبلی داده ها نیست و داده ها می توانند گروههای مختلفی را شکل دهند که دارای ویژگیهای مشتر ک هستند. در این شیوه مجموعه ای از موارد برچسب گذاری نشده

[1] Machine Learning
[2] Supervised Learning
[3] Unsupervised learning
[4] Item

(دسته بندی نشده) در اختیار ما قرار می گیرد و سعی ما بر این است که ساختار داده ها را به گونه ای منطقی یاد گرفته و آن ها را توصیف کنیم (جین، مارتی و فلین،۱۹۹۹)[۱].

به منظور دستیابی به این دو شیوه یادگیری، معمولا چهار روش داده کاوی بکار گرفته می شود که عبارتند از: دسته بندی[۲]، خوشه بندی[۳]، قوانین تلازمی[۴]، و مصور سازی.[۵]

دسته بندی: دسته بندی بعنوان فرایندی از یادگیری نظارت شده بخش مهمی از مبحث داده کاوی است. در این روش یک مورد از داده بر اساس مشخصه هایش به یک دسته از سری دسته های از پیش تعریف شده منسوب می شود. تحلیل عملکرد ژن ها بر اساس دسته های از پیش تعریف شده توسط زیست شناسان ، نمونه ای از کاربرد دسته بندی است. در پزشکی نیز دسته بندی بیماری ها و روش های درمانی به روش های مختلف متداول است و از این دسته بندی ها در تشخیص و درمان بیماری ها بسیار استفاده می شود. به عنوان مثال دسته بندی شکستگی های مچ پا به چندین روش صورت گرفته که هر یک کاربردها و محدودیتهای خاص خود را دارد. به طور کلی نکته جالب توجه در بکارگیری هر دسته بندی این است که یک دسته بندی خاص ممکن است با گذشت زمان و به واسطه پیشرفت های حاصل دیگر نتواند همچون گذشته کارایی داشته باشد و لازم باشد تا با دسته بندی ها دیگر جایگزین شود. در این رابطه، در دسته بندی شکستگی های مچ پا، به عنوان مثال لاگ هانسن شکستگی ها را بر اساس موقعیت پا در هنگام ضربه دیدن و جهت نیروی وارد آمده به مچ پا به ۴ گروه تقسیم کرد. دسته بندی لاگ هانسن از شکستگی ها با روی کار آمدن متدهای جدید جراحی منسوخ شد. در سیستم AO/OTA با دسته بندی شکستگی ها بر اساس سطح شکستگی در استخوان فیبولا آن ها را به سه دسته A,B, C تقسیم کرد که هریک بعدا به گروه ها و زیر گروه های بیشتر بسط داده شدند. پیچیدگی این دسته بندی و متکی بودن آن به روش های عکسبرداری گران قیمت و دور از دسترس، نیز به عدم استفاده از این دسته بندی منجر شد. در دسته بندی دیگر پیشنهاد شد که شکستگیها به دو دسته کلی ثابت و غیر ثابت تقسیم شوند. از محاسن استفاده از این سیستم، تشخیص شیوه درمانی مناسب و پیش بینی

[1] Jain, Murty and Flynn, 1999
[2] Classification
[3] Clustering
[4] Association rules
Visualization[5]

چگونگی بهبود عضو شکسته شده است. به طور مثال مشخص شده است که شکستگیهای ثابت نیاز کمتری به جراحی و فیکسیشن داخلی دارند و تنها با جا انداختن قسمت شکسته از روی پوست و فیکسیشن خارجی مانند گچ و آتل بندی قابل درمان هستند. در حالیکه شکستگی های ناثابت نیاز به جراحی و فیکسیشن داخلی با پیچ و پلاک دارند (دیویدویچ و اگال،۲۰۱۰)[1].

استفاده از طرح های رده بندی در کتابخانه ها نیز نمونه ای از کاربرد دسته بندی در حوزه کتابداری است. به طور کلی، در طرح های رده بندی مختلف نظیر طرح رده بندی دیویی و کنگره دسته بندی موضوعات، خواه بر مبنای نظری یا بر اساس پشتوانه انتشاراتی، از پیش صورت گرفته است و یادداشت های دامنه همراه با توضیحات و مثال های مختلف در متن این طرح ها قرار گرفته تا به کتابدار نشان دهد که چگونه کتاب های کتابخانه خود را در دسته مربوطه قرار دهد یا به بیان دیگر رده مربوط به کتاب در دست فهرست را شناسایی نماید.

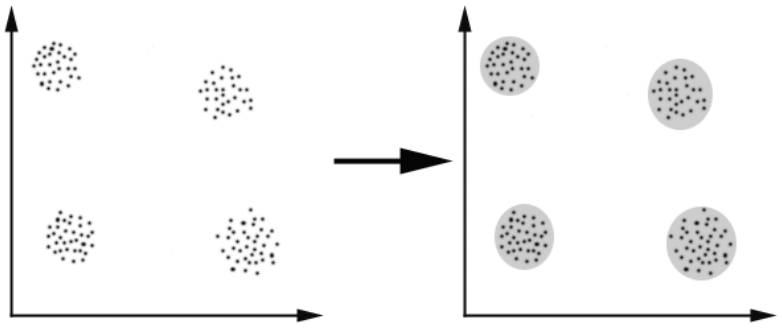
در سیستم رده بندی دیویی که اغلب برای رده بندی کتاب ها در کتابخانه های عمومی استفاده می شود به عنوان مثال دسته بندی موضوعات بر اساس رشته های دانشگاهی صورت گرفته است. به گونه ای که در این نظام کل دانش بشر به صورت فلسفی به ده رده اصلی (شامل کلیات، فلسفه، علوم اجتماعی، زبان، علوم خالص، علوم عملی، هنر، ادبیات، و جغرافیا) تقسیم شده و سپس هر یک از این ده رده به ده زیررده خاص تر و باز هر زیر رده به ده زیر رده فرعی تر تقسیم شده است. سپس تقسیم بندی موضوعات در هر زیر رده از عام به خاص تا جایی ادامه پیدا می کند که کتاب های مربوط به یک موضوع خاص در کنار سایر کتاب های مربوط به آن موضوع با نظم سلسله مراتبی قرار گیرند. بدین ترتیب مراجعه کننده ی کتابخانه می تواند با مراجعه به قفسه ها تمام کتاب های مربوط به یک حوزه ی موضوعی را کنار هم بیابد.

نکته مهم در دسته بندی یا در واقع رده بندی کتاب در کتابخانه ها ، توجه به اصل ویژگی است. بدین معنی که یک کتاب باید تا حد امکان در رده خاصتر قر ار گیرد و از قرار دادن آن در رده های عامتر اجتناب گردد. به عنوان مثال کتابی که در زمینه کبوتر ها نگاشته شده است، باید تا حد امکان و در صورتی که طرح رده بندی اجازه می دهد، در رده خاص تر کبوترها قرار گیرد، نه در رده کلی تر پرندگان. نظم سلسله مراتبی موضوعات نیز سبب می شود که کتاب های مربوط به موضوعات عام تر شماره رده ی کلی تر و کتاب های مربوط به

[1] Davidovitch and Egol,2010

موضوعات خاص تر شماره رده ی ریزتر دریافت کنند. بدین ترتیب در قفسه/های مربوط به یک حوزه موضوعی ابتدا کتاب هایی که به موضوع درسطح عام تر پرداخته اند قرار می گیرند و به دنبال کتاب های عام مربوط به یک حوزه موضوعی، کتاب های مربوط به موضوعات خاص ترمی آیند. ساختار سلسله مراتبی در طرح رده بندی دیویی بدین معنی است که هر موضوع در آن تابع و جزئی از موضوع های مافوق خود است. به این ترتیب، هر یادداشتی که به ذات یک موضوع مربوط می شود، به تمام اجزای تابع آن موضوع نیز قابل تعمیم است. این نکته در دسته بندی تمام اشیاء قابل ملاحظه است. بگونه ای که به عنوان مثال هر کبوتری دارای ویژگیهای کلی پرندگان است.

خوشه بندی: تحلیل خوشه ای در واقع نوعی تکنیک دسته بندی است که به ایجاد گروه های متجانس در مجموعه ای از داده های پیچیده کمک می کند(بورگن و بارنت،۱۹۸۷،ص.۴۵۶)[1] در خوشه بندی، اشیاء بر پایه ی میزان شباهت یا فاصله مشخصه هایشان به گروه های مختلف دسته بندی می شوند (شکل ۲). به عنوان مثال در خوشه بندی افراد بر اساس بهره هوشی آن ها، متغیر (مشخصه) مورد بررسی هوش بهر آن ها است. به طوری که افرادی که بهره هوشی آن ها بیشتر به هم نزدیک باشد در یک خوشه قرار می گیرند. به عبارت دیگر هر چه تفاوت بهره هوشی دو فرد کمتر باشد، شباهت آن ها بر اساس این مشخصه بیشتر، یا به عبارتی عدم شباهت (فاصله) بین آن ها کمتر است.

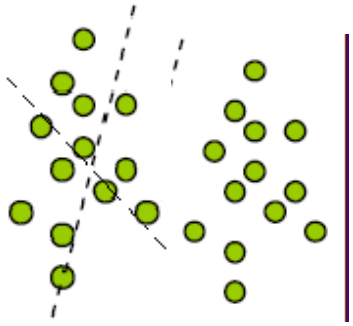


شکل ۴: در این شکل نمونه‌ای از اعمال خوشه‌بندی روی یک مجموعه از داده‌ها مشخص شده است که از معیار فاصله(Distance) به عنوان عدم شباهت(Dissimilarity) بین داده‌ها استفاده شده است.

[1] Borgen and Barnett, 1987, p. 456

خوشه بندی روشی در تحلیل اکتشافی داده ها[1] است. فرایند خوشه بندی، خوشه ها را بدون شناخت قبلی از محتوای مدارک یا از دسته هایی که احتمالا وجود دارد تولید می کند. این روش در مواردی به کار می رود که رکوردهای از پیش دسته بندی شده در اختیار ما نیستند و اشیاء بر اساس شباهتشان به گروه ها یا به عبارت دیگر خوشه های مختلف تقسیم می شوند. فرضیه خوشه بندی که به خوبی توسط ریجزبرگن (۱۹۷۹) تعریف شده است می گوید که از طریق فرایند خوشه بندی، اشیاء مشابه بر حسب مشخصه/های مشترکشان در یک گروه قرار می گیرند. هر گروه یا خوشه اشیائی را در بر می گیرد که به سایر اعضای آن گروه شباهت دارد ، اما با اعضای سایر خوشه ها مشابه نیست(وایوز و همکاران،۲۰۰۸).

از دیدگاه داده کاوی، خوشه بندی یک روش یادگیری غیر نظارت شده است و اغلب به روش خودکار[2] صورت می گیرد. زیرا خوشه بندی به روش دستی نیازمند صرف وقت بسیار زیاد از طرف متخصصین موضوعی است و در نتیجه بسیار پرهزینه می باشد. در استفاده از روش دستی همچنین دانش قبلی فرد ممکن است به سوگیری در فرایند خوشه بندی منجر شود. به علاوه، ارتباط اشیاء در داده های اولیه ممکن است به گونه ای باشد که خوشه های آن به راحتی توسط انسان قابل تشخیص نباشد (شکل۵). از اینرو، ارزیابی روابط معنایی بین موضوعات ممکن است همیشه به سادگی امکان پذیر نباشد. به بیان دیگر، ممکن است گستره ای از مدارک وجود داشته باشد که به صورت طبیعی نتوان آن ها را در هیچ سلسله مراتب موضوعی جای داد. این مسئله به خصوص زمانی که تعداد مدارک زیاد باشد و مشخصه های متعدد ازآن ها مد نظر باشد بروز می کند.



شکل ۵: داده خوشه بندی شده در مقابل داده خام

[1] Exploratory data analysis
[2] Automatic Clustering Analysis (ACA,(

خوشه بندی به روش خودکار به کشف خصایص مهم در نمونه های با ابعاد زیاد و کشف گروهها و طبقات جدید کمک می کند. ضمنا، در دسته بندی موضوعات به صورت دستی متخصصین موضوعی صرفا به صورت ذهنی و بر اساس دانش و آموخته های خود از مباحث مطروحه در منابع عمل می کنند، در حالیکه، با بکارگیری روش های خوشه بندی خودکار می توان ارتباط پنهان بین اصطلاحات موجود در مجموعه ای از مدارک را یافت که به صورت ذهنی یا با استفاده از روش های دیگر قابل شناسایی نیستند. به عنوان مثال ارتباط بین بیماری Raynaud و اسید های چرب موجود در روغن ماهی، استروژن و آلزایمر، و سردردهای میگرنی و عفونت مننژیتی برای اولین بار در یک مطالعه خوشه بندی روشن شد (بنرجی،۱۹۹۸)[1]. مطالعات خوشه بندی همچنین به بررسی سیر تحول موضوعات کمک نموده و نشان می دهند که یک مفهوم یا موضوع یا نظریه خاص کی و در کجا پیدایش یافته است.

از خوشه بندی به عنوان مثال در حوزه تجارت و بازاریابی در مدیریت روابط مشتری[2]، و تقسیم بازار[3] ، در حوزه زیست شناسی برای دسته بندی گیاهان و جانوران و در حوزه اطلاع رسانی برای خوشه بندی مدارک[4] استفاده می شود. خوشه بندی مدارک به تعیین شباهت بین مدارک با تحلیل ارتباط موضوعی اصطلاحات موجود در متن می پردازد. این شیوه می تواند به کشف شباهت هایی میان مدارک کمک کند که در روش های دیگر از قبیل شباهت سنجی مدارک بر اساس هم پوشانی استناداتشان ممکن نیست. در این زمینه سوانسون طی پژوهشهای متعدد، نشان داد که این شیوه می تواند به پیدایش فرضیه های جدید منجر شود. به این وسیله او توانست به عنوان مثال به ارتباط بین کمبود منیزیم و سردردهای میگرنی پی ببرد (سوانسون،۱۹۸۷؛سوانسون،۱۹۹۱).

قوانین تلازمی: قوانین تلازمی، اولین بار توسط آگراوال و ریکانت (۱۹۹۴)[5] مطرح شد. این تکنیک اساسا برای پیدا کردن روابط معنادار بین موارد یا متغیرهایی که به طور همزمان در پایگاه های اطلاعاتی رخ می دهند، به کار گرفته می شود و زمانی می تواند مفید واقع گردد که وابستگی های مختلف بین موارد در یک مجموعه بزرگ داده ها وجود داشته باشد. از این تکنیک در استخراج دانش از وبلاگ ها بسیار استفاده می شود (لی و

[1] Banerjee,1998
[2] Management of customers' relationships
[3] Market Segmentation
[4] Document Clustering
[5] Agrawal and Srikant 1994

همکاران،۲۰۰۲)[1] و در کشف الگوهای وابستگی میان محصولات در حوزه تجارت الکترونیک کاربرد فراوان دارد.

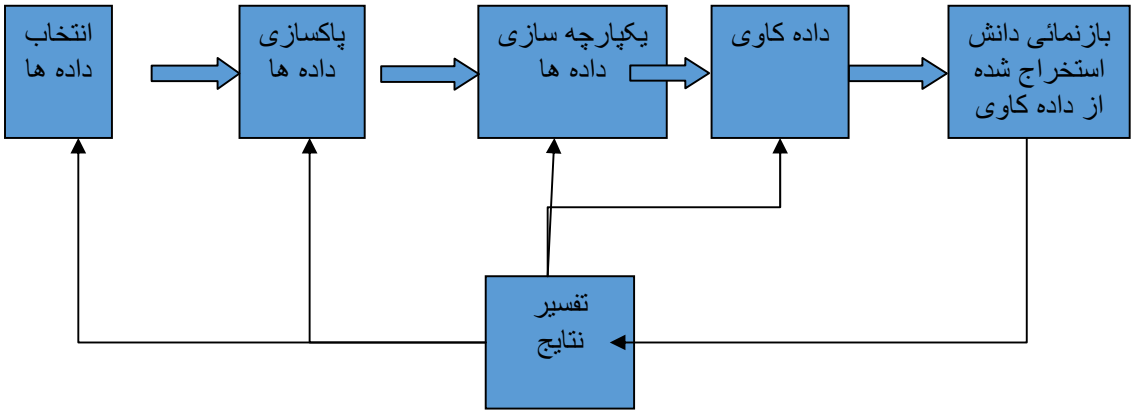
مصور سازی عینی: مصور سازی عینی بر این فرض استوار است که افراد می توانند ساختار موجود در مجموعه ای از داده ها به شکل بصری را بهتر درک کنند. از اینرو با نشان دادن داده ها به شکل بصری می توان به تعامل مستقیم فرد با مجموعه داده ها و استنتاج و کشف دیدگاه های تازه از آن ها کمک کرد. این تکنیک به ویژه زمانی سودمند واقع می شود که شناخت کمی از داده ها وجود دارد و خواننده نمی داند که به دنبال کشف چه چیزی است. تحلیل هم استنادی نمونه ای از کاربرد مصورسازی عینی است که در بحث وب کاوی به آن اشاره شد.

مراحل مختلف فرایند داده کاوی:

- **پاک سازی داده ها**: یا به عبارت دیگر از بین بردن نویز و ناسازگاری داده ها. در این مرحله داده های غیر معتبر شامل داده های دارای نویز و داده های ناقص از مجموعه داده ها خارج می شوند.
- **یکپارچه سازی داده ها**: در این مرحله، منابع چندگانه داده ای با هم ترکیب می شوند.
- **انتخاب داده ها** : داده های مرتبط به فرایند داده کاوی از سایر داده ها جدا می شود. این مبحث را می توان بخشی از فرایند کاهش اطلاعات نیز دانست.
- **تبدیل داده ها**: در این مرحله داده ها از طریق خلاصه سازی، همسان سازی یا محاسبه مقادیر تجمعی، به قالبی که برای داده کاوی قابل استفاده باشد تبدیل می شوند.
- **داده کاوی**: بخش اصلی فرایند که در آن با استفاده از روش ها و تکنیکهای خاص ، موارد یا به عبارتی الگوها استخراج می شوند.
- **ارزیابی موارد**: در این مرحله صحت موارد بر اساس معیارهای جذابیت تشخیص داده می شود.

[1] Lee et al.,2002

- **بازنمایی دانش**: در این بخش به منظور ارائه دانش استخراج شده به کاربر ، از یک سری ابزارهای مصور سازی عینی استفاده می گردد.



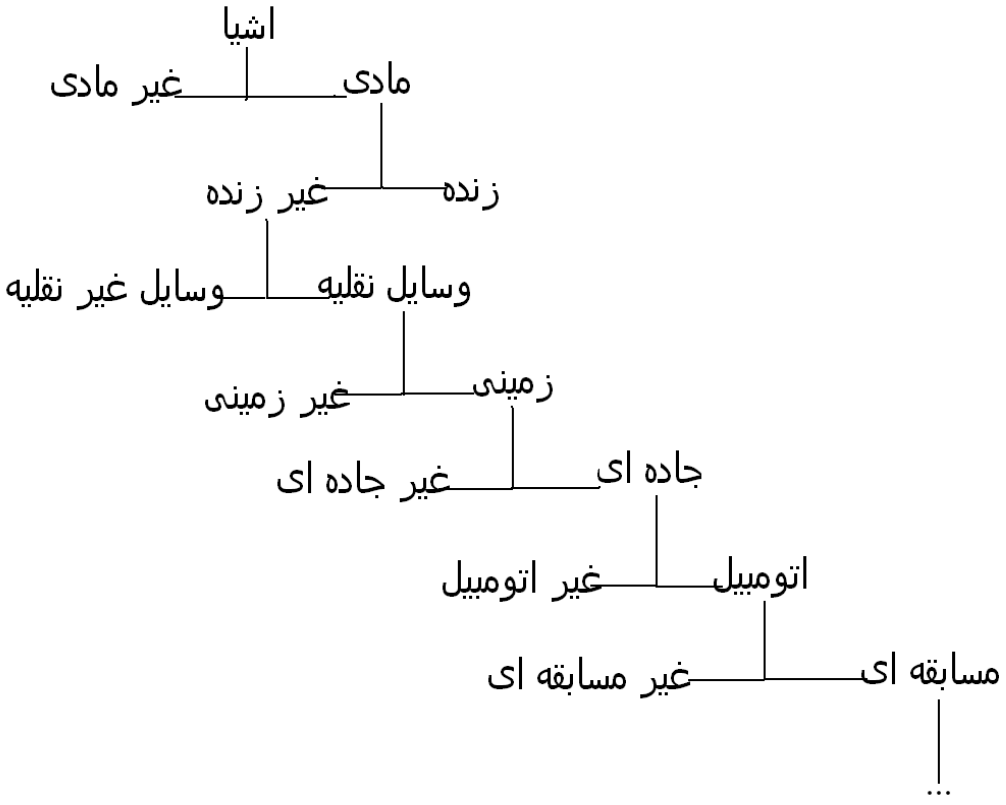
شکل ۵: مراحل مختلف فرایند داده کاوی

توضیحات بیشتر در خصوص مراحل مختلف داده کاوی و روشهای اجرا در فصول بعد دنبال خواهد شد.

مفهوم شیئ

اصطلاح شیئ زیربنایی ترین مفهوم در خوشه بندی است زیرا خوشه بندی عملا به دسته بندی اشیاء می پردازد. از این رو باید مشخص کرد که منظور از شیئ چیست و در اصل به چه چیزی شیئ گفته می‌شود. عملا به هر موجود ذهنی و عینی می توان اصطلاح شیئی را اطلاق کرد.

هر شیئ زیر مجموعه ای از یک دسته عام تر محسوب می شود. مثلا، اتومبیل زیر مجموعه وسایل نقلیه و وسایل نقلیه زیر مجموعه اشیای غیر زنده و اشیای غیر زنده زیر مجموعه اشیای مادی و آن نیز در مجموعه کل اشیا قرارمی‌گیرد. تصویر زیر این منظور را بهتر نمایش می‌دهد.



تصویر: نمونه ای از سلسله مراتب دسته بندی اشیاء

دسته بندی اشیا به طبقات ریزتر به آن‌ها هویتی دقیق‌تر می‌دهد. اشیایی که در طبقات پایین تر قرار می گیرند در برابر طبقات بالاتر یک مصداق یا نمونه محسوب می‌شوند. مثلا اتومبیل پراید مصداق یا نمونه ای از اتومبیل است. پراید تولید ۱۳۹۰ خود مصداق یا نمونه ای از یک پراید است. هر یک از پرایدهای تولیدی ۱۳۹۰ خود مصداقی از یک پراید تولیدی سال ۱۳۹۰ است. به عنوان مثال دیگر کبوترها مصداقی از پرندگان هستند و کبوترهای چاهی مصداقی از کبوترها هستند. کبوترهای چاهی دارای هویتی خاص تر در مقایسه با گروه کلی کبوترها هستند و کبوترها به نوبه خود هویتی دقیق تر در مقایسه با گروه کلی تر پرندگان دارند.

مشخصه اشیاء

ویژگی مشترک اشیاء این است که دارای یک سری مشخصه هستند و این مشخصه ها اساس خوشه بندی است. اتومبیل نمونه‌ای مشخص از یک شیئ است که سرعت، قدرت موتور، قطر لاستیک‌ها، استحکام بدنه، کارخانه سازنده، مارک و غیره از مشخصه های آن به شمار می آید. اگر مثلا قدرت موتور دو اتومبیل با هم فرق کند، از نظرقدرت، آن دو اتومبیل متمایز دانسته می شوند و در دو دسته مختلف جای می گیرند.

اگر چه مجموعه مشخصه های یک شیئی به آن هویت می بخشند، اما اگر حتی یک مشخصه تغییر کند، هویت شیئی تغییر می یابد. البته، با تغییر یک مشخصه نمی توان به شیئ هویت دلخواه داد. مثلا، خودرویی که هویت یک اتومبیل مسابقه‌ای را دارد، دارای مجموعه ای از مشخصه ها است که روی هم رفته به آن هویت اتومبیل و مسابقه ای بودن می دهد. بنابراین، یک مشخصه مانند سرعت بالا به تنهایی نمی‌تواند به اتومبیل هویت مسابقه‌ای بودن بدهد بلکه باید از مشخصه های دیگری مانند وزن مناسب، استحکام مناسب و بسیاری از چیزهایی دیگر برخوردار باشد. پس ، اگر اتومبیل یکسری از مشخصه های لازم را داشته باشد، هویت اتومبیل مسابقه ای را پیدا می کند و در دسته اتومبیل‌های مسابقه ای و در غیر این صورت در دسته اتومبیل های غیر مسابقه ای جای می‌گیرد.

همان طور که یک شیئ دارای مشخصه است، هر مشخصه در جای خود می تواند یک شیئی محسوب شود. مثلا، کتاب مشخصه‌ای مانند نویسنده دارد و بخشی از هویت خود را از آن می گیرد. نویسنده در عین این که مشخصه محسوب می شود، به نوبه خود می تواند یک شیئ هم باشد. یکی از مشخصه هایی که به یک نویسنده هویت می بخشد، کتاب‌هایش است. از این رو، دو شیئ گوناگون می‌توانند مشخصه ای برای یکدیگر به شمار بآیند. همان گونه که مشخصه ای به نام کتاب در جای خود یک شیئ است و مشخصه ای به نام نویسنده دارد و نویسنده نیز در جای خود شیئ است و با مشخصه ای به نام کتاب هویت می‌یابد. برای استفاده های بعدی، این رابطه را قاعده "دوطرفه شیئ–مشخصه" می نامیم.

معمولا هر شیئ بیش از یک مشخصه دارد اما برای دسته بندی اشیاء و قرار دادن آن ها در دسته های ریزتر، معمولا یک مشخصه از مجموع مشخصه‌های آن‌ها را در نظر می‌گیریم. به عبارتی، اشیای درون یک دسته را تنها بر اساس شباهتشان در یک مشخصه ی مشترک تقسیم بندی می‌کنیم. از این رو، در فرآیند خوشه بندی یکی از مسائلی که مطرح است، برگزیدن مشخصه ای است که اشیاء بر اساس میزان شباهت یا عدم شباهتشان در آن مشخصه خوشه بندی شوند. این نشان می‌دهد که اگر اشیای درون یک دسته را بر اساس شباهت در مشخصه

دیگرشان دسته بندی می‌کردیم، خوشه های دیگری حاصل می‌شد. مثلا اگر اتومبیل ها را بر اساس مشخصه نام کشور سازنده دسته بندی کنیم، دسته هایی مانند اتومبیل های ایرانی، ژاپنی، آلمانی و غیره ایجاد می‌شود. حال اگر اتومبیل ها را بر اساس مشخصه نام کارخانه سازنده دسته‌بندی می کردیم، دسته هایی چون ایران خودرو، سایپا، بنز، تویوتا، مزدا و غیره به وجود می‌آمد. نتیجتا، اشیا را می‌توان به تعداد مشخصه‌هایی که دارند چندین بار دسته بندی کرد. البته خوشه بندی باید بر اساس مشخصه ای صورت بگیرد که نتایج آن به خلاصه سازی و تحلیل اطلاعات و داده های مورد نظر ما کمک کند. مثلا اگر بخواهیم اتومبیل ها را بر اساس قدرت آن ها خوشه بندی کنیم، حجم موتور اتومبیل ها را به عنوان مشخصه مورد نظر بر می گزینیم و آن ها را بر اساس شباهتشان در این مشخصه دسته بندی می‌کنیم.

اشیاء یک دسته وقتی بر اساس میزان شباهت در یک مشخصه تقسیم بندی شوند، چند دسته کوچکتر را به وجود می‌آورند که اصطلاحا به آن ها خوشه می‌گوییم. اگر دسته عامتر را والد این دسته ها بدانیم، مسلما اشیاء درون این دسته ها مشخصه های والد خود را به ارث خواهند برد. مثلا دسته ای به نام اتومبیل ها، والد اتومبیل های مسابقه ای است. بنابراین، اتومبیل های مسابقه ای تمامی مشخصه های اتومبیل را دارند و تنها بر اساس شباهتشان در یک مشخصه (مسابقه ای بودن) در یک دسته قرار گرفته اند. از این رو، وقتی که به دسته بندی اتومبیل ها به مسابقه ای و غیر مسابقه ای می‌پردازیم، پذیرفته ایم که اشیائی که می‌خواهیم دسته بندی کنیم، اتومبیل هستند. به عبارتی این نکته را به رسمیت شناخته ایم که این اشیا تمامی مشخصه‌های دسته بالاتر را به ارث برده اند. مثلا همه آن‌ها بیش از سه چرخ دارند، دارای فرمان هستند، گاز و ترمز و سایر مشخصه‌های مشترک اتومبیل ها را دارند.

گاهی دسته بندی را براساس یک مشخصه انجام می دهیم و پس از آن تحقیق می کنیم تا ببینیم که آیا اعضای دسته های حاصل از نظر دارا بودن مشخصه ای ثانوی نیز با هم اشتراک دارند و آیا میزان آن مشخصه ثانوی در دسته های مختلف به یک اندازه است. مثلا افراد را می‌توان از نظر مشخصه بزهکاری به دو دسته بزهکار و غیر بزهکار تقسیم کرد. ممکن است افراد هر یک از این دو دسته در مشخصه ای دیگر مانند ملیت نیز با هم مشترک باشند. مثلا ممکن است بزهکاران همگی یا اکثرا از یک کشور باشند. در این صورت بزهکار بودن و ملیت را باید دو مشخصه تابعی یا مرتبط خواند که می توان از مقدار یکی، مقدار مشخصه دیگری را تعیین کرد. از این رو مقدار مشخصه ها را می توان اجزا و پارامترهای یک تابع دانست. اصطلاحا، مشخصه یا مشخصه هایی که مقدار آن مشخص است، مثل مشخصه بزهکاری، ورودی تابع و مقدار حاصله از محاسبه روی این مقادیر، مثل

به دست آوردن ملیت افراد، را خروجی آن تابع می‌نامند. تابع ساده y =۲x که یک تابع خطی است، می‌تواند این گفته را بهتر نمایش دهد. همان گونه که می بینیم، مقدار خروجی تابع y با مقدار x در ارتباط است. به زبان دیگر، مقدار مشخصه y تابع مقدار مشخصه x است و اگر مثلا مقدارx برابر با ۵ باشد، مقدار y برابر با ۱۰ خواهد بود.

داده ها

هر شیئ را می‌توان یک داده در نظر گرفت. علاوه بر آن، هر مشخصه از یک شیئ را نیز می‌توان یک داده محسوب کرد که داده اصلی (شیئ) را توصیف می‌کند. مثلا اتومبیل داده ای (شیئ) است که داده (مشخصه) حجم موتور، قدرت آن را نشان می‌دهد. هر انسان داده ای است که داده سال تولد نمایانگر سن او و داده قد نمایانگر طول او است. این که هر مشخصه خود یک داده است، نشان می‌دهد که می‌توان مقداری را به آنان نسبت داد. فرضا، حجم موتور اتومبیل که یک مشخصه است، می‌تواند مقداری برابر با ۱۸۰۰، ۲۰۰۰، ۲۸۰۰ سی–سی و بیشتر یا کمتر باشد.

نتیجه دیگری که از داده دانستن اشیا و همین طور داده دانستن مشخصه های اشیاء می‌توان گرفت این است که می‌توان روی آن‌ها محاسبه و پردازش انجام داد. این که چه محاسباتی می‌توان با هدف دسته بندی روی مشخصه ها انجام داد، به نوع داده بستگی دارد.

انواع داده ها:

داده ها به طور کلی به چهار نوع تقسیم می شوند: اسمی، رتبه ای، فاصله ای و نسبی.

ساده ترین نوع داده، *داده اسمی* است. داده های اسمی به داده هایی اطلاق می شود که می توان به آن ها عدد یا اسمی را نسبت داد، اما این اعداد و اسامی منتسب، صرفا قراردادی است. مثلا تقسیم کلاس‌های اول یک مدرسه به کلاس اول–الف، اول–ب، اول–پ و غیره نشان می‌دهد که داده ای (شیئ) به نام کلاس، داده ای (مشخصه‌ای) با عنوان نام کلاس دارد که از نوع اسمی است و مقدار آن الف، ب، پ و غیره است. علاوه بر این، می‌توان به هر کلاس، عددی را به جای حروف الفبا نسبت داد، مثل کلاس اول–یک، اول–دو، اول–سه. این مثال نشان می‌دهد که روی این گونه اعداد قراردادی نمی‌توانیم عملیات ریاضی مثل جمع، تفریق، ضرب و

تقسیم را انجام دهیم. تنها کاری که می‌توان روی آن ها انجام داد، شمردن تعداد آن‌ها است که اصطلاحا به آن تعیین فراوانی می‌گویند.

نوع دیگر داده، داده های رتبه ای است. این داده ها برتری یا رتبه یک شیئ را در بین اشیاء دیگر از نظر یک مشخصه نشان می‌دهد. به این نوع از داده ها نیز می‌توان یک عدد منتسب کرد. اما بر خلاف داده های اسمی نمی توان هر عددی که خواستیم به یک داده نسبت بدهیم. در مثال قبل، وقتی که کلاسی اول- یک نام دارد، اگر به آن کلاس نام اول – دو را می دادیم و به کلاسی دیگر اول –یک می گفتیم، اشکالی ایجاد نمی شد. اما در این نوع داده ها قاعده ای برای منتسب کردن اعداد به یک داده وجود دارد. برای این کار، یک شیئ از نظر میزان دارا بودن یک مشخصه بررسی می شود و اشیاء به ترتیب از زیاد به کم یا از کم به زیاد رتبه بندی می‌شوند. معروفترین کاربرد این نوع داده در مسابقات ورزشی است. به عنوان مثال به ورزشکاران یا تیم های ورزشی بر اساس رکورد یا میزان امتیازات کسب شده، رتبه اول، دوم، و... داده می شود.

ویژگی مشترک داده های اسمی و رتبه ای در این است که نمی توان روی آن ها عملیات ریاضی (جمع، تفریق، ضرب و تقسیم) انجام داد. از این رو این دو نوع داده را داده های کیفی می‌نامند.

نوع پیشرفته تر داده، داده فاصله ای است. همان گونه که از اسم این نوع داده بر می‌آید، با آن می توان فاصله دو داده را از هم سنجید. امکان سنجیدن فاصله یا تفاوت دو داده از این نوع با عملیات ریاضی تفریق امکان پذیر است. به عنوان مثال با منها کردن دمای یک شهر از شهر دیگر می‌توان تفاوت دمای آن دو شهر را سنجید. بنا بر این، در صورت داشتن چنین داده هایی، دسته بندی آن ها بر مبنای فاصله بین دو داده که از طریق انجام عملیات تفریق محاسبه می شود، صورت می‌گیرد.

آخرین نوع داده، داده نسبی است. همان گونه که از نام این نوع داده برمی‌آید، می‌توان برای مقایسه دو داده از این نوع، نسبت بین آن ها را سنجید. برای سنجیدن نسبت بین دو شیئ از نوع نسبی، مقدار یکی را بر دیگری تقسیم می‌کنیم. مثلا وقتی که شخصی ۱۸ سال و شخص دیگر ۹ سال دارد، شخص اول دو برابر شخص دیگر سن دارد، یعنی نسبت سن او با دیگری ۲ است. تفاوت این نوع از داده ها با نوع قبلی(داده فاصله ای) در این است که در نوع قبلی نمی توان نسبت را حساب کرد. مثلا وقتی که دمای شهری ۱۸ و دمای شهری دیگر ۹ است، نمی توان گفت که شهر اول، دو برابر شهر دیگر گرم است. دلیل آن این است که درجه حرارت، نقطه

صفر مطلق ندارد ومی‌تواند بالا یا پائین صفر باشد. در اساس تفاوت داده های فاصله ای با نسبی در آن است که داده های فاصله ای صفر مطلق ندارند.

ویژگی مشترک داده های نوع فاصله ای و نسبی در این است که می توان روی هر دو عملیات تفریق و جمع را انجام داد. از این رو این دو نوع داده را داده های کمی می نامند.

مقدار داده‌ها:

یک مشخصه، به عنوان یک داده، از دو بخش نام مشخصه و مقدار مشخصه تشکیل شده است. مشخصه رنگ اتومبیل را در نظر بگیریم، نام این مشخصه رنگ است و مقدار آن می‌تواند بین قرمز، سیاه، سفید و سایر رنگ ها متغیر باشد یا مقدار مشخصه ای مثل عمر اتومبیل می‌تواند بین صفر و بیشتر متغیر باشد. آن چه که باعث می‌شود تا اشیاء از والد خود جدا شده و در دسته‌های متمایزتر قرار بگیرند، همین مقداری است که در یک مشخصه کسب می‌کنند.

مقدار مشخصه ها به طور کلی می‌تواند از نوع رشته‌ای، عددی (صحیح و اعشاری) و بولی باشد. در مثال رنگ اتومبیل مقدار مشخصه، قرمز، سیاه سفید و غیره بود که از نوع رشته ای است. مقدار مشخصه ی عمر اتومبیل عددی است که می تواند بین صفر و بیشتر باشد. و مقدار مشخصه کیسه هوا می‌تواند وجود یا عدم وجود آن در اتومبیل باشد که مقدار این مشخصه بولی است.

انواع متغیرها در ارتباط با نوع داده: متغیر رشته‌ای نوعی متغییر است که مقدار آن می‌تواند عدد، حرف، علائم یا ترکیبی از آن‌ها باشد. مثلا عبارت "ali110*&" یک رشته محسوب می‌شود که ترکیبی از حروف (ali) ، اعداد (110) و علائم (&*) است. طول یک رشته می‌تواند از صفر تا بی‌نهایت باشد. این نوع از متغییرها تفاوتی بین اعداد، حروف و علائم قائل نیستند. لذا اگر در یک سیستم، عددی را به عنوان مقدار یک متغییر رشته‌ای قرار دهیم، آن را همانند حروف و علائم را به عنوان بخشی از یک متن در نظر می‌گیردکه انجام عملیات ریاضی روی آن‌ها ممکن نیست.

مقدار داده های نوع اسمی، معمولا به صورت رشته ای است. وقتی که حتی عددی را به یک داده اسمی نسبت می‌دهیم، آن عدد به علت خاصیت رشته ای بودن قابل محاسبه نیست. مثل فصل های کتاب که به عنوان داده

ای اسمی از اعداد ۱، ۲، ۳ و غیره تشکیل شده است (فصل ۱، فصل ۲ و فصل۳) که قابل جمع، تفریق، ضرب و تقسیم نیستند. مقدار داده های نوع رتبه ای نیز به صورت رشته ای است. زیرا روی این نوع داده ها نیز نمی توان عملیات ریاضی جمع، تفریق، ضرب و تقسیم را انجام داد.

اعداد صحیح متغییرهایی هستند که تنها اعداد صحیح را به عنوان مقدار خود دریافت می‌کنند. اعداد صحیح اعدادی هستند که اعشار ندارند و مقدار آن بین ∞+ تا ∞- قرار دارد. ۰، ۱ و ۱- نمونه‌ای از این اعداد هستند. اعداد اعشاری نیزمی‌توانند بین ∞+ تا ∞- قرار گیرند. تفاوت آن‌ها با نوع قبل در این است که اعشار را نیز قبول می‌کنند. این نوع از مقدار(صحیح و اعشاری) مناسب داده های فاصله ای و نسبی است. زیرا می‌توان روی این نوع از داده ها عملیات ریاضی انجام داد.

متغییرهای بولی، نوعی دیگر از متغییرها هستند که مقدار آن‌ها بیش از دو حالت ندارد: صحیح و غلط. در مثال تقسیم بندی اتومبیل ها از نظر داشتن یا نداشتن کیسه هوا، مقداری که این مشخصه پیدا می‌کند، داشتن کیسه هوا (صحیح) یا نداشتن آن (غلط) است. بنا بر این اتومبیل هایی که کیسه هوا دارند در یک دسته و آن هایی که ندارند در دسته دیگر قرار می‌گیرند. این نوع از مقدار، مناسب هر نوع از انواع چهارگانه داده (اسمی، رتبه ای، فاصله ای و نسبی) است.

نکته مهم دیگر درباره نوع داده ها این است که هر داده نوع بالاتر را می‌توان به داده های نوع پائین تر تبدیل کرد. اگر داده های اسمی، رتبه ای، فاصله ای و نسبی را به ترتیب از پائین به بالا در نظر بگیریم، داده های نسبی را می توان به سه نوع قبلی(فاصله ای، رتبه ای یا اسمی)، فاصله ای را به دو نوع قبلی (رتبه ای یا اسمی) و رتبه ای را به نوع قبلی (اسمی) تبدیل کرد. مثلا درآمد افراد، داده ای از نوع نسبی است که اگر بخواهیم آن را به رتبه ای تبدیل کنیم، می توانیم افرادی که از لحاظ در آمدی در یک محدوده خاص قرار می گیرند، را رتبه بندی کنیم، مثل کسانی که تا یکصد هزار تومان درآمد دارند، کسانی که بین یکصد تا دویست هزار تومان و الی آخر.

نکته دیگری که با مقدار دهی به داده ها در ارتباط است، نوع داده ها از جنبه **ساده و تکرار پذیر بودن آن ها است.** مشخصه های ساده آن هایی است که تنها یک مقدار دارند. مثلا، مقدار مشخصه حجم موتور اتومبیل تنها یک چیز می‌تواند باشد: ۱۸۰۰ ، ۲۰۰۰، ۲۸۰۰ سی سی و غیره. همین طور، مشخصه سال تولید اتومبیل تنها یک مقدار دارد، مثل سال ۱۳۹۰. در برابر مشخصه هایی است که آرایه ای از مقادیر را دریافت می کنند، یعنی

بیش از یک مقدار می‌گیرند. اگر نویسنده را به عنوان یکی از مشخصه های کتاب در نظر بگیریم، یک کتاب می‌تواند یک یا بیش از یک نویسنده داشته باشد که نام هر نویسنده یک مقدار از آرایه مقادیر مربوط به آن مشخصه است. بنابر این هر مشخصه ای که بتواند آرایه ای از مقادیر یا بیش از یک مقدار دریافت کند، یک مشخصه تکرار پذیر نامیده می‌شود.

تشکیل ماتریس متقارن

ماتریس متقارن:

پس از شناخت نوع داده، نوبت به تشکیل ماتریس متقارن می رسد که به آن ماتریس هم جواری گفته می‌شود. ماتریس جدولی است که از n سطر و m ستون تشکیل شده است. پس می توان گفت که ماتریس آرایه ای از اعداد است که می‌تواند یک یا بیش از یک بعد داشته باشد. همان گونه که جدول پائین نشان می‌دهد، وقتی که از آرایه صحبت می‌کنیم، منظورمان نمایش اعداد به صورت سطر و ستون در یک جدول است. در هر ماتریس تعداد هر یک از ستون ها و سطرهامی‌تواند بین ۱ و بیشتر متغیر باشد. بنابر این هر ماتریس می‌تواند حداقل یک بعدی (۱× ۱) باشد. در ماتریس هایی که برای خوشه بندی به کار می‌رود، تعداد سطرها و ستون ها(n و m) برابر است و به همین دلیل این ماتریس ها، ماتریس متقارن، مربع یا n × n نیز نامیده می‌شوند.

در ماتریس های متقارن اگر دو شیئ با نام های x و y داشته باشیم، عملا شباهت آن دو در هر دو صورت (y and x) و(x and y) یکی است. پس در مقایسه جفتی برای تشکیل ماتریس متقارن اگر sim به معنای شباهت باشد، چنین فرض می‌شود:

$$sim(x, y) = sim(y, x)$$

.

	شیئ۱	شیئ۲	شیئ۳	شیئ۴
شیئ۱				

شیئ۲				
شیئ۳				
شیئ۴				

همان گونه که در جدول بالا با رنگ های سفید و خاکستری مشخص شده است، مقدار هر خانه با مقدار خانه قرینه اش یکی و عناوین سطرها و ستون ها مشابه هم است.

برای تعیین اعداد درون جدول و ایجاد یک ماتریس متقارن، ابتدا باید شباهت یا تفاوت داده های مورد نظر(اشیاء) را به صورت جفتی یا دو به دو مقایسه کرد. این که چگونه می‌توان شباهت یا تفاوت دو داده را به صورت جفتی محاسبه کرد، به نوع داده وابسته است. سه نفر را در نظر بگیرید که می‌خواهیم آن ها را بر اساس شباهت یا تفاوت میزان قد دسته بندی کنیم که داده ای نسبی است. اگر قد آن ها به ترتیب ۱۷۰، ۱۸۰ و ۲۰۰ سانتی متر باشد، باید تفاوت قد آن‌ها را به ترتیب زیر مقایسه کنیم:

۱- تفاوت قد نفر اول با نفر اول(۰ = ۱۷۰- ۱۷۰)، نفر اول با نفر دوم (۱۰ - = ۱۸۰-۱۷۰)، و نفر اول با نفر سوم(۳۰ - =۲۰۰-۱۷۰)

۲- تفاوت قد نفر دوم با نفر اول(۱۰- = ۱۸۰ - ۱۷۰)، نفر دوم با نفر دوم (۰=۱۸۰ - ۱۸۰)، و نفر دوم با نفر سوم (۲۰ = ۱۸۰-۲۰۰)

۳- تفاوت قد نفر سوم با نفر اول (۳۰ = ۱۷۰ - ۲۰۰)، نفر سوم با نفر دوم (۲۰ = ۱۸۰ - ۲۰۰)، و نفر سوم با نفرسوم (۰ = ۲۰۰ - ۲۰۰).

برای به تصویر کشیدن این مقایسه ها می‌توانیم از جدولی متشکل از سه سطر و سه ستون بهره ببریم. این جدول نشان می‌دهد که برای سنجش شباهت سه مورد باید نه (۳×۳)مقایسه صورت گیرد.

جدول: ماتریس متقارن ۳×۳ از مقایسه دو به دوی قد سه نفر

	نفر اول	نفر دوم	نفر سوم
نفر اول	۰س	۱۰س	۳۰س
نفر دوم	۱۰س	۰س	۲۰س
نفر سوم	۳۰س	۲۰س	۰س

داده های اسمی:

برای تشکیل یک ماتریس متقارن از داده های اسمی می توان بر میزان اشتراک بین دو شیئ توجه کرد. در این صورت سه روش متصور است: فراوانی هم رخدادی، جاکاردی و کوسینوسی.

فراوانی هم رخدادی. در صورتی که مشخصه مورد نظر، داده ای از نوع اسمی و تکرار پذیر باشد، یکی از راه ها تعیین فراوانی هم رخدادی دو به دوی مقادیر موجود است. وقوع یک رخداد مشابه برای دو شیئ را هم رخدادی می گویند. مثلا وقتی دو مشتری چند محصول مشابه می خرند، رخدادی مشترک به وقوع پیوسته است. تعداد محصولات مشابهی که آن دو خریده اند، عملا فراوانی هم رخدادی است. در این جا مشتری یک شیئ است و محصولات خریداری شده توسط او یک مشخصه محسوب می شود. مثال دیگر، کتاب به عنوان شیئ و نویسنده به عنوان یک مشخصه از کتاب است. فرض کنید سه نویسنده داریم که تعدادی کتاب نوشته اند. اگر آن ها را به ترتیب " نویسنده اول"، "نویسنده دوم" و "نویسنده سوم" بنامیم، و از میان این کتاب ها، روی ۵۰ کتاب نام نویسنده اول، روی ۴۰ کتاب نام نویسنده دوم و روی ۹۰ کتاب نام نویسنده سوم آمده باشد، باید نه مقایسه جفتی از روش تعیین فراوانی هم رخدادی به شکل زیر انجام دهیم.

۱- فراوانی هم رخدادی نویسنده اول با نویسنده اول، نویسنده اول با نویسنده دوم و نویسنده اول با نویسنده سوم.

۲- فراوانی هم رخدادی نویسنده دوم با نویسنده اول، نویسنده دوم با نویسنده دوم و نویسنده دوم با نویسنده سوم.

۳- فراوانی هم رخدادی نویسنده سوم با نویسنده اول، نویسنده سوم با نویسنده دوم، نویسنده سوم با نویسنده سوم.

جدول: ماتریس فراوانی هم رخدادی سه نویسنده

	نویسنده اول	نویسنده دوم	نویسنده سوم
نویسنده اول	۵۰ کتاب	۱۰ کتاب	۲۰ کتاب
نویسنده دوم	۱۰ کتاب	۴۰ کتاب	۷ کتاب
نویسنده سوم	۲۰ کتاب	۷ کتاب	۹۰ کتاب

در جدول بالا، برچسب سطرها و ستون ها مشخصه هایی هستند که به صورت دو به دو مورد مقایسه قرار گرفته‌اند و اعداد درون خانه های جدول، فراوانی اشیائی است که از مقایسه جفتی با روش هم رخدادی تعیین شده است. سطر و ستون اول از جدول بالا نشان می‌دهد که ۵۰ کتاب است که روی آن نام نویسنده اول آمده است ، ۱۰ کتاب است که روی آن هم نام نویسنده اول و هم نام نویسنده دوم، ۲۰ کتاب که روی آن هم نام نویسنده اول و هم نام نویسنده سوم آمده است. سطرها و ستون های دیگر نیز به همین صورت هم رخدادی ها را نشان می‌دهد. وقتی که مقدار هم رخدادی دو نویسنده را تعیین می‌کنیم، . منظور این نیست که هریک از این دو نویسنده به تنهایی، فقط این تعداد کتاب نوشته اند، چرا که این تعداد می تواند برای هر یک از آنها به تنهایی بیشتر هم باشد. اما ما در هر زمان دو نویسنده را مد نظر داریم تا مقایسه جفتی انجام دهیم.

در خصوص هویت اشیاء گفتیم که هویت هر شیئ را مشخصه‌هایش تعیین می‌کند و هر مشخصه بر اساس "قاعده دوطرفه شیئی-مشخصه" نیز در جای خود می‌تواند یک شیئ باشد. در مثال جدول قبل، نویسنده یک شیئ محسوب شد که کتاب یکی از مشخصه های آن بود. بر اساس این قاعده، هر کتاب می‌تواند خود یک شیئ منظور شود و نویسندگان به عنوان یکی از مشخصه‌های آن به حساب آید. جدول زیر در مقایسه با جدول قبلی چنین موردی را به تصویر می‌کشد. در این جا نیز نه (۳× ۳) مقایسه جفتی از روش تعیین فراوانی هم رخدادی به شکل زیرصورت می‌گیرد:

۱- فراوانی هم رخدادی کتاب اول با کتاب اول، کتاب اول با کتاب دوم، کتاب اول با کتاب سوم.

۲- فراوانی هم رخدادی کتاب دوم با کتاب اول، کتاب دوم با کتاب دوم، کتاب دوم با کتاب سوم.
۳- فراوانی هم رخدادی کتاب سوم با کتاب اول، کتاب سوم با کتاب دوم، کتاب سوم با کتاب سوم.

جدول: ماتریس فراوانی هم رخدادی سه کتاب

	کتاب اول	کتاب دوم	کتاب سوم
کتاب اول	۴نویسنده	۲نویسنده	۱ نویسنده
کتاب دوم	۲ نویسنده	۳ نویسنده	۰ نویسنده
کتاب سوم	۱ نویسنده	۰ نویسنده	۵ نویسنده

در جدول بالا نیز سه شیئ تصور شده است که کتاب هستند و بر اساس فراوانی هم رخدادی مشخصه نویسنده به صورت جفتی مقایسه شده اند. سطر و ستون اول از جدول بالا نشان می‌دهد که وقتی کتاب اول و اول (اول با خودش) را به صورت جفتی مقایسه کنیم، می‌بینیم که ۴ نویسنده مختلف آن کتاب را نوشته اند، وقتی که کتاب اول و دوم را به صورت جفتی مقایسه می‌کنیم، می‌بینیم که تعداد ۲ نویسنده از این دو کتاب مشترک(مثل هم) و اگر کتاب اول و سوم را مقایسه کنیم، تعداد ۱ نویسنده از این دو کتاب مشترک هستند. سطرها و ستون های دیگر نیز به همین صورت هم رخدادی نویسندگان کتاب ها را نشان می‌دهد.

روش جاکاردی: از آنجا که مشخصه های تکرار پذیر بیش از یک عضو دارند، می‌توان قواعد مجموعه ها را روی آن ها پیاده کرد.که تعیین میزان اشتراک و اجتماع نمونه ای از آن است. فرضا، یک متن را در نظر بگیریم که مشخصه ای تکرار پذیر به نام واژه دارد. یعنی از واژه های گوناگون تشکیل شده است که مجموعه این واژه ها، واژگان آن نامیده می‌شود. اگر واژگان دو متن را با هم مقایسه کنیم، به احتمال زیاد تعدادی از واژه ها در هر دو متن مشترکا به کار رفته و در برابر تعدادی فقط در یکی از آن دو متن مورد استفاده قرار گرفته است. بنابراین، برای تعیین میزان شباهت این دو متن دو پارامتر وجود دارد: ۱. تعداد واژه های مشترک دو متن و ۲.

مجموع کل واژگان دو متن. تعیین تعداد واژه های مشترک دو متن ساده است، برای تعیین مجموع کل واژگان دو متن، باید تعداد واژه های مشترک را با تعداد واژه های غیر مشترک جمع زد:

مجموع کل واژگان دو متن = تعداد واژه های مشترک دو متن + تعداد واژه های غیر مشترک هر متن

با این کار، عملا اجتماع(⋃) مقادیر دو مجموعه (متن) محاسبه می شود. نهایتا، نسبت بین کل واژگان دو متن و تعداد واژه های مشترک آنان، مقداری را به دست می‌دهد که معیاری دیگر برای مقایسه جفتی مشخصه های اسمی تکرار پذیر است:

$$Jac(x_i, y_j) = \frac{n(x_i \bigcap y_j)}{n(x_i \bigcup y_j)}$$

در فرمول بالا x_iشیئ اول و y_j شیئ دوم، n تعداد مقادیر، $\bigcap$ علامت اشتراک و $\bigcup$ علامت اجتماع فرض شده است و می‌توان گفت مقدار شباهت دو شیئ x_i و y_j عبارت از تعداد مقادیر مشترک بین دو شیئ تقسیم بر مجموع تعداد مقادیر آن دو است. این فرمول را می‌توان به صورت دیگری نگاشت که نشان دهد، چگونه می‌توان مستقیما اشتراک و اجتماع بین مقادیر دو شیئ را حساب کرد:

$$Jac(x_i, y_j) = \frac{\sum x_i y_j}{\sum x_i + \sum y_j + \sum x_i y_j}$$

در آمار و ریاضیات علامت سیگما ($\sum$) به معنای مجموع تعداد است. بنابراین اشتراک بین مقادیر دو شیئ عبارت است از مجموع تعداد مقادیر مشترک بین دو شیئ ($\sum x_i y_j$) و اجتماع آن ها با جمع کردن مجموع تعداد مقادیر شیئ x_i و y_j و مجموع تعداد مقادیر مشترک بین دو شیئ($\sum x_i y_j$)، یعنی ($\sum x_i + \sum y_j + \sum x_i y_j$) به دست می‌آید.

روش محاسبه فوق، روش Jaccard نامیده می‌شود. در این روش، اگر تعداد مقادیر دو مجموعه مورد مقایسه یکسان باشد، نسبت حاصله، نماینده مناسبی برای سنجش شباهت یا عدم شباهت است، اما مساله این جا است که تعداد اعضای دو مجموعه مورد مقایسه همیشه یکسان نیست. به همین دلیل، این روش را می‌توان اصلاح کرد.

روش کوسینوسی: روش کوسینوسی، نوع اصلاح شده جاکاردی است. در این روش، به جای این که مقسوم علیه مستقیما اجتماع دو مجموعه، یعنی به صورت $n(x_i \bigcup y_j)$ باشد، به شکل زیر است:

$$Cos(x_i, y_j) = \frac{n(x_i \bigcap y_j)}{\sqrt{[n(x_i)]^2 \times [n(y_j)]^2}}$$

در روش کوسینوسی، صورت کسر همانند روش جاکاردی است. در مخرج کسر، تعداد x_i و y_j، هر کدام به توان دو می‌رسد و سپس در هم ضرب می‌شود و از نتیجه حاصلضرب، جذر گرفته می‌شود. این عمل را می‌توان به صورت فرمول زیر نیز نمایش داد.

$$Cos(x_i, y_j) = \frac{\sum x_i y_j}{\sqrt{\sum x_i^2 \sum y_j^2}}$$

در هر حال، مقدار حاصله از دو روش، عددی بین ۰ و ۱ است که ۰ عدم شباهت کامل و ۱ شباهت کامل است. به عبارت دیگر، اگر مقدار ۰ بشود یعنی هیچ یک از اعضای دو مجموعه با یکدیگر شباهت ندارند و برعکس، مقدار ۱ نشان می دهد که تمامی مقادیر دو مجموعه عین هم است. باید توجه داشت که مقدار حاصله تنها برای ماتریس های متقارن مناسب است زیرا در این نوع ماتریس ها، مقدار حاصل از سنجش شباهت مورد اول با دوم، با شباهت مورد دوم با اول یکی فرض می‌شود.

برای درک بهتر مطالب بالا، شباهت دو شیئی فرضی را به دو روش یاد شده در بالا محاسبه می‌کنیم. فرض می کنیم که یک شیئ به نام x و شیئ دیگری به نام y داریم که می‌خواهیم آن ها را از نظر یک مشخصه اسمی تکرار پذیر مقایسه کنیم. مشخصه مورد نظر در شیئ x آرایه ای از ۵ مقدار a، f، p، r و w و در شیئ y آرایه ای از ۴ مقدار a، f، h و s است. این مطلب را می‌توان به شکل زیر نمایش داد:

$$y = \{a, f, h, s\} \quad و \quad x = \{a, f, p, r, w\}$$

مقدار a و f در هر دو مجموعه بالا وجود دارد، بنابراین، اشتراک این دو مجموعه ۲ عضو است. پس

$$Jac(x_i, y_j) = \frac{n(x_i \cap y_j)}{n(x_i \cup y_j)} = \frac{\sum x_i y_j}{\sum x_i + \sum y_j + \sum x_i y_j} = \frac{2}{5+4+2} = \frac{2}{11} =\sim 0.18$$

و به روش کسینوسی

$$Cos(x_i, y_j) = \frac{n(x_i \cap y_j)}{\sqrt{[n(x_i)]^2 \times [n(y_j)]^2}} = \frac{\sum x_i y_j}{\sqrt{\sum x_i^2 \sum y_j^2}} = \frac{2}{\sqrt{5^2 \times 4^2}} = \frac{2}{\sqrt{25+16}} = \frac{2}{\sqrt{41}} =\sim 0.31$$

نتیجه حاصله از روش جاکارد حدودا ۰/۱۸ و با روش کوسینوسی حدودا ۰/۳۱ به دست آمد که نشان می‌دهد، مقادیر به دست آمده از این دو روش فرق می‌کند.

اگر شیئ دیگری به نام z را در نظر بگیریم و مقادیر آن را آرایه ای از سه عضو a، j و k فرض کنیم، می توانیم دوماتریس متقارن سه بعدی ترسیم کنیم:

z = {a, j, k}

جدول: ماتریس حاصل از محاسبه دو به دوی مشخصه ها به روش جاکاردی D

	x	y	z
x	۱	۰/۱۸	۰/۱۴
y	۰/۱۸	۱	۰/۱۷
z	۰/۱۴	۰/۱۷	۱

جدول: ماتریس حاصل از محاسبه دو به دوی مشخصه ها به روش کوسینوسی E

	x	y	z
x	۱	۰/۳۱	۰/۱۷
y	۰/۳۱	۱	۰/۲۰
z	۰/۱۷	۰/۲۰	۱

اگر مقادیر خانه های متناظر دو جدول بالا را در نظر بگیریم، خانه ای که کمترین مقدار را در جدول بالا دارد، خانه متناظرش نیز کمترین مقدار را در جدول پائین دارا است. همین طور، خانه ای که بیشترین مقدار را در جدول بالا دارد، خانه متناظرش بیشترین مقدار را در جدول پائین دارا است. اگر سایر خانه های متناظر در دو جدول را مقایسه کنیم، پی خواهیم برد که رتبه خانه های متناظر از لحاظ مقدار یکی است. بنابراین، به کارگیری هر یک از این دو روش، عملا تاثیر زیادی بر نتیجه نهایی نخواهد داشت.

داده های رتبه ای: همان گونه که قبلا ذکر شد داده های رتبه ای نوع دیگری از داده ها هستند. مقایسه دو به دوی مشخصه هایی که ازنوع رتبه ای هستند، مستقیما صورت نمی گیرد. به عنوان مثال اگر توجه کنیم، می بینیم که کسب رتبه در مسابقات، با کسب مقادیر دیگری در ارتباط است. وزنه برداری که رتبه اول را کسب کرده است، قبلا وزنه ای را بلند کرده است که وزن آن یک مقدار بوده است. چون مقدار وزنه او از لحاظ سنگینی بیشتر از بقیه بوده است، توانسته است مقام اول را کسب کند. پس رابطه ای بین کسب یک مقدار و کسب مقام یا رتبه وجود دارد. حال برای دسته بندی وزنه برداران مقام دار، می‌توان میزان وزنه هایی را که بالا برده‌اند در نظر گرفت و هر فرد را با سایر افراد دو به دو مقایسه کرد.

مثال دیگر، رتبه بندی دانشجویان بر اساس معدل درسی است. آن چه که باعث شده است که آن ها شاگرد اول یا چندم باشند، مقدار معدل آن ها است. پس باید آن ها را بر اساس معدلشان دو به دو با هم مقایسه کرد. نمونه دیگر تیم های فوتبال است که رتبه آن ها در لیگ، براساس امتیازشان در بازی ها به دست می‌آید.

وزن وزنه ها، معدل دانشجویان و امتیاز تیم‌های فوتبال همگی مقادیر عددی هستند. از این رو می‌توان گفت که کسب رتبه از سوی اشیاء، ما حصل مقادیرعددیی است که در یک مشخصه کسب کرده اند. همان طور که قبلا آمد، تنها داده های فاصله ای و نسبی است که مقادیر عددی دریافت می‌کنند. به همین علت، وقتی می خواهیم

اشیاء رتبه دار را دسته بندی کنیم، باید از روش هایی بهره ببریم که مخصوص داده های نسبی و فاصله ای یا اسمی است است.

داده های فاصله ای:. وقتی که می خواهیم مقدار عددی دو شیئ از نوع فاصله ای را مقایسه کنیم، کافی است که از عمل تفریق بهره ببریم و آن دو را با هم مقایسه کنیم. به عنوان نمونه، فاصله سن دو نفر به را حتی با کم کردن سن یکی از دیگری ممکن است. وقتی که سن یک نفر ۴۰ سال و سن دیگری ۱۰ سال است، فاصله سن این دو را می توان از طریق عملیات تفریق محاسبه و نتیجه ۳۰ را حاصل کرد. فرض کنیم که می‌خواهیم شهرها را بر اساس متوسط میزان دمای سالانه، دسته بندی کنیم. اگر تعداد شهرهای مورد مقایسه سه شهر و میانگین دمای شهراول، دوم و سوم، به ترتیب ۱۰، ۲۵ و ۳۰ درجه سانتیگراد باشد، نه مقایسه جفتی به شرح زیر خواهیم داشت:

۱. متوسط دمای شهر اول با اول، اول با دوم، اول با سوم.
۲. متوسط دمای شهر دوم با اول، دوم با دوم، دوم با سوم.
۳. متوسط دمای شهر سوم با اول، سوم با دوم،سوم با سوم.

جدول ماتریس حاصل از محاسبه دو به دوی دمای سه شهر F

	شهر اول	شهر دوم	شهر سوم
شهر اول	۰ = |۱۰ – ۱۰|	۱۵ = |۱۰ – ۲۵|	۲۰ = |۱۰ – ۳۰|
شهر دوم	۱۵ س = |۲۵ – ۱۰|	۰ = |۲۵ – ۲۵|	۵ = |۳۰ – ۲۵|
شهر سوم	۲۰ = | ۳۰ – ۱۰|	۵ = |۳۰ – ۲۵|	۰ = |۳۰ – ۳۰|

همان گونه که در جدول آمده است، قدر مطلق نتیجه تفریق مد نظر است، یعنی علامت مثبت یا منفی عدد حاصل در نظر گرفته نمی شود.

داده های نسبی: علاوه بر تفریق، می‌توان برای داده های نسبی روش دیگری را به کار بست و آن مقایسه جفتی بر اساس نسبت بین مقادیر دو شیئ است. برای این کار باید مقدار یکی را بر دیگری تقسیم کرد. در تفریق، جابجایی دو عدد باعث می شود که علامت عدد حاصل بین مثبت و منفی تغییر کند. مثلا ۵+ = ۵ – ۱۰ و ۵- = ۱۰ – ۵ است. در تقسیم نیز جابجایی اعداد بر نتیجه تاثیر می گذارد، یعنی اگر اعداد مثال قبلی را در نظر بگیریم و برای تعیین نسبت بین دو عدد، آن ها را جابجا کنیم، هر بار نتیجه مختلفی به دست می‌آید: $\frac{10}{5}=2$ و $\frac{5}{10}=\frac{1}{2}$. نتیجه های حاصل عکس هم است. بنابر این، برای حل مساله باید یک رویه انتخاب کرد: الف) تقسیم عدد کوچک تر بر بزرگتر یا ب) تقسیم عدد بزرگتر بر کوچکتر.

فرض کنیم سه نفر را داریم که به ترتیب ۳، ۱۵ و ۳۰ سال سن دارند و می خواهیم سن آن ها را به صورت جفتی از طریق روش الف و ب مقایسه کنیم. در هر حال نه مقایسه برای هر روش خواهیم داشت:

۱. نسبت سن نفر اول با اول، اول با دوم، اول با سوم.
۲. نسبت سن نفر دوم با اول، دوم با دوم، دوم با سوم.
۳. نسبت سن نفر سوم با اول، سوم با دوم، سوم با سوم

جدول: مقایسه دو به دوی داده های نسبی بر اساس تقسیم عدد کوچک بر بزرگ (روش الف) H

	نفر اول	نفر دوم	نفر سوم
نفر اول	$\frac{3}{3}=1$	$\frac{3}{15}=\frac{1}{5}$	$\frac{3}{30}=\frac{1}{10}$
نفر دوم	$\frac{3}{15}=\frac{1}{5}$	$\frac{15}{15}=1$	$\frac{15}{30}=\frac{1}{2}$
نفر سوم	$\frac{3}{30}=\frac{1}{10}$	$\frac{15}{30}=\frac{1}{2}$	$\frac{30}{30}=1$

جدول: مقایسه دو به دوی داده های نسبی بر اساس تقسیم عدد بزرگ بر کوچک (روش ب)I

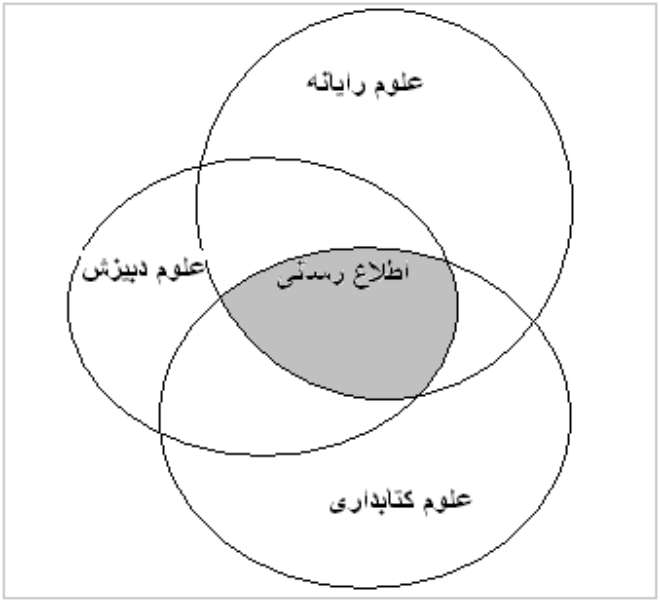
	نفر اول	نفر دوم	نفر سوم
نفر اول	$\frac{3}{3}=1$	$\frac{15}{3}=5$	$\frac{30}{3}=10$
نفر دوم	$\frac{15}{3}=5$	$\frac{15}{15}=1$	$\frac{30}{15}=2$
نفر سوم	$\frac{30}{3}=10$	$\frac{30}{15}=2$	$\frac{30}{30}=1$

با منطبق کردن دو جدول بالا می توان به مقایسه نتایج حاصل از دو روش الف و ب پرداخت. در هر دو جدول، نتیجه مقایسه جفتی هر شیئ با خودش برابر با ۱ است اما در جدول الف هر چه دو مورد شبیه تر باشند، نتیجه حاصله عددی بزرگتر خواهد بود که حداکثر مقدار آن ۱ و حداقل آن مقداری بیشتر از ۰ است . در جدول روش ب، هر چه دو مورد شبیه تر باشند، نتیجه حاصله عددی کوچکتر خواهد بود که حداقل مقدار حاصله ۱ (نشانه بیشترین شباهت) و حداکثر مثبت بی نهایت خواهد بود. اگر مقدار حاصله را x بنامیم، این مساله را می توان چنین بیان کرد: در حالت الف، $۱ \leq x > ۰$ و در حالت ب، $+\infty > x > ۱$.

اکنون نوبت آن است که نگاهی کلی به نتایج جداول حاصل از روش های مختلف مقایسه جفتی بیندازیم. در برخی از روش ها، هر چه مقدار حاصله از مقایسه جفتی بیشتر باشد، شباهت بیشتر (جداول .....) و در برخی، هر چه این مقدار کمتر باشد، شباهت بیشتر است(جداول .....). به حالت اول، اصطلاحا میزان یا مقیاس شباهت و به حالت دوم میزان یا مقیاس عدم شباهت گفته می‌شود.

خوشه بندی در بازیابی اطلاعات

بازیابی اطلاعات را می‌توان معادل بازیابی مدارک دانست. زیرا بازیابی اطلاعات، شاخه ای از علم اطلاع رسانی است و این علم، شکل توسعه یافته علم دبیزش یا دکومانسیون محسوب می شود. علم دبیزش، فناوری‌های موجود در حیطه علم رایانه و هم چنین مبانی و فنون علم کتابداری را به خدمت می گیرد و از این هم پوشانی، علم اطلاع رسانی ظهور می یابد.



تصویر: هم پوشانی علومی که اساس علم اطلاع رسانی را شکل می دهند.

علم دبیزش حیطه فعالیت خود را گردآوری، ارزیابی، تجزیه و تحلیل، ذخیره، بازیابی و اشاعه *مدارک* می‌داند. اگر این فعالیت ها را به حوزه اطلاع رسانی تعمیم بدهیم، آن گاه حیطه فعالیتی آن گردآوری، ارزیابی، تجزیه و تحلیل، ذخیره، بازیابی و اشاعه *اطلاعات* خواهد بود. علاوه بر آن، نیاز به استفاده از اصطلاح عامتر اطلاعات به جای اصطلاح مدرک از آن جا ناشی می شود که تعریفی جامع و مانع از مدرک وجود ندارد. از این رو،

تعیین یک مرز دقیق برای جدا سازی اشیائی که باید مدرک خوانده و اشیائی که باید از این شمول خارج شوند، دشوار است.

رانگاناتان مدرک را خرده اندیشه ای می داند که روی یک سطح صاف نگاشته شده، قابل حمل است و ارزش نگهداری در طول زمان را دارد. از این نگاه، به دشواری می توان چند رسانه ای ها و منابع دیجیتالی را مدرک نامید.

ابرامی اصطلاح کتاب را به جای مدرک به کار برده است. از نظر او "کتاب یک رسانه گروهی است که در آن مطالبی ضبط شده، قابل انتقال است، و بازیابی مطالب آن از نظر زمان و مکان محدود نباشد". گاه منابعی وجود دارد که برای مخاطبین محدود و مخصوصی نگاشته یا تولید شده‌اند، که طبق این تعریف از شمول کتاب(مدرک) خارج می شوند. تعاریف دیگر نیز محدودیت هایی را دربر دارد که سبب می شود تا برخی از منابع اطلاعاتی، مثل اطلاعات دیوار نبشته ها و سنگ نبشته ها مدرک خوانده نشود.

البته، برخی از تعاریف سعی کرده اند تا محدوده ای وسیع تر از منابع اطلاعاتی را مدرک بدانند. .... اظهار می دارد که هر چه حاوی اطلاعات باشد و بتوان از آن اطلاعاتی را به دست آورد یک مدرک نامیده می شود. از این رو، یک بز کوهی در یک باغ وحش یک مدرک خوانده می شود. زیرا می توان از آن اطلاعاتی را درباره این نوع از حیوانات مشاهده و کسب کرد. لوی ... مدرک را نماینده و انعکاس دهنده سخن انسان می داند. او مدرک را نماینده ای می داند که به جای انسان سخن می گوید و با هدفی مشخص و برای مخاطبین خاصی به وجود آمده است. در این تعریف حتما نیازی نیست که مدرک رسانه ای گروهی باشد و می تواند، مثلا، نامه ای شخصی باشد. مدارک را می توان از نظر نوع رسانه ، مانند کاغذی، الکترونیکی، مغناطیسی و غیره نیز محدود کرد. با این همه، هر تعریف محدودیت های خاص خود را ایجاد می کند.

اگر بخواهیم به صورت مصداقی بیان کنیم، در این نوشتار، اصطلاح مدرک به منابع اطلاعاتی چون صفحات وب، مقالات مجلات و کنفرانس‌ها، پایان نامه ها، پروانه های ثبت اختراع، استاندارد نامه ها، گزارش های فنی، نوشته های شخصی و دولتی و شبیه آن ها اطلاق می شود. به عبارتی، هر چه که قابلیت ذخیره شدن در رایانه به شکل متن داشته باشد، در این نوشتار مدرک دانسته می شود.

مشخصه های مدرک

از نگاه خوشه بندی، مدارک دسته ای از اشیاء هستند که می توان روی آن ها فنون دسته بندی خودکار اطلاعات را به کار بست. بنابراین، باید به هر مدرک به عنوان عضوی متعلق به دسته بزرگ اشیا نگریست و مشخصه‌هایی برای آن قائل شد. گروه ها و نهادهای مختلفی سعی کرده اند تا این مشخصه ها را مشخص و استاندارد کنند که هر یک از آن ها برای هدف خاصی به این کار پرداخته اند. در این راستا، کتابخانه کنگره آمریکا استاندارد MARC را ارائه کرده است. این استاندارد، اجزای اطلاعاتی (مشخصه های) یک مدرک را تعیین می کند. از طریق این اجزا می توان مقادیر مربوط به مشخصه های یک مدرک را نشان داد و آن ها را مبادله و تفسیر کرد. این اجزای اطلاعاتی در جدول ..... آورده شده است.....

نهاد دیگری که در زمینه تعیین مشخصه های مدارک فعالیت دارد، موسسه دوبلین کور است. این موسسه نیز به تعیین مشخصه های مدارک پرداخته است و اصطلاحا، این مشخصه ها را عناصر فراداده می نامد. به عبارتی، مدرک یک داده و هر یک ازمشخصه های آن، یک عنصر فراداده فرض شده است. فراداده را باید داده‌هایی درباره یک داده یا اطلاعاتی درباره یک اطلاع خواند. عناصری چون عنوان، خالق اثر، موضوع و امثال آن‌ها داده هایی (مشخصه ها) هستند که اطلاعاتی درباره مدرک می دهند. دوبلین کور تا کنون پانزده عنصر اصلی را مشخص و ارائه کرده است که عناصر ساده نامیده می شوند. علاوه بر نام مشخصه، مقادیر آن را نیز می توان با کمک استانداردی خاص نوشت. مثال واضح آن تاریخ (سال، روز و ماه) است که می توان آن را به اشکال گوناگون نگاشت که در ایران معمولا به صورت "روز/ماه/سال" نوشته می شود. در این رابطه، در طرح دوبلین کور، برای برخی ازعناصر به طور مشخص ذکر شده که چه طرح ها و استانداردهایی ممکن است در تعیین مقدار یک عنصر استفاده شود که به آن ها طرح های نشانه گذاری گفته می شود. مثلا مقدار عنصر موضوع که یک مشخصه از مدرک است، می تواند کدهای رده بندی مثل رده کنگره و دیویی، یا اصطلاحات کنترل شده با واژگان هایی مثل سرعنوان های موضوعی پزشکی (MeSH) و سرعنوان های موضوعی کتابخانه کنگره(LCSH) باشد.

علاوه بر آن، هر عنصر می‌تواند دارای زیر مجموعه ای از عناصر خاص تر باشد که اصطلاحا به آن ها پالایشگر گفته می شود. مثلا عنصری مانند "توصیف اثر" که یکی از عناصر پانزده گانه است، می تواند دارای زیر مجموعه ای از دو عضو "فهرست مندرجات" و "چکیده" باشد. به عبارتی، این دو عضو، عناصری هستند که زیر مجموعه عنصر "توصیف اثر" قرار می گیرند.

همان گونه که دیدیم، روی مشخصه های مدرک توافق نظر وجود ندارد. یعنی، هر موسسه یا شخص بر اساس هدفی که دارد مشخصه هایی را تعیین و تحت نام هایی خاص مانند اجزای کتابشناختی، عناصر فراداده یا نام هایی از این قبیل ارائه می کند. از این رو، تعداد مشخصه ها در هر طرح و استانداردی متفاوت است. مثلا، استاندارد مارک مشخصه های مدرک را بسیار ریز تعیین کرده است و به این جهت، تعداد مشخصه های اصلی و فرعی مدرک به ۳۵۰ مورد بالغ می شود. در برابر، دوبلین کور نگاهی عام به این مساله دارد و تعداد این مشخصه ها را پانزده مورد تعیین کرده است.

هر شاخه از حوزه علوم کتابداری و اطلاع رسانی به طور خاص به یک یا چند مشخصه مدرک توجه ویژه دارد. مثلا کتابسنجی و علم سنجی توجه ویژه ای به مشخصه استناد دارد. مقدار این مشخصه، منابعی است که در بخش منابع و مآخذ (یا همان بخش کتابشناسی یا کتابنامه) در انتهای هر فصل یا در انتهای مدرک، و یا به صورت زیر نویس آورده شده است. پدیدآور مدرک، مشخصه مورد توجه دیگری در این شاخه از علم است. اگر پژوهشگران این شاخه بخواهند از فنون خوشه بندی بهره ببرند، مسلما این دو مشخصه را بیشتر از همه مورد توجه قرار خواهند داد.

در شاخه بازیابی اطلاعات، به عنوان شاخه ای دیگر از علوم کتابداری و اطلاع رسانی، به طور ویژه به مشخصه "موضوع" توجه می شود. علاوه بر آن، واژه ها و واژگان به کار رفته در کل متن نیز، به عنوان مشخصه ای از یک مدرک، در این شاخه از علم مورد توجه است. زیرا، واژه، یکی از مبناهای تعیین و تشخیص موضوع است.

مشخصه موضوع، علاوه بر این نام، با عناوینی چون اصطلاح موضوعی، اصطلاح نمایه ای، سرعنوان موضوعی، توصیفگر، کلید واژه و حتی رده موضوعی شناخته می شود. نقطه مشترک بین این عناوین این است که همه این ها، واژه یا عباراتی هستند که به طور بسیار مختصر و در کوتاه ترین شکل ممکن، محتوای یک مدرک یا نیاز اطلاعاتی یک کاربر را بیان می کنند. .... درباره این اصطلاحات صحبت خواهیم کرد و هم چنین به فنون و روش های تعیین موضوعات خواهیم پرداخت.

مراحل خوشه بندی مدارک

در حوزه بازیابی اطلاعات، برای خوشه بندی مدارک یا موضوعات مراحل گوناگونی طی می شود: ۱. استخراج مدارک، ۲. استخراج اصطلاحات موضوعی، ۳. تشکیل ماتریس های متقارن

استخراج مدارک

در خوشه بندی، قبل از هر چیز باید مدارک مرتبط با موضوع یا زمینه موضوعی مورد نظر مشخص و استخراج شود. ساده ترین روش برای برگزیدن مدارک مرتبط، جستجوی موضوعی یا کلید واژه ای در پایگاه های اطلاعاتی یا صفحات وب است. مدارکی که حاصل این جستجوها هستند، می تواند در فرآیند خوشه بندی مورد استفاده قرار گیرد. اما یک جستجو می تواند منجر به بازیابی ده ها، صدها، هزاران و میلیون ها مدرک شود. در این حالت، تعداد مقایسه های دو به دوی یک مدرک با مدارک دیگر یا یک موضوع با موضوعات دیگر، بسیار بالا خواهد رفت که در عمل بسیار وقت گیر و حتی ناممکن خواهد بود. برای کاهش تعداد مقایسه ها می توان با ترفندهای گوناگون تنها مدارک شاخص تر را بازیابی کرد. برای این کار، یکی از راه حل های معمول، برگزیدن آن مدارکی است که به موضوع یا حوزه موضوعی مورد نظر ما مرتبط تر هستند. مثلا اگر بخواهیم مدارک مربوط به بیماری های روانی را خوشه بندی کنیم، با حجم بزرگی از مدارک رو به رو خواهیم شد، بنا براین، آن مدارکی را بر می گزینیم که موضوع اصلی آن ها بیماری های روانی است، نه هر مدرکی که تنها اشاره ای به این مساله داشته باشد. یکی از مزایای این کار بالا بردن دقت در بازیابی مدارک و تولید خوشه های دقیق از مدارک بازیابی شده است. منظور از دقت در بازیابی، بالا بودن نسبت تعداد مدارک مربوط، به تعداد کل مدارک بازیابی شده است که اصطلاحا به آن ضریب دقت گفته می شود:

مدارک بازیابی شده مرتبط / کل مدارک بازیابی شده = ضریب دقت

البته با این روش میزان بازیافت کاهش می یابد. یعنی احتمال این که تعدادی از مقالات مرتبط را نیز از دست بدهیم وجود دارد، در واقع، همیشه بین دقت و بازیافت نتیجه عکس وجود دارد. با بالا رفتن دقت، بازیافت کاهش می یابد و همین طور بازیافت بالا مستلزم دقتی پائین است. به عبارتی، با بالا بردن دقت، نسبت مدارک مرتبط بازیابی شده به کل مدارک موجود پائین می آید:

ضریب بازیافت= مدارک بازیابی شده مرتبط / کل مدارک مرتبط موجود در مجموعه مدارک

برای بالا بردن دقت بازیابی و در نتیجه کاهش تعداد مدارک بازیابی شدهراه های گوناگونی وجود دارد که در زیر به برخی از آن ها می پردازیم:

جستجو درعنوان: وجود اصطلاح مورد نظر در عنوان مدرک، شاخصی خوب برای مرتبط بودن آن مدرک به حوزه موضوعی مورد نظر ما و بالا بردن دقت بازیابی است. معمولا اگر یک اصطلاح موضوعی در عنوان یک مدرک وجود داشته باشد، احتمال ارتباط آن مدرک با موضوع یا حوزه موضوعی مورد نظر بیشتر می شود. فرضا اگر به دنبال مقالاتی باشیم که در حوزه درمان بیماری های روانی هستند، اگر یک اصطلاح درمانی، مثل "خانواده درمانی"، در عنوان مقاله به کار رفته باشد، به احتمال زیاد آن مقاله ارتباط زیادی با حوزه درمان بیماری های روانی دارد. در این جا از کل مقالاتی که می تواند در حوزه یاد شده باشد، تنها مقالاتی بازیابی می شود که به احتمال بیشتر با این حوزه مرتبط هستند.

جستجو در کلید واژه ها: معمولا در هر مدرک به دنبال چکیده، بخشی به نام کلید واژه ها وجود دارد که حاوی موضوعات داده شده به آن مدرک توسط نویسنده است. این کلید واژه ها می تواند مبنایی برای برگزیدن مدارک باشد. مدرکی که موضوع مورد نظر ما به عنوان کلید واژه آن محسوب می شود، به احتمال بیش تر با نیاز اطلاعاتی ما مرتبط است و این امر می تواند دقت جستجوی ما را افزایش دهد.

چکیده:

این نوشتار، با معرفي پروتكل پيشنهادي PMH ، بر آن است تا نشان دهد كه چگونه مي توان از اين پروتكل براي ايجاد فهرستگان اسناد آرشيوي بهره گرفت. هدف نهائي اين پروتكل، تسهيل در انتقال اطلاعات به صورت فراداده در محيط شبكه است. PMH ، با بهره گيري از عناصر فراداده دوبلين كور، قالب (فرمت) XML (زبان نشانه گذاري گسترش پذير) و پروتكل HTTP اين امر را ميسر مي سازد. بدين منظور، مقاله به تعريف فراداده و معرفي عناصر فراداده اي دوبلين كور مي پردازد و ساختار پيشنهادي ركورد فراداده را براساس الگوي پيشنهادي در قالب XML نشان مي دهد. همچنين، توضيحاتي درباره شش فرمان قراردادي براي درخواست فراداده از طريق پروتكل HTTP ارائه مي كند، و نشان مي دهد كه چگونه دارندگان اسناد، مي توانند نقش داده فرآور و متوليان امور اسناد، نقش خدمت فرآور را ايفا كنند. در پايان، به مزاياي اصلي و جنبي استفاده از اين پروتكل اشاره شده است.

كليدواژگان: فهرستگان ها/ اسناد الكترونيكي/ انتقال/ پروتكل OAI، PMH/ زبان هاي نشانه گذاري/ فراداده ها/ پروتكل دوبلين كور/ زبان نشانه گذاري گسترش پذير/ پروتكل فرا متن

نمونه ای از یک چکیده همراه با کلید واژه

جستجو در توصیفگرها: روش موثر دیگر برای کاهش مدارک بازیابی شده، برگزیدن مدارکی است که قبلا توسط نمایه سازان، موضوعات مورد نظر ما را دریافت کرده اند. معمولا سازمان هایی که به تهیه پایگاه های اطلاعاتی اشتغال دارند، از متخصصان موضوعی یا کتابداران و اطلاع رسانان برای موضوع دادن به مدارک آن پایگاه کمک می گیرند. به این موضوعات اصطلاحا توصیفگر، اصطلاح موضوعی، سرعنوان موضوعی، نمایه موضوعی و مواردی از این قبیل گفته می شود.

ERIC #:	
A unique accession number assigned to each record in the database; also referred to as ERIC Document Number (ED Number) and ERIC Journal Number (EJ Number).	EJ912314
Title:	
The name assigned to the document by the author. This field may also contain sub-titles, series names, and report numbers.	Teacher Qualifications and School Climate: Examining Their Interrelationship for School Improvement
Authors:	
Personal author, compiler, or editor name(s); click on any author to run a new search on that name.	DeAngelis, Karen J.; Presley, Jennifer B.
Descriptors:	
Terms from the Thesaurus of ERIC Descriptors; used to tag materials by subject to aid information search and retrieval. Click on a	Teacher Qualifications; Academic Achievement; Organizational Climate; Educational Improvement; Performance Factors; Educational Environment; Predictor Variables; Correlation; Schematic Studies; Data Analysis

نمونه ای از یک رکورد کتابشناختی در پایگاه اطلاعاتی ERIC همراه با توصیفگرهای آن

در بالا، بخشی از یک رکورد اطلاعاتی از پایگاه اطلاعاتی ERIC آورده شده است. در بخش Descriptors، موضوعاتی آمده که توسط نمایه سازان برای این مقاله برگزیده شده است. این اصطلاحات را می توان از جمله موضوعات شاخص برای این مدرک برشمرد.

.

استخراج اصطلاحات موضوعی

پس از آن که مدارک مورد نظر را استخراج کردیم، باید اصطلاحات موضوعی مدارک را نیز شناسایی و استخراج کنیم. عنوان بخش مهمی از مدرک است که از واژه های به کار رفته در آن می توان اصطلاحات موضوعی را استخراج کرد. در کنار عنوان، می توان چکیده مدرک را نیز در نظر گرفت و اصطلاحات موضوعی را در آن بخش جست. از متن هر مدرک نیز می توان اصطلاحات موضوعی را به دست آورد، اما احتمال مناسب بودن اصطلاحاتی که از متن حاصل شده است متن کمتر است. از این رو، استخراج اصطلاحات از عنوان و چکیده توصیه می شود. همین طور می توان کلید واژه های داده شده توسط نویسنده یا موضوعات تعیین شده توسط نمایه سازان را به عنوان اصطلاحات موضوعی در نظر گرفت.

نمایه سازی دستی، نیمه خودکار و خودکار

همان گونه که دیدیم، اصطلاحات موضوعی اساس کار خوشه بندی در حوزه بازیابی اطلاعات است. بنابراین، قبل از هر چیز باید آموخت که چگونه می توان آن ها را تعیین و از مدارک استخراج کرد. به عمل تعیین و استخراج موضوعات، اصطلاحا نمایه سازی موضوعی گفته می شود. موضوعاتی که به در نمایه سازی موضوعی تعیین می شوند، عملا اصطلاحاتی هستند که در کوتاه ترین حالت ممکن بتوانند یک مدرک را توصیف کنند، به همین جهت به آن ها توصیفگر نیز گفته می شود. روش های فراوانی برای استخراج موضوعات وجود دارد. این روش ها را می توان به سه گروه دستی، نیمه خودکار و خودکار تقسیم کرد.

برای استخراج موضوعات یک مدرک می توان به سه روش عمل کرد:

- استخراج موضوعات به روش دستی
- **استخراج توصیفگرها و کلید واژه های موجود**
- **استخراج موضوعات به روش نمایه سازی خودکار**

استخراج موضوعات به روش دستی

منظور از روش دستی، موضوع دادن به مدارک مستقیما توسط اشخاص است. انتساب کلید واژه به مدارک می تواند توسط نویسندگان ، متخصصان موضوعی، یا حرفه‌مندان کتابداری و اطلاع رسانی(نمایه سازان) صورت گیرد.

در روش نیمه دستی معمولا موضوعات مدارک را رایانه مشخص می کند، ولی انتخاب و گزینش نهایی موضوعات مستخرج از طریق رایانه به افراد واگذار می شود. در حالیکه، در روش خودکار، تمامی فرآیند مطالعه و استخراج و تعیین نهایی اصطلاحات موضوعی توسط رایانه و بدون دخالت انسانی صورت می گیرد.

بر اساس خطمشی های نمایه سازی، روند کار برای تعیین اصطلاحات در هر یک از سه روش دستی، نیمه خودکار و خودکار فرق می کند. دو مشی برای این کار محتمل است: اول، تعیین اصطلاحات موضوعی بدون تغییر شکل آن ها و دقیقا مانند آن چه که در متن ظاهر شده است که به این نوع از موضوعات، اصطلاحات طبیعی گفته می شود. دوم، بیان اصطلاحات از طریق واژه هایی که الزاما عین واژه های به کار رفته در متن نیست که این ها را اصطلاحات مصنوعی می خوانند. اصطلاحات طبیعی معمولا حاصل روش های خودکار و نیمه خودکار و اصطلاحات مصنوعی بیشتر حاصل نمایه سازی دستی است.

اصطلاحات مصنوعی را می توان به صورت کنترل شده در آورد. برای این کار از ابزارهایی مانند اصطلاحنامه(تزاروس) و فهرست سرعنوان های موضوعی استفاده می شود که لیستی از موضوعات استاندارد شده برای موضوع دادن به مدارک را عرضه می کنند و علاوه بر آن روابط معنایی بین موضوعات را مشخص می کنند. به این صورت که در زیر هر موضوع مجاز (انتخاب شده)، اصطلاحات مترادف، خاص تر، عام تر و مرتبط با آن، مانند شکل .....، نیز نشان داده می شود. نمایه‌ساز پس از بررسی مدرک و تعیین موضوعات آن، به این منابع رجوع می کند، اگر موضوعی که به ذهن نمایه ساز آمده است، دقیقا عین اصطلاح موجود در اصطلاحنامه یا فهرست سرعنوان های موضوعی باشد، آن را به عنوان موضوع آن مدرک بر می گزیند، در غیر این صورت باید اصطلاحی را به کار ببرد که در آن منابع به آن ارجاع داده شده است. مثلا نمایه سازی پس از مطالعه یک مدرک موضوع تعلیم و تربیت را برای آن در نظر می گیرد و سپس برای استاندارد کردن موضوعی که به ذهنش رسیده است به فهرست سرعنوان های موضوعی فارسی مراجعه می کند و در این فهرست، به ترتیب الفبا در زیر حرف "ت" به دنبال اصطلاح "تعلیم و تربیت" می گردد و در آن جا می بیند که این منبع از

این اصطلاح به "آموزش و پرورش" ارجاع داده است، لذا اصطلاح آموزش و پرورش را، که یک اصطلاح انتخاب شده است، به عنوان موضوع آن مدرک بر می گزیند.

ارجاع فوق در فهرست سرعنوان های موضوعی به این شکل آمده است:

تعلیم و تربیت

نک آموزش و پرورش

Article I. Data Processing

Descriptor Details

Record Type:	
Indicates the status of a term: Main (term used for subject indexing); Synonym	Main
(not used for indexing – *see* Use Term); or Dead (no longer used for indexing; *see* Scope Note for more information).	
Scope Note:	Systematic handling, manipulation, and computation of information by machines
A definition or description of what the term covers; may clarify an ambiguous term or indicate a special meaning in the field of education.	
Category:	Information/Communications Systems
Any of 41 broad subject areas in the ERIC *Thesaurus*; each Descriptor is assigned to a category.	
Broader Terms:	Information Processing;
Suggested additional or alternate search terms that are less specific than the original term searched.	
Narrower Terms:	Natural Language Processing;
Suggested additional or alternate search terms that are more specific than the original term searched.	
Related Terms:	Automation; Calculators; CD ROMs (2004); Computer Centers; Computer Interfa

نمونه ای از یک مدخل در تزاروس Eric

استخراج توصیفگرها و کلید واژه های موجود

مسلما کسانی که می خواهند خوشه بندی کنند، امکان نمایه سازی حجم بزرگی از مدارک را به صورت دستی و کنترل شده ندارند، اما می توانند از اصطلاحاتی که دیگران (نویسندگان و نمایه سازان) تعیین کرده اند برای این کار بهره ببرند. کلید واژه هایی که نویسندگان معمولا به مدارک خود می دهند، وسیله ای مناسب برای این کار است. علاوه بر آن، رکوردهای موجود در پایگاههای اطلاعاتی که حاوی اطلاعات کتابشناختی مانند نویسنده، عنوان، ناشر، وغیره هستند، در بردارنده توصیفگرها (موضوعات کنترل شده ای) هستند که قبلا توسط نمایه سازان به این مدارک داده شده است. مثلا، رکوردهای کتابشناختی مدلاین که یک پایگاه اطلاعاتی حاوی منابع اطلاعاتی مربوط به پزشکی است، توسط نمایه سازان و با استفاده از فهرست سرعنوان های موضوعی پزشکی، معروف به MeSH، موضوع داده شده اند که می توان از آن موضوعات در خوشه بندی استفاده کرد. برای این کار کافی است که رکوردهای مرتبط از پایگاه یا پایگاه های مورد نظر را ذخیره و با روش های پردازش روی متن، موضوعات اختصاص داده شده به آنها را استخراج کرد. در زیر به چگونگی این کار می پردازیم:

معمولا، همانگونه که در شکل...نشان داده شده است، کلید واژه ها در پایان چکیده می آیند که با علائمی چون کاما، نقطه کاما، خط تیره و علائم دیگر از یکدیگر جدا می شوند. توصیفگرها نیز در رکورد های کتابشناختی مانند شکل .... و .....، وجود دارند. این توصیفگرها در فایل های متنی مانند فایل های XML،HTML ، متن خالص (plaint text) قرار دارند که باید با روش های پردازش متن(توسط برنامه های رایانه ای موجود مانند..... یا برنامه هایی که خود به این منظور با زبان های برنامه نویسی تهیه نموده ایم) ا یک به یک جدا و برای فرآیندهای بعدی به شکلی مناسب ذخیره شوند. در ادامه، به عنوان نمونه نشان داده می شود که چگونه با پردازش متن، می توان توصیفگرهای رکوردهای مدلاین را استخراج کرد:

```
PMID- 21343569
OWN - NLM
STAT- MEDLINE
DA  - 20110223
DCOM- 20110224
IS  - 1538-3598 (Electronic)
IS  - 0098-7484 (Linking)
VI  - 305
IP  - 8
DP  - 2011 Feb 23
TI  - "Disappointing" trial results offer hope for older
women with breast cancer.
PG  - 765-6
FAU - Voelker, Rebecca
AU  - Voelker R
LA  - eng
PT  - News
PL  - United States
TA  - JAMA
JT  - JAMA : the journal of the American Medical
Association
JID - 7501160
RN  - 0 (Bone Density Conservation Agents)
RN  - 0 (Diphosphonates)
RN  - 0 (Estrogens)
RN  - 0 (Imidazoles)
RN  - 118072-93-8 (zoledronic acid)
SB  - AIM
SB  - IM
MH  - Aged
MH  - Bone Density Conservation Agents/*therapeutic use
MH  - Breast Neoplasms/*drug therapy
MH  - Clinical Trials, Phase III as Topic
MH  - Combined Modality Therapy
MH  - Diphosphonates/*therapeutic use
MH  - Disease Progression
MH  - Disease-Free Survival
MH  - Estrogens/blood
MH  - Female
MH  - Humans
MH  - Imidazoles/*therapeutic use
MH  - Menopause
MH  - Neoplasm Recurrence, Local/prevention & control
MH  - Treatment Failure
EDAT- 2011/02/24 06:00
MHDA- 2011/02/25 06:00
CRDT- 2011/02/24 06:00
AID - 305/8/765 [pii]
AID - 10.1001/jama.2011.172 [doi]
PST - ppublish
SO  - JAMA. 2011 Feb 23;305(8):765-6.
```

نمونه ای از یک رکورد مدلاین

شکل بالا نمونه ای از یک رکورد مدلاین است. یکی از فیلدهای آن با برچسب MH مشخص و نیز تکرار شده است. این فیلد حاوی موضوعات کنترل شده ای است که توسط نمایه سازان کتابخانه ملی پزشکی آمریکا به بیش از هفده میلیون رکورد اطلاعاتی داده شده است. می توان مجموعه ای از رکوردها را که از مدلاین ذخیره شده است توسط یک نرم افزار پردازش متن مثل ..... بررسی و موضوعات را از آنها استخراج کرد و عملیات مورد نظر را روی موضوعات استخراج شده انجام داد.

به این منظور، باید برنامه ما از ابتدا هر رکورد را خط به خط بخواند. هنگامی که به خطی رسید که با - MH شروع شده است، رشته بعد از آن را در جایی شبیه جدول بانک اطلاعاتی یا دداشت کند و در ستون دیگر مقابل آن شماره رکورد مدرک را بیاورد تا برای فرآیندهای بعدی، مانند تهیه واژگان موضوعی و تعیین فراوانی هم رخدادی آماده و قابل استفاده باشد.

پس از استخراج موضوعات هر مدرک، باید واژگانی از موضوعات به دست آمده ایجاد کرد. برای ایجاد این واژگان موضوعی، همه موضوعات مدارک را در یک جا جمع می کنیم. اگر موضوعی در بیش از یک مدرک وجود دارد، تنها یک بار حساب می شود. پس از این کار، سیاهه ای از موضوعات، شبیه جدول زیر به دست می آید که واژگان موضوعی مدارک ما محسوب می شود.

اصطلاح موضوعی	تعداد مدارک
بودجه	۱۲۷
امر به معروف و نهی از منکر	۱۲۷
روزنامه نگار ی	۱۲۶
د استانهای عربی	۱۲۶
د استانهای حماسی	۱۲۶
بهد اشت شخصی	۱۲۶
امنیت ملی	۱۲۶

عروض فارسی	۱۲۵
سی (زبان برنامه نویسی کامپیوتر)	۱۲۵
سرگرمی‌ها	۱۲۵
زبان شناسی	۱۲۵
تمدن اسلامی	۱۲۵
همسرگزینی	۱۲۴
دوستی	۱۲۴
حقوق اساسی	۱۲۴
تکنولوژ ی آموزشی	۱۲۴
تلویزیون	۱۲۴
کتابهای تصویر ی	۱۲۳
کار آموز ی (پزشکی)	۱۲۳
نمایشنامه آمریکایی	۱۲۳
نقشه برد ار ی	۱۲۳
گیاه درمانی	۱۲۳
شهرد ار ی	۱۲۳
رو انشناسی اجتما عی	۱۲۳
حزب توده ایر ان	۱۲۳
تحولات اجتما عی	۱۲۳
تاسیسات	۱۲۳
چین	۱۲۳
بیماریهای رو انی	۱۲۲

نقد ادبی	۱۲۱
عمر ان منطقه ای	۱۲۱
عشق (عرفان)	۱۲۱

نمونه ای از یک واژگان موضوعی همراه با تعداد مدارک حاوی هر موضوع

معمولا فراوانی بخش بزرگی از موضوعات حاصل شده برابر با یک است، یعنی بسیاری از موضوعات تنها در یک مدرک ظاهر شده اند. بخش بزرگ دیگری از موضوعات نیز وجود دارد که تنها در معدودی از مدارک مورد استفاده قرار گرفته‌اند و نهایتا تعداد کمتری از موضوعات است که فراوانی بالایی دارند. این حالت، همانند قاعده وضعیت واژه‌های به کار رفته در یک متن است که اگر طبق روش زیپف بر اساس تعداد تکرارشان رتبه‌بندی شوند یک سوم پایانی سیاهه واژگانی، معمولا دربردارنده واژه‌هایی است که تنها یک بار مورد استفاده قرار گرفته‌اند و در آغاز سیاهه، واژه هایی قرار دارند که فراوانی استفاده از آن ها بالاست اما تعداد چنین واژه هایی به نسبت اندک است.

تشکیل ماتریس متقارن

مشخصه موضوع، داده ای اسمی و تکرار پذیر است. از این رو، یک مدرک می تواند بیش از یک موضوع داشته باشد. به زبان دیگر، مشخصه موضوع می تواند آرایه ای از مقادیر را کسب کند. هر مقدار از این آرایه، یک رشته محسوب می شود. مثلا، اگر مقاله ای داشته باشیم که سه کلید واژه به آن منتسب کرده باشند، آرایه ای از سه عضو داریم که همگی از نوع رشته ای هستند.

به علت تکرار پذیر و اسمی بودن مشخصه موضوع، برای مقایسه دو به دوی مقادیر موضوعی دو مدرک، می توان از روش های تعیین "فراوانی هم رخدادی" و "اشتراک موضوعی" بین دو مدرک بهره برد. فرض کنیم که پنج مدرک داریم و می خواهیم از نظر مشخصه مدرک، آن ها را به صورت جفتی با روش اول مقایسه کنیم. هر یک از این مدارک آرایه ای از موضوعات دارند که می تواند با آرایه دیگر کاملا مشابه یا کاملا نا مشابه و یا تنها در تعدادی از موضوعات مشابه باشد. برای درک بهتر این مطلب، در جدولی به شکل ماتریس نامتقارن، آرایه ای فرضی از موضوعات آن پنج مدرک را نشان می دهیم:

جمع	مدرک۵	مدرک۴	مدرک۳	مدرک۲	مدرک ۱	
۳	۰	۰	۱	۱	۱	موضوع ۱
۳	۰	۱	۰	۱	۱	موضوع ۲
۲	۰	۱	۰	۱	۰	موضوع ۳
۲	۱	۱	۰	۰	۰	موضوع ۴
۳	۱	۰	۱	۰	۱	موضوع ۵
۴	۱	۰	۱	۱	۱	موضوع ۶
۲	۰	۰	۱	۱	۰	موضوع ۷
۳	۱	۱	۰	۰	۱	موضوع ۸
۱	۰	۰	۱	۰	۰	موضوع ۹
۴	۱	۱	۰	۱	۱	موضوع ۱۰
	۵	۵	۵	۶	۶	جمع

در جدول بالا، مجموعه ای از پنج مدرک فرض شده است. ستون های این جدول با نام های "مدرک ۱" تا "مدرک ۵" نام گذاری شده اند. این پنج مدرک، مجموعا ده موضوع تولید کرده‌اند که آن ها نیز در هر ردیف از جدول، از "موضوع ۱" تا "موضوع ۱۰" نامگذاری شده اند. اعداد ۱ و ۰ در خانه های جدول، به ترتیب نشانگر وجود و عدم وجود موضوع در مدرک است. بر اساس قاعده دو طرفه شیئ–مشخصه، از جدول بالا می توان دو نوع ماتریس ایجاد کرد که یکی به خوشه بندی مدارک و دیگری به خوشه بندی موضوعات نظر دارد.

خوشه بندی مدارک

در بازیابی اطلاعات گاه نیاز است که مدارک بازیابی شده را بر اساس شباهت موضوعی دسته بندی کنیم که به آن خوشه بندی مدارک می گویند. برای این کار می توان هر مدرک را با مدارک دیگر دو به دو مقایسه کرد و میزان موضوعات مشترک آن مدارک را به دست آورد. این کار، عملا تعیین تعداد موضوعاتی است که در هر دو مدرک مورد مقایسه به صورت مشترک رخ داده است.

به عنوان نمونه، جدول زیر، ماتریسی ۵ × ۵ است که بر اساس تعیین تعداد موضوعات مشترک هر مدرک با مدرک دیگر از جدول ..... حاصل شده است.

	مدرک ۱	مدرک ۲	مدرک ۳	مدرک ۴	مدرک ۵
مدرک ۱	۶ موضوع	۴ موضوع	۳ موضوع	۳ موضوع	۴ موضوع
مدرک ۲	۴ موضوع	۶ موضوع	۳ موضوع	۳ موضوع	۲ موضوع
مدرک ۳	۳ موضوع	۳ موضوع	۵ موضوع	۰ موضوع	۲ موضوع
مدرک ۴	۳ موضوع	۳ موضوع	۰ موضوع	۵ موضوع	۳ موضوع
مدرک ۵	۴ موضوع	۲ موضوع	۲ موضوع	۳ موضوع	۵ موضوع

در جدول بالا، اعداد درون خانه ها نشان می دهد که هر مدرک با مدرک دیگر در چند موضوع مشترک است، به عبارتی، فراوانی هم رخدادی مدارک چقدر است. مثلا، اگر به خانه هایی که از تقاطع مدرک ۳ و مدرک ۴ به دست آمده است نگاه کنیم، می بینیم که نتیجه محاسبه هم رخدادی این دو مدرک صفر است. به عبارتی هیچ موضوعی بین این دو مشترک نیست.

علاوه بر تعیین تعداد هم رخدادی، می توان از روش جاکاردی و کوسینوسی نیز مدارک را مقایسه کرد اما در این جا به جای تعداد، نسبت اشتراک موضوعی هر مدرک با مدرک دیگر سنجیده می شود :

	مدرک ۱	مدرک ۲	مدرک ۳	مدرک ۴	مدرک ۵
مدرک ۱	اشتراک موضوعی $\frac{6}{6}=1$	$\frac{4}{8}=0.5$			
مدرک ۲	اشتراک موضوعی $\frac{4}{8}=0.5$				
مدرک ۳	اشتراک موضوعی $\frac{4}{8}=0.5$				

مدرک ۴					
مدرک ۵					

در جدول بالا از فرمول $Jac(x_i, y_j) = \frac{n(x_i \cap y_j)}{n(x_i \cup y_j)} = \frac{\sum x_i y_j}{\sum x_i + \sum y_j + \sum x_i y_j}$ استفاده شده و با توجه به جدول .... اعداد درون ماتریس مشخص شده است. اگر با فرمول کوسینوسی اعداد درون خانه های ماتریس را محاسبه کنیم، جدول زیر حاصل خواهد شد:

	مدرک ۱	مدرک ۲	مدرک ۳	مدرک ۴	مدرک ۵
مدرک ۱	$\frac{6}{6}=1$ اشتراک موضوعی				
مدرک ۲	$\frac{4}{8}=0.5$ اشتراک موضوعی				
مدرک ۳	$\frac{4}{8}=0.5$ اشتراک موضوعی				
مدرک ۴					
مدرک ۵					

مقایسه جفتی در جدول بالا از فرمول زیر صورت گرفته است.

$$Cos(x_i, y_j) = \frac{n(x_i \cap y_j)}{\sqrt{[n(x_i)]^2 \times [n(y_j)]^2}} = \frac{\sum x_i y_j}{\sqrt{\sum x_i^2 \sum y_j^2}}$$

برای تعیین میزان شباهت بین مدارک و خوشه بندی آن ها، می توان خانه های هر یک از سه ماتریس بالا را به عنوان نقاط برداری در نظر گرفت اما ماتریس های حاصل از روش های جاکاردی و کوسینوسی برای خوشه بندی مدارک مناسب تر است.

خوشه بندی موضوعات

علاوه بر دسته بندی مدارک، بر اساس قاعده دو طرفه مشخصه–شیئ می توان موضوعات مستخرج از مدارک را نیز دسته بندی کرد که به این کار خوشه بندی موضوعی گفته می شود. نتیجه این کار چیزی شبیه ایجاد یک طبقه بندی موضوعی است. در این حالت، قبلا باید موضوعات مورد نظر برای دسته بندی را مشخص کرد. مسلما تعداد موضوعات استخراج شده از مجموعه ای از مدارک زیاد خواهد بود، برای کاهش تعداد موضوعات، همان گونه که با ترفندهایی مدارک مرتبط تر را استخراج می کردیم، باید سعی شود موضوعات مرتبط تر به حوزه موضوعی برگزیده شود. برای گزینش موضوعات مرتبط تر می توان از متخصصان موضوعی کمک گرفت. وی می تواند در فهرست موضوعات استخراج شده بگردد و آن موضوعاتی که از نظر او ارتباط بیشتری دارند، برگزیند. این کار در صورتی میسر است که قرار نباشد تمام فرآیند به صورت خودکار صورت بگیرد. اما، اگر بخواهیم گزینش موضوعات برای خوشه بندی را کاملا به صورت خودکار انجام دهیم (مثل خوشه بندی موضوعات در یک موتور جستجوی اینترنی) یکی از را ه ها ی ممکن، توجه به فراوانی موضوعات است. یعنی آن موضوعاتی را برگزینیم که در مدارک بیشتری ظاهر شده است. برای خوشه بندی موضوعات، می توان حدی را تعیین کرد. مثلا، ممکن است صد موضوع مرتبط برگزیده شود.

علاوه بر روش یاد شده، می توان موضوعاتی را برگزید که وزن بیشتری نسبت به سایرین دارند. این در صورتی ممکن است که وزن سنجی اصطلاحات موضوعی به روش تعیین "وزن اصطلاح برای مجموعه ای از مدارک" صورت گرفته باشد. در این جا نیز می توان اصطلاحات موضوعی را بر اساس وزنشان از زیاد به کم مرتب کرد. پس از آن، می توان از لحاظ تعداد، حدی را در نظر گرفت و موضوعات را از سیاهه اصطلاحات مستخرج، به تعداد تعیین شده برگزید.

پس از گزینش موضوعات برای دسته بندی، مانند قبل می توان سه نوع ماتریس ایجاد کرد:

۱. ماتریس هم رخدادی موضوعات
۲. ماتریس اشتراک بین موضوعات از روش جاکاردی
۳. ماتریس اشتراک بین موضوعات از روش کوسینوسی

اگر جدول .... را مبنا قرار دهیم، این ماتریس ها به شکل زیر خواهند بود:

جدول: ماتریس هم رخدادی موضوعات مدارک

	موضوع۱	موضوع۲	موضوع۳	موضوع۴	موضوع۵	موضوع۶	موضوع۷	موضوع۸	موضوع۹	موضوع۱۰

	۲ مدرک	۱ مدرک	۱ مدرک	۲ مدرک	۳ مدرک	۲ مدرک	۱ مدرک	۰ مدرک	۱ مدرک	۲ مدرک	۳ مدرک
موضوع۱										۳ مدرک	۲ مدرک
موضوع۲									۳ مدرک	۲ مدرک	
موضوع۳											
موضوع۴											
موضوع۵											
موضوع۶											
موضوع۷											
موضوع۸											
موضوع۹											
موضوع۱۰											

دو جدول دیگر نیز برای مقایسه جفتی موضوعات می توان ایجاد کرد که یکی از فرمول جاکاردی و دیگری از فرمول کوسینوسی بهره می برد:

	موضوع۱	موضوع۲	موضوع۳	موضوع۴	موضوع۵	موضوع۶	موضوع۷	موضوع۸	موضوع۹	موضوع۱۰
موضوع۱										
موضوع۲										
موضوع۳										
موضوع۴										
موضوع۵										
موضوع۶										

موضوع۷										
موضوع۸										
موضوع۹										
موضوع۱۰										

	موضوع۱	موضوع۲	موضوع۳	موضوع۴	موضوع۵	موضوع۶	موضوع۷	موضوع۸	موضوع۹	موضوع۱۰
موضوع۱										
موضوع۲										
موضوع۳										
موضوع۴										
موضوع۵										
موضوع۶										
موضوع۷										
موضوع۸										
موضوع۹										
موضوع۱۰										

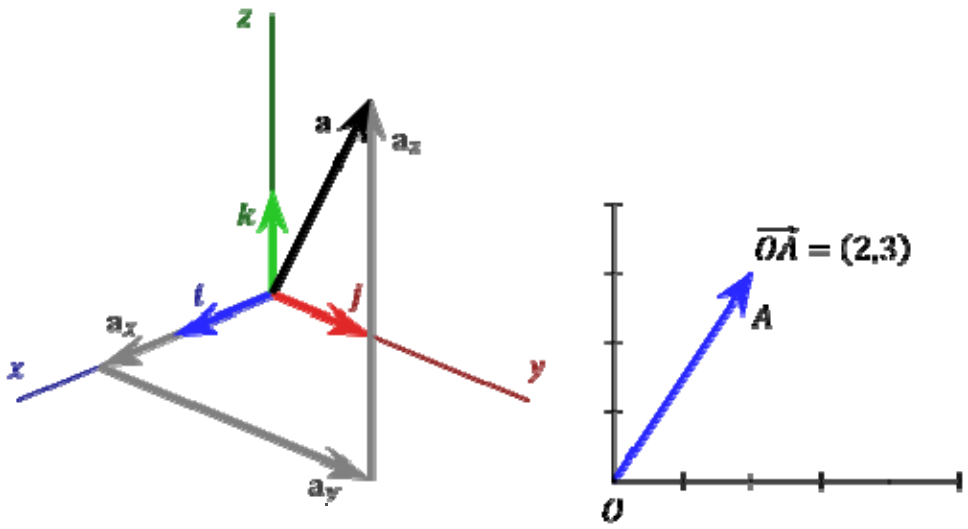
روش تعیین فراوانی هم رخدادی بین مدارک، نسبت به دو روش دیگر برای خوشه بندی موضوعات مناسب تر است.

تعیین فاصله یا شباهت

پس از ایجاد ماتریس به یکی از روش های بالا،نوبت به تعیین فاصله هندسی بین مدارک یا موضوعات می رسد تا بر آن اساس خوشه های مورد نظر را ایجاد کرد. به طور کلی دو نوع فاصله هندسی وجود دارد: کوسینوسی و اقلیدسی.

مقایسه برداری: عملا، هدف از مقایسه دو به دوی مشخصه ها، به دست آوردن اعداد داخل خانه های ماتریس هم جواری (متقارن) است. مجموع مقایسه های یک شیئی با اشیاء دیگر تشکیل یک بردار هندسی می‌دهد. از نگاه ماتریسی، اعداد هر ردیف از ماتریس هم جواری نقاط یک بردار را تشکیل می‌دهند.

بردار، خطی است که در فضا از دو نقطه یا بیشتر عبور می‌کند. اگر این خط با عبور از دو نقطه در فضا تشکیل شود، بردار حاصله دو بعدی و اگر از بیش از دو نقطه در فضا عبور کند، بردار چند بعدی نامیده می‌شود.



تصویر: بردار دو بعدی **تصویر: بردار سه بعدی**

درک و ترسیم بردارهای بیش از دو بعد کمی دشوار است. در سمت راست برداری دو بعدی ترسیم شده که از نقطه صفر در تقاطع محورهای x و y شروع شده و در محل تقاطع ۲ از محور xها و ۳ از محورy ها به پایان رسیده است. در برابر، در تصویر سمت چپ، برداری از سه بعد x، y و z نمایش داده شده است و خطی که

از تقاطع نقاط x، y و z حاصل شود، برداری حاصل از سه نقطه در فضا است. بردارهای حاصله از مقادیر درون خانه های ماتریس، اساس کار سنجش شباهت بین اشیاء و نهایتا خوشه بندی آن ها است که در فصل های .....و ..... به طور مفصل به آن خواهیم پرداخت.

روش‌های خوشه‌بندی

مقدمه

خوشه بندی[1] فرآیند گروه بندی[2] یا دسته بندی مجموعه ای از اشیاء[3]- داده ها یا الگوها که معمولاً به صورت برداری ارائه می شود- در دسته ها(کلاس‌ها یا خوشه‌ها) است به طوری‌که اشیای موجود در یک خوشه به هم شباهت[4] دارند در حالی‌که اشیای موجود در خوشه های متفاوت غیر شبیه‌اند[5].

نکته بسیار مهمی که در انتخاب روش خوشه بندی باید مد نظر قرار گیرد این است که هیچ تکنیک یکتای خوشه بندی وجود ندارد که برای انواع متفاوت ساختار داده‌ای عمومیت داشته باشد. به علاوه مجموعه متنوعی از تکنیک‌ها برای نمایش داده، اندازه‌گیری شباهت و دسته بندی وجود دارند که منجر به تولید انواع متنوعی از روش‌هایخوشه‌بندی
می‌شوند.

در عمل، هر تکنیک خوشه‌بندی در مورد شکل خوشه‌ها، معیار شباهت و روش دسته بندی به کار گرفته شده فرض های وابسته به خود را دارد. لازم به ذکر است که هر تکنیک خوشه‌بندی فواید و مضرات خود را دارد که وابسته به اندازه داده و توزیع داده‌ها است. در نتیجه، این مجری تکنیک یا کاربر است که معیارهای متفاوت خوشه‌بندی را تعیین می‌کند و تکنیک مناسب خوشه‌بندی را به گونه‌ای انتخاب می‌کند که نیازهای وی را به بهترین وجه ممکن برآورده کند. به علاوه هر چه مجری دانش بیشتری در مورد تکنیک خوشه‌بندی، داده‌های تولید شده و دامنه کابرد داشته باشد امکان موفقیت وی بیشتر می شود.[2]

بنابراین خوش‌بندی یک فرآیند ذهنی(وابسته به طرز تفکر فرد) است به این معنا که یک مجموعه داده در کاربردهای متفاوت به گونه‌های متفاوت افراز می‌شود.

۲-۴ تعاریف و نمادها

[1] -clustering
[2] -classification
[3] - objects
[4] -similar
[5] -dissimilar

در ذیل نمادها و اصطلاحاتی که از آنها در این بخش استفاده می‌شود آمده است:

- $\mathbf{x}$(یا هر حرف انگلیسی کوچک) که نماینده یک الگو، بردار ویژگی، نقطه داده‌ای، شی داده‌ای یا مشاهده است.
- $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_d})$ برداری با $\mathbf{d}$ اندازه‌گیری است. که هر اندازه‌گیری یک مشخصه یا ویژگی نامیده می‌شود. مقدار $\mathbf{i}$ امین ویژگی یا مشخصه با $\mathbf{x_i}$ نمایش داده می‌شود.
- $\mathbf{d}$ ابعاد فضای ویژگی(مشخصه) است.
- $\mathbf{X} = \{\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_N}\}$ مجموعه داده یا مشخصه ها را نشان می‌دهد. $\mathbf{x_i} = (\mathbf{x_{i1}}, \mathbf{x_{i2}}, \ldots, \mathbf{x_{id}})$ $\mathbf{i}$امین نقطه داده یا بردار ویژگی را نشان می‌دهد و $\mathbf{x_{ij}}$ مقدار $\mathbf{j}$ امین مشخصه از $\mathbf{x_i}$ را مشخص می‌کند.
- $\mathbf{N}$ تعداد نقاط داده در مجموعه داده را نشان می‌دهد.
- خوشه با نماد $\mathbf{C}$ مشخص شده است.
- $\mathbf{C} = \{\mathbf{C_1}, \mathbf{C_2}, \ldots, \mathbf{C_N}\}$ مجموعه ای از خوشه‌ها را نمایش می‌دهد.
- $\mathbf{K}$ تعداد خوشه‌ها در مجموعه داده است.
- $\mathbf{x'}$ ترانهاده بردار $\mathbf{x}$ است.(بردار سطری ۱در d برای x به یک بردار ستونی d در ۱ تبدیل شده است.)

۳-۴

۴-۳-۱ رویکردهای اساسی در خوشه بندی

کاهش بعد داده‌ها[1]: در بسیاری از موارد حجم داده های موجود بسیار زیاد است و پردازش آنها بسیار وقت گیر می باشد. خوشه بندی، داده‌ها را به چندین بخش افراز می‌کند. در این مرحله خوشه‌بندی، به جای پردازش کل مجموعه داده، مجموعه داده اولیه به صورت فشرده در آورده می‌شود.

تولید فرض[2]: گاهی برای استنتاج فرض‌هایی حول داده اولیه از تحلیل خوشه‌ای استفاده می شود. برای مثال در یک پایگاه داده دو دسته اصلی مشتری بر اساس سن و زمان خرید آنها وجود دارند. فرض های زیر ممکن است از این پایگاه داده استنتاج گردد: افراد جوان عصرها خرید می‌کنند و افراد پیر در صبح ها خرید می‌کنند.

[1] - data reduction
[2] - hypothesis generation

آزمایش فرض[1]: در این حالت تحلیل خوشه‌ای برای بررسی صحت فرض‌های از قبل تعیین شده به کار می رود. برای مثال فرض "افراد جوان صبح ها خرید می‌کنند" را در نظر بگیرید. یک راه برای تعیین درستی این فرض به کارگیری تحلیل خوشه‌ای برای مجموعه داده ذخیره شده است. فرض بر این است که جزئیات داده برای هر مشتری به صورت سن، شغل، میزان درآمد، زمان خرید و ... باشد اگر بعد از تحلیل خوشه‌ای این فرض که افراد جوان در عصر خرید می‌کنند به دست آید، تحلیل خوشه ای درست جواب می‌دهد.

پیش بینی[2]: در این حالت تحلیل خوشه‌ای انجام می‌شود و خوشه‌های نتیجه بر اساس مشخصه ای خاص تعیین می‌شوند. در مرحله بعد داده های جدید و ناشناخته می توانند بر اساس شباهت در خوشه‌ای خاص قرار گیرند. برای مثال، تحلیل خوشه‌ای می‌تواند بر روی مجموعه داده‌ای از بیماران از یک نوع بیماری عفونی به کار رود. نتیجه تعدادی خوشه از بیماران بر اساس واکنش آنها به دارویی خاص است. بنابراین برای یک بیمار جدید می‌توان خوشه ای که وی در آن قرار می‌گیرد را تعیین نمود و بر اساس آن داروی موردنظر برای بیمار را تجویز کرد.

۱- ۲-۳-۴

۴-۴ اجزاء و وظایف خوشه‌بندی

گام‌های زیر در خوشه‌بندی به صورت زیر می‌آیند:[1,4]

۱. **نمایش مشخصه**[3]: هدف نهایی از این مرحله سازماندهی مشاهدات و ساخت مجموعه ای از بردارهای مشخصه است که به عنوان ورودی به الگوریتم داده می‌شود. این گام شامل چندین وظیفه کوچکتر است.

۲. **تعیین نوع ویژگی یا مشخصه**[4]: انواع ویژگی‌ها یا کیفی یا کمی‌اند. داده‌های کیفی مانند ارتفاع که اعداد به آنها نسبت داده می‌شود و قابل پردازش و انجام عملیات هستند. داده‌های کیفی مانند رنگ که مقداری که دریافت می‌کنند رشته‌ای است و عملیات ریاضی را مانند جمع بر روی آنها نمی‌توان اجرا کرد. این دو نوع داده می توانند به زیر نوع‌های زیر تقسیم گردند. زیرنوع‌های کمی عبارتند از: ۱- پیوسته (مانند ارتفاع) ۲- گسسته (مانند تعداد مشتریان) ۳- داخلی (مدت زمان یک رخداد). زیرنوع-

[1] -hypothesis testing
[2] - prediction
[3] -feature representation
[4] - attribute types

های کیفی از قبیل: ۱- داده‌های اسمی(رنگ) ۲- ترتیبی (رتبه در مسابقات ورزشی). بعد از تعیین نوع مشخصه‌ها، مقیاس مشخصه‌ها و تعداد هریک باید تعیین شود.

۳. **انتخاب ویژگی**[1]: منظور، انتخاب مشخصه‌های مرتبط و تاثیرگذار از کل مجموعه داده اولیه است که با خوشه بندی ارتباط نزدیکی دارد.

۴. **استخراج ویژگی**[2]: هدف در اینجا محاسبه و پردازش مجموعه جدیدی از مشخصه ها است که از مجموعه داده اولیه با استفاده از تبدیلاتی مانند تحلیل اجزای اصلی[3][5] حاصل می‌شود و به صورت خودکار کیفیت خوشه‌بندی را بالا می برد.

نقش کاربر یا مجری در این مرحله بسیار حیاتی است. او بایستی مشاهدات و حقایق را در مورد داده برای ایجاد بهترین مجموعه ممکن از مشخصه‌ها در راستای خوشه‌بندی جمع آوری نماید. این مرحله نقش بسیار مهمی بر روی نتایج خوشه‌بندی ایفا می‌کند.

۵. خوشه‌بندی

این مرحله اشاره به انتخاب روش اندازه‌گیری شباهت، معیار خوشه‌بندی و الگوریتم‌های خوشه‌بندی برای مجموعه دادگان در دسترس دارد. در عمل روش اندازه‌گیری شباهت و معیار خوشه‌بندی نوع الگوریتم خوشه بندی را توصیف می کند.

- فاصله یا شباهت معیاری است که به صورت جفت جفت بر روی داده ها اعمال می شود و هدف از این معیار این است که نشان دهد دو شیء یا بردارهای مشخصه آنها چقدر به هم شباهت دارند.
- معیار خوشه بندی به صورت یک تابع هزینه بیان می‌شود برای مثال مجموع مربعات فاصله. معیار خوشه بندی براساس شکل و نوع خوشه هایی که در مجموعه داده مورد انتظار است انتخاب می شود.
- خوشه بندی به روش‌های گوناگونی اعمال می‌شود. همانطور که قبلاً ذکر شد شباهت و معیار خوشه بندی الگوریتمی که برای افراز داده به کار می رود را مشخص می نماید. با این وجود تصمیمات دیگری نیز باید اتخاذ شود. یکی از این تصمیمات تعیین عضویت هر شیء است. در بعضی از مواقع خوشه ها به سختی[4] انتخاب می‌شوند به این معنا که افرازهای داده به صورتی است که هر داده تنها در یک خوشه

[1]-feature selection
[2] -feature extraction
[3] -PCA(principal component analysis)
[4] -hard

قرار می گیرد. در بعضی دیگر از حالات خوشه ها راحت[1] تعیین می شوند به این معنا که در این افرازها یک داده ممکن است به چندین دسته تعلق داشته باشد. همچنین در افراز فازی هر داده با یک درجه عضویت به هر خوشه تعلق دارد.

۶. اعتبار(صحت) نتایج

بعد از اعمال الگوریتم‌های خوشه بندی بایستی اعتبار و صحت نتایج بررسی گردد. صحت خوشه بندی بررسی نتیجه الگوریتم خوشه بندی است. روش‌های صحت خوشه‌بندی معمولاً بر اساس معیاری بهینه، یا روش-های آماری برای جواب به این سوال که آیا نتیجه خوشه بندی معنادار است یا نه، انجام می‌شود. در عمل سه نوع روش تعیین صحت وجود دارد:[6]

- آزمون بیرونی: به مقایسه ساختار خوشه بندی و ساختار قبلی داده می‌پردازد.
- آزمون دورنی: به بررسی ساختار درونی خوشه ها می‌پردازد.
- آزمون نسبی: به مقایسه دو ساختار می پردازد و شایستگی نسبی هر یک را بررسی می‌کند.

۷. تفسیر نتیجه

در بسیاری از مواقع کارشناسان به ادغام نتیجه خوشه بندی با شواهد تجربی دیگر می‌پردازند که این نوع تجزیه و تحلیل منجر به نتیجه‌گیری بهتر می‌گردد.

۴-۵ طرح خوشه بندی

در این فصل به بررسی روش‌های خوشه‌بندی بر اساس تکامل طبیعی مسائل خوشه‌بندی و کاربردهای آن می‌پردازیم و سپس در هریک از این روش‌ها زیر نوع ها تعریف می‌شود و خصوصیات هریک، نقاط ضعف و قدرت شرح داده می‌شود.

روش‌های کلاسیک و نوین خوشه بندی، عملاً با مجموعه دادگان کم(با فضای داده کمتر) سروکار دارد و تمام مشخصه‌های داده‌ای که در رابطه با آن خوشه است را دربردارد، به این معنا که هر خوشه شامل زیر مجموعه ای از نقاط داده است که مجموعه کاملی از مشخصه ها در آن وجود دارد. به همین دلیل به این گونه دسته بندی، خوشه بندی با ابعاد کامل داده اطلاق می شود.

به علت پیشرفت‌های فناوری در زمینه گردآوری داده، امروزه خوشه بندی با داده هایی با ابعاد بالا معنا پیدا نموده است. برای دسته بندی این نوع داده ها با ابعاد زیاد روش‌های خوشه‌بندی با ابعاد کامل[1] کارایی

[1] -soft

ندارند، اولین دلیل، وجود مجموعه مشخصه های زیاد در رابطه با داده است، در صورتی که تمام مشخصه ها در هر خوشه مناسب نمی‌باشد. همچنین وجود مشخصه های نامرتبط باعث از بین بردن تمایل خوشه‌بندی و گمراه کردن الگوریتم های خوشه بندی با ابعاد کامل می‌شود که از طریق پوشش داده های نویز دار و بی ربط به وجود می‌آید. دومین چالش در رابطه با داده‌های با ابعاد بالا این است که نقاط با ابعاد بالا تمایل به داشتن مسافت مساوی با یکدیگر دارند از این‌رو معیار فاصله در رابطه با مجموعه کامل از ابعاد بی معنی قلمداد می‌شود.

۴-۶ دسته بندی اشیاء

تا کنون آموختیم که پس از تشکیل ماتریس متقارن اولیه، باید هر شیئ را یک بردار در نظر بگیریم و فاصله بین آن ها را با روش هایی مانند روش اقلیدسی و کوسینوسی محاسبه کنیم. پس از این کار می توان بر اساس فاصله های حاصل، اشیاء را دسته بندی کرد.:[5,4,2] همانند سنجش فاصله بین بردارها، روش دسته بندی اشیاء نیز متفاوت است. هر روش برای هدفی خاص مناسب است و مزایا و معایب مربوط به خود را دارد. از دیدگاه سنتی این روش ها را می توان به دو دسته سلسله مراتبی و افرازی تقسیم کرد. علاوه بر آن، در سال‌های اخیر روش‌های نوین دیگری ظهور یافته اند. از این رو، در این کتاب روش های موجود دسته بندی را به دو گروه اصلی سنتی و نوین تقسیم می کنیم که هر کدام نیز می توانند دارای زیر مجموعه هایی باشند :

- سنتی
- سلسله مراتبی
  - تراکمی
    - روش تک اتصال
    - روش اتصال کامل
    - روش اتصال میانگین
    - روش مرکزی
    - روش میانی
    - روش وارد[۲]
  - تقسیمی

[1] - full dimensional
[2] -Ward

- غیر سلسله مراتبی
  - میانگین K(K-Means)
- نوین
  - مبتنی بر چگالی
  - فازی
  - شبکه ها
  - گراف‌ها

روش های سنتی:

ویژگی مشترک روش های سنتی این است که هر شیئ تنها در یک خوشه قرار می گیرد. در کل می توان این روش ها رادر دو گروه سلسله مراتبی و غیر سلسله مراتبی قرار داد.

۴-۶-۱ روش سلسله مراتبی

در این روش، مانند یک درخت، هر شاخه کوچکتر جزئی از یک شاخه بزرگتر است و نهایتا، همه این ها به صورت سلسله مراتبی به تنه آن درخت وصل می شوند. نتیجه خوشه بندی به روش سلسله مراتبی را می توان به همین شکل در نظر گرفت که اشیاء به شکل یک نمودار درختی به صورت بازگشتی در خوشه های کوچک و کوچکتر قرار می گیرند که اصطلاحا به آن دندوگرام می گویند.

مزیت روش خوشه بندی سلسله مراتبی این است که از طریق آن می توان رابطه سلسله مراتبی بین اشیاء را کشف کرد و هم چنین راحت تر می توان میزان شباهت بین اشیاء را به صورت تصویری دید. به عبارتی، در یک نمودار درختی، هر چه عمق دهانه دو شیئ کمتر باشد، به راحتی می توان شدت شباهت آن دو را درک کرد. در تصویر پائین، مثلا مورد ۳ و ۳۰ نسبت به مورد ۲۴ و ۲۷ شباهت بیشتری دارند، زیرا عمق دهانه خوشه اولی از خوشه نامبرده دومی کمتر است. علاوه بر آن، مثلا می توان درک کرد که خوشه حاوی مورد ۲۴ و ۲۷ با خوشه مورد ۳ و ۳۰ چگونه با هم رابطه پیدا می کنند.

شکل- نمودار دندوگرام، در این نمودار محور افقی نشان دهنده نقاط داده و محور عمودی نمایان کننده شباهت بین نقاط داده می باشد

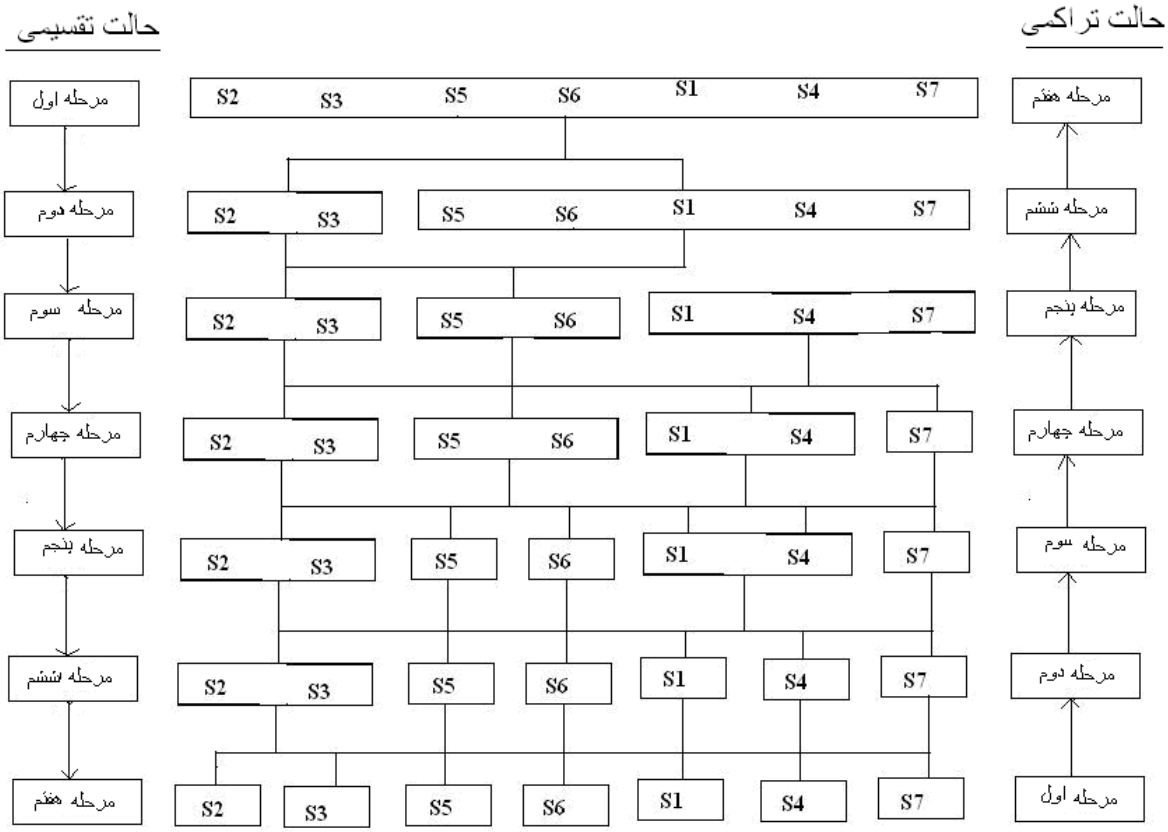
مزیت دیگر روش خوشه بندی به صورت سلسله مراتبی این است که از قبل نباید تعداد خوشه ها را تعیین کرد، در حالی که در روش های دیگر، باید از قبل دانست که تعداد خوشه ها چه اندازه باید باشد. اما تصمیم درباره این که ریز کردن تعداد خوشه ها تا چه میزان و مرحله ای صورت گیرد، بحث انگیز است. مشکل دیگر روش سلسله مراتبی هنگامی ظاهر می شود که تعداد اشیاء مورد نظر در ماتریس بالا باشد. زیرا در این حالت به یک پردازنده بسیار قوی نیاز است. وقتی تعداد اشیاء بالا باشد، پس از طی مراحل خوشه بندی و ایجاد نمودار درختی، عملا تفسیر و تحلیل خوشه ها نیز دشوار می شود. از این رو، این روش هنگامی موثر است که تعداد اشیاء به اندازه ای باشد که در کار پردازش اختلالی ایجاد نشود و امکان تفسیر و تحلیل خوشه های حاصله وجود داشته باشد.

در روش سلسله مراتبی، ایجاد خوشه ها به دو صورت اصلی صورت می گیرد:

۱-رویکرد تراکمی (پائین به بالا)

۲- رویکرد تقسیمی (بالا به پائین)

در هر دو رویکرد، کار دسته بندی اشیاء در خوشه های مناسب با توجه به ماتریس حاصل از مرحله قبل صورت می گیرد. در زیر به هر یک از این دو رویکرد و روش های مربوط به آن می پردازیم:



شکل- دو رویکرد الگوریتم‌های سلسله مراتبی

رویکرد تراکمی[1]

در این رویکرد، ابتدا هر شی یا داده یک خوشه محسوب می شود و به تدریج این خوشه های ریزتر با هم ادغام می‌شوند تا این که همه اشیا در یک خوشه قرار بگیرند. گاه این کار تا این مرحله صورت نمی گیرد و تنها تا وقتی ادامه می یابد که به تعداد خوشه های مورد نظر برسد. خوشه بندی با این رویکرد، به چند حالت ممکن است که هر کدام دارای مزایا و معایبی است:

[1] -agglomerative

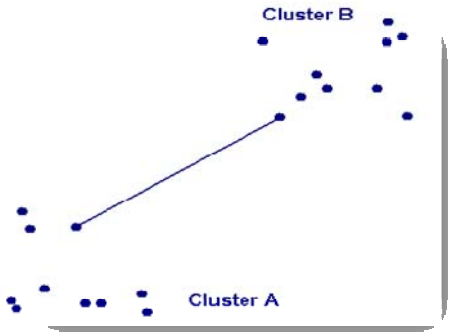
- روش تک اتصال
- روش اتصال کامل
- روش اتصال میانگین
- روش مرکزی
- روش میانی
- روش وارد[1]

اکثر الگوریتم های سلسله مراتبی مبتنی بر متریک های تک اتصال و اتصال کامل هستند. الگوریتم های مبتنی بر متریک اتصال کامل تمایل به تولید خوشه های محدود و متراکم و جمع و جور دارند در حالیکه متریک تک اتصال گرایش به تولید خوشه های دورافتاده و کشیده(دراز) دارد. بنابراین یکی از خصوصیات الگوریتم های سلسله مراتبی توانایی آنها در مقابله با خوشه هایی با شکل‌های هندسی متفاوت است.

۱- روش تک اتصال[2](روش نزدیکترین همسایه)

بر اساس این روش، شاخص میزان شباهت، کوچکتر بودن عددی است که از سنجش فاصله بین یک شیئ با اشیاء دیگر حاصل می شود. مثلا اگر صد شیئ داشته باشیم، صد فاصله حاصل می شود. در نهایت، کوچکترین عدد حاصل نشان می دهد کدام دو مورد از میان این صد مورد نسبت به هم بیشترین نزدیکی را دارند و همین طور به ترتیب هر چه میزان فاصله بین دو شیئ بیشتر شود، شباهت آن دو کمتر خواهد شد.

[1] -Ward
[2] -single link



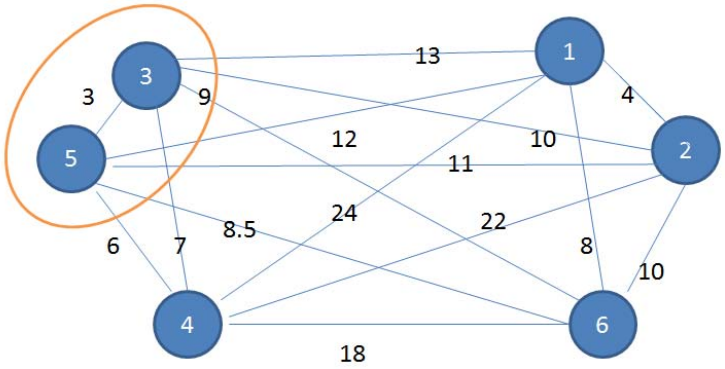
شکل- روش تک اتصال

همان گونه که گفتیم، بر اساس الگوریتم‌های سلسله مراتبی تراکمی، ابتدا هر شیئ به تنهایی یک خوشه محسوب می شود، سپس در هر مرحله بر اساس متریک تعریف شده به هم اتصال می‌یابند تا کل داده در یک خوشه قرار گیرد و کار خوشه بندی به انتها می‌رسد. برای مثال[9]، اجرای این کار با روش نزدیک ترین همسایگی از طریق ماتریس زیر به صورت زیر خواهد بود:

در ماتریس زیر، صرف نظر از اعداد خانه های قطر جدول که صفر است، کوچکترین مقدار ۳ است که از سنجش فاصله بین شیئ ۳ و ۵ حاصل شده است. بنا براین، در مرحله اول، از میان همه این اشیاء، مورد ۳ و ۵ به هم شبیه تر هستند. بنا بر این، در مرحله بعد، این دو تشکیل یک خوشه را می دهند.

	۱	۲	۳	۴	۵	۶
۱	۰	۴	۱۳	۲۴	۱۲	۸
۲		۰	۱۰	۲۲	۱۱	۱۰
۳			۰	۷	۳	۹
۴				۰	۶	۱۸
۵					۰	۸,۵
۶						۰

شکل- ماتریس برای ۶ نمونه داده



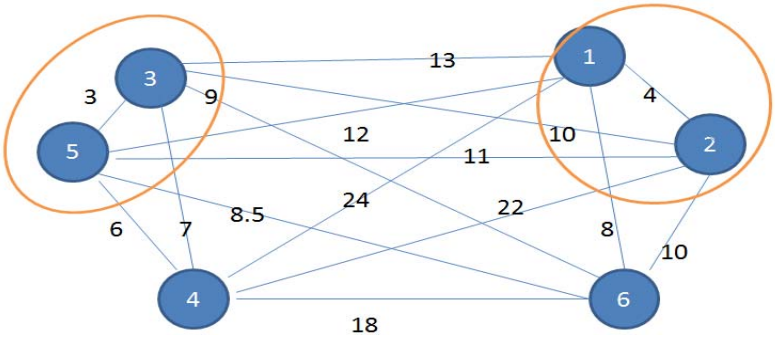
شکل- خوشه بندی سلسله مراتبی روش تک اتصال- داده‌های ۳ و ۵ دارای کمترین فاصله(بیشترین شباهت) هستند در یک دسته قرار می‌گیرند.

وقتی که اشیاء ۳ و ۵ در یک دسته قرار گرفتند، برای سنجش فاصله این دسته دو عضوی با سایر اشیاء نیز کوچکترین فاصله دخیل است. بنا براین، باید فاصله این دسته را با تمامی اشیاء دیگر بسنجیم تا هر کدام کمترین فاصله را با این دسته کسب کرد وارد این دسته شود و تشکیل یک خوشه سه عضوی بدهد. این خوشه سه عضوی به همین نحو باید در مرحله بعدی عضو چهارم خود را پیدا کند و این عمل آنقدر ادامه یابد که همه اشیاء در یک دسته قرار گیرند. برای سنجیدن فاصله هر دسته با اشیاء دیگر، باید مانند مثال زیر عمل کرد:

مثلا، قصد داریم که فاصله دسته {۳ و ۵} را از نقطه ۱ پیدا کنیم. در این جا، دو فاصله وجود دارد که فاصله شیئ ۱ با شیئ ۳ برابر با ۱۳ و فاصله شیئ ۱ با شیئ ۵ برابر با ۱۲ است. از این میان فاصله این دسته دو عضوی با شیئ ۱ برابر با ۱۲ است، چون ۱۲ از ۱۳ کوچکتر است. فاصله دسته {۳ و ۵} با اشیاء دیگر یک به یک سنجیده می شود و نهایتا آن شیئ که نسبت به بقیه فاصله کمتری با این دسته دارد، وارد آن دسته می شود. پس از آن، دسته سه عضوی حاصل، به همین نحو، عضو چهارم خود را پیدا می کند.

	۱	۲	(۳ و ۵)	۴	۶
۱	۰	۴	۱۲	۲۴	۸
۲		۰	۱۰	۲۲	۱۰
(۳ و ۵)			۰	۶	۸,۵
۴				۰	۱۸
۶					۰

شکل- ماتریس به روز شده بعد از دسته بندی ۳ و۵ در یک خوشه



شکل-گراف به روز شده بعد از دسته بندی ۲ و۱ در یک خوشه

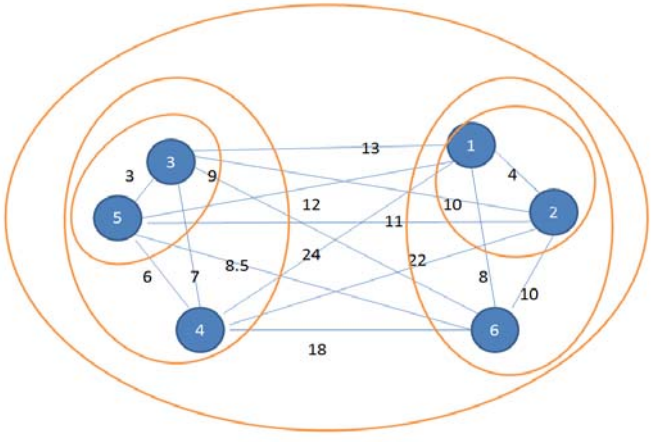
	(۱ و ۲)	(۳ و ۵)	۴	۶
(۱ و ۲)	۰	۱۰	۲۲	۸
(۳ و ۵)		۰	۶	۸,۵
۴			۰	۱۸
۶				۰

شکل- ماتریس به روز شده بعد از دسته بندی ۲ و۱ در یک خوشه

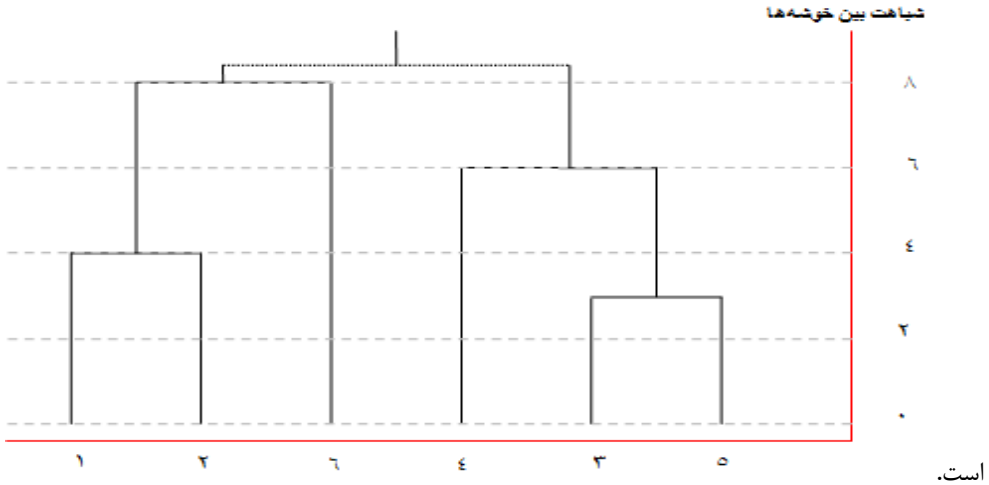
این مراحل تکرار می‌شود تا در نهایت تمام داده‌ها در دو دسته قرار گیرند و به یک ماتریس دو در دو برسیم.

	(۱، ۲ و ۶)	(۳، ۴ و ۵)
(۱، ۲ و ۶)	۰	۸.۵
(۳، ۴ و ۵)		۰

شکل-ماتریس دو در دو که ارتباط بین دو خوشه با داده های{۱و۲و۶} و {۳و۴و۵} را نشان می‌دهد.



شکل- خوشه‌بندی سلسله مراتبی تراکمی به روش تک اتصال- نهایتا یک خوشه به وجود می‌آید و یک ماتریس یک در یک با یک خوشه ایجاد می‌شود که شامل تمام داده‌ها



است.

شکل-نمودار دندوگرام روش تک اتصال برای ۶ نمونه داده

- **Advantages:** Theoretical properties, efficient implementations, widely used. No cluster centroid or representative required, so no need arises to recalculate the similarity matrix.
- **Disadvantages:** Unsuitable for isolating spherical or poorly separated clusters

۲-روش اتصال کامل[1](روش دورترین همسایه)

بر خلاف روش نزدیک ترین همسایه، مبنای دسته بندی در این جا، بزرگترین فاصله است. اگر بخواهیم بر اساس ماتریس ؟؟؟؟، خوشه بندی سلسله مراتبی را با به کار گیری روش دورترین همسایگی انجام دهیم، مانند روش نزدیک ترین همسایگی، باید ابتدا کمترین فاصله یا همان کوچکترین عدد ماتریس را مورد توجه قرار بدهیم که اشیاء ۳ و ۵ انتخاب می‌شوند و در یک دسته قرار می‌گیرند. سپس این ماتریس باید به روز شود. در اینجا بیشترین فاصله بر اساس متریک تعریف شده انتخاب می‌شود. برای مثال فاصله دسته {۳و۵} از شیئ در مقایسه با شیئ ۳ برابر با ۶ و در مقایسه با شیئ ۵ برابر با ۷ است. در این جا میزان فاصله بین شیئ ۴ با دسته {۳و۵} برابر با ۷ یعنی بزرگترین عدد از میان این دو عدد است.

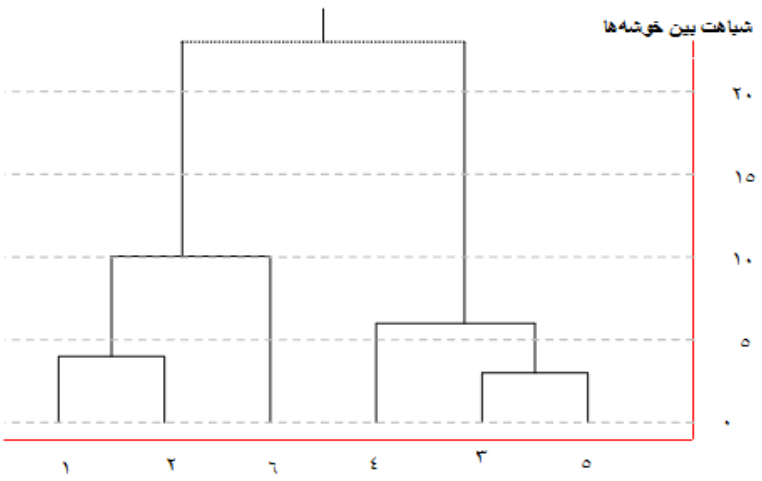
بعد از آن، کمترین فاصله بین اشیاء ۱ و ۲ وجود دارد که برابر با ۴ است، لذا این دو در یک دسته قرار می گیرند. مثلا، برای محاسبه فاصله دسته {۱ و ۲} با شیئ ۳، مجددا فاصله بین شیئ ۱ و ۳ و هم چنین

[1] - complete link

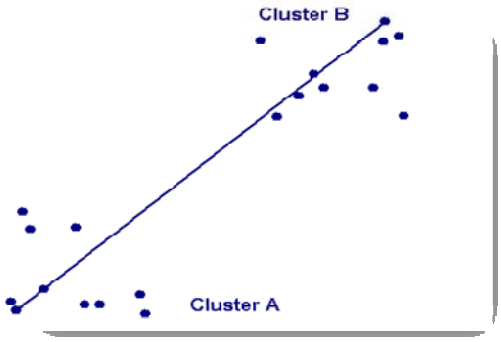
فاصله بین شیئ ۲ و ۳ سنجیده می شود. بزرگترین عددی که از این دو مقایسه حاصل شود، فاصله بین دسته {۱ و ۲} با شیئ ۳ محسوب خواهد شد و این کار تا پایان ادامه می یابد.

	۱	۲	(۳ و ۵)	۴	۶
۱	۰	۴	۱۳	۲۴	۸
۲		۰	۱۱	۲۲	۱۰
(۳ و ۵)			۰	۷	۹
۴				۰	۱۸
۶					۰

شکل- ماتریس حاصل از دسته بندی نقاط ۳و ۵ در یک خوشه وبه روز رسانی ماتریس



شکل- نمودار دندوگرام روش اتصال کامل برای ۶ نمونه داده



شکل-روش اتصال کامل

- **Advantages:** Good results from (Voorhees) comparative studies.
- **Disadvantages:** Difficult to apply to large data sets since most efficient algorithm is general HACM using stored data or stored matrix approach.

۳-روش اتصال میانگین[1]

همان گونه که از نام این روش بر می آید، ملاک خوشه بندی در آن میانگین فاصله ها است. برای به دست آوردن میانگین فاصله، ابتدا باید فاصله هر یک از اعضای خوشه مورد نظر را با اعضای خوشه دیگر جمع کرد. مثلا، اگر ۳ شیئ در یک خوشه و ۲ شیئ در خوشه دیگر باشد، در این جا شش فاصله داریم که با یکدیگر جمع می شوند. بنا بر این، برای محاسبه میانگین باید حاصل جمع آن شش عدد بر عدد ۶ تقسیم شود. به عبارتی، مخرج کسر، حاصل ضرب تعداد اعضای خوشه اول در خوشه دوم است. این نکته را می توان با فرمول زیر نمایش داد:

$$d_{AB} = \frac{\sum_{i \in A, j \in B} d_{ij}}{N_A N_B}$$

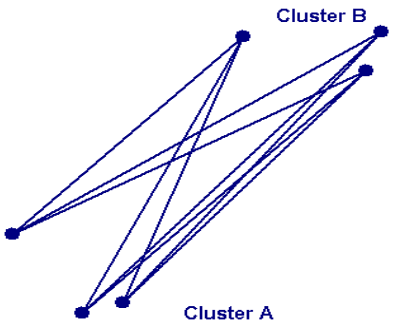
در بالا، A و B دو خوشه متفاوت در نظر گرفته شده است. متغیر i نشانگر هر یک از اعضای خوشه A و متغیر j نشانگر هر یک از اعضای خوشهٔ B است. بنابر این، d_{ij} فاصله شیئ i از خوشه A تا شیئ j در خوشه B است. هم چنین، N_A تعداد اعضای خوشهٔ A و N_B تعداد اعضاء خوشهٔ B محسوب می شود.

[1] -average link

بر اساس روش سلسله مراتبی تراکمی، ابتدا باید هر یک از اشیاء را یک خوشه در نظر بگیریم. حال اگر مجددا ماتریس ؟؟؟؟؟ را نمونه قرار دهیم، طبق فرمول بالا، فاصله بین شیئ ۳ و ۵ برابر با عدد ۳ است و چون تعداد هر کدام یک است بنا بر این مخرج کسر نیز ۱ می شود. از این رو، کوچکترین عدد حاصل، همانند روش های دیگر برابر با ۳ است. بنابر این، کمترین فاصله و به عبارتی، بیشترین شباهت بین شیئ ۳ و ۵ وجود دارد و در مرحله اول، این دو شیئ تشکیل یک خوشه بزرگتر می دهند. حال اگر بخواهیم فاصله دسته {۳ و ۵} را طبق این روش با شیئ ۶ محاسبه کنیم، فاصله شیئ ۳ و ۵ برابر با ۸/۵ و شیئ ۳ و ۶ برابر با ۹ است. از این رو، طبق فرمول بالا، نتیجه به صورت زیر خواهد بود:

$$d\{3,5\}6 = \frac{8.5+9}{2\times 1} = 8.75$$

طبق حاصل بالا، فاصله بین دسته {۳ و ۵ } با شیئ ۶ برابر با ۸/۷۵ است. حال فاصله این دسته با اشیاء دیگر به همین نحو سنجیده می شود. بالاخره، آن شیئ که کوچکترین فاصله را کسب کند، وارد دسته {۳ و ۵} می شود و تشکیل یک خوشه سه عضوی می دهد. در مرحله بعد، فاصله این خوشه سه عضوی به همین نحو با اشیاء باقیمانده سنجیده می شود، تا یک دسته با چهار شیئ تشکیل شود. با توجه به فرمول بالا، این عمل آن قدر صورت می گیرد تا در مرحله آخر، یک ماتریس دو در دو حاصل شود.



شکل- روش اتصال میانگین

- **Advantages:** Ranked well in evaluation studies
- **Disadvantages:** Expensive for large collections

۶-خوشه‌بندی با روش Ward

این روش در سال ۱۹۶۳ توسط Ward معرفی شده است و به همین جهت با نام وی شناخته می شود. در این روش، ابتدا میانگین فاصله اشیاء یک خوشه محاسبه می شود و سپس همانند روش محاسبه واریانس،

تفاضل فاصله هر شیئ با آن میانگین سنجیده می شود. به عبارتی، روش وارد بر مبنای مجموع مربعات تفاضل هر داده از یک خوشه با بردار میانگین آن خوشه استوار است. این مفهوم را می توان به صورت فرمول زیر نمایش داد:

$$\mathrm{ESS} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

در فرمول بالا، ESS[1] برابر با مربع خطای استاندارد است. x_i نشانگر یک شیئ و n برابر با تعداد اشیاء در یک خوشه است. بنا بر این، $\sum_{i=1}^{n} x_i^2$ برابر با مربع مجموع مقادیر فاصله ها و $\frac{1}{n}(\sum_{i=1}^{n} x_i)^2$ برابر با مربع میانگین فاصله ها است. بر این اساس، اگر مجددا ماتریس ؟؟؟؟ را در نظر بگیریم، در مرحله اول، با فرمول بالا، فاصله هر شیئ با اشیاء دیگر سنجیده می شود. دو شیئ که بر اساس فرمول بالا کمترین مقدار ESS را کسب کنند، شبیه ترین اشیاء محسوب می شوند. در مراحل بعدی، مجددا مقدار ESS آن دو شیئی با سایر اشیاء محاسبه می شود. ترکیب این دو، با هر شیئ که کمترین مقدار ESS را کسب کند، تشکیل یک خوشه سه عضوی می دهد. این مرحله تا زمانی که همه اشیاء در یک خوشه قرار بگیرند ادامه می یابد.

خلاصه این که بر اساس روش Ward باید مراحل زیر را طی کرد:

۱. هر شیئ به عنوان یک خوشه در نظر گرفته شود.
۲. به ازاء تمام جفت خوشه‌های ممکن، آن دو خوشه‌ای انتخاب شوند که ESS کمتری دارند.
۳. دو خوشه ای که انتخاب شده اند با هم ترکیب شوند.
۴. تا زمانی که همه اشیاء در یک خوشه قرار نگرفته اند، یا تعداد خوشه‌ها به تعداد مورد نظر نرسیده است، مراحل ۲ و ۳ تکرار می‌شوند[10].

باید توجه داشت که روش وارد منجر به ایجاد خوشه های کم تعداد می شود و این شاید به تفسیر داده ها کمک بیشتری بکند.

- **Advantages:** Good at recovering cluster structure, yields unique and exact hierarchy
- **Disdvantages:** Sensitive to outliers, poor at recovering elongated clusters

[1] Error Sum Squares

رویکرد تقسیمی[1]

در این روش، تمام داده ها یا اشیاء ابتدا در یک دسته قرار می گیرند و به تدریج آن دسته به دسته های ریزتر تقسیم می شود. در هر تکرار، خوشه های بزرگ تر براساس معیارهایی به خوشه های کوچکتر شکسته می‌شوند تا زمانی که هر خوشه تنها حاوی یک داده یا شی باشد و یا شرط ایجاد خوشه های کوچکتر به پایان برسد.

۴-۶-۲ روش غیر سلسله مراتبی (افرازی)

در روش افرازی، بر خلاف روش سلسله مراتبی، از قبل باید تعداد خوشه ها را مشخص و سپس اقدام به خوشه بندی کرد. تفاوت اساسی دیگر روش سلسله مراتبی با افرازی در این است که حاصل آن به جای ایجاد نمودار درختی (دندوگرام) ، خوشه های مجزا از یکدیگر است. به عبارتی، حاصل این روش، خوشه هایی است که بین اعضای آن ارتباطی وجود ندارند. برای ایجاد این گونه خوشه هایی، باید برخی از اشیاء را به عنوان نماینده برگزید و شباهت اشیاء دیگر را با آن ها سنجید. بر این اساس، هر شیئ در خوشه مربوط به خود جای می گیرد. بنا بر این، اشیاء نماینده، عملاً مراکز خوشه ها هستند.

غالباً روش‌های افرازی روند دو مرحله ای را دنبال می‌کنند که به تدریج کیفیت خوشه بندی را افزایش می-دهد:

۱. در اولین گام، اشیاء نماینده از راه‌های متفاوت انتخاب می‌شوند. یکی از این راه‌ها، تعیین تعداد مشخصی از اشیاء به روش تصادفی و یکی دیگر از راه‌ها، انتخاب به روش جستجو است. در روش انتخابی، اشیاء نماینده آن هایی هستند که در مرکز سایر اشیاء قرار می گیرند. به عبارتی، اشیائی به عنوان نماینده انتخاب می شوند که مقدار مجموع مربعات فاصله آنها از دیگر اشیاء، از همه کمتر باشد.

۲. در گام بعد، اشیاء بعدی به خوشه‌هایی تعلق می گیرند که کمترین فاصله را با نماینده آن دارند.

همان گونه که می بینم، خوشه بندی افرازی، مستلزم دانش قبلی از مسئله است که غالباً موجود نیست. یکی دیگر از ضعف‌های این روش‌ها عدم تاوانایی آنها در ساخت خوشه هایی با شکل دلخواه است. این روش‌ها معمولاً قادر به ساخت خوشه هایی هستند که به خوبی از هم جدا شده اند و شکل فشرده و محدب دارند.

[1] -decisive

روش‌های افرازی دارای مزیت هایی هستند، من جمله کاربرد آن ها برای هنگامی که تعداد اشیاء مورد مطالعه بسیار بالا است. مزیت دیگر این الگوریتم‌ها، سادگی نسبی آنها است. برخلاف روش‌های سلسله مراتبی که خوشه ها بعد از اجرای الگوریتم مورد بازبینی قرار نمی‌گیرند، روش‌های افرازی به تدریج توسعه می یابند و منجر به خوشه هایی با کیفیت بالا می شوند.

۴-۶-۲-۱ روش kmeans

روش kmeans [7]ساده ترین و رایج ترین الگوریتم در خانواده الگویتم های افرازی به شمار می رود. حاصل این روش، تابعی از مربعات خطا E(X) است که به صورت زیر تعریف می شود:

$$E(X) = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|^2$$

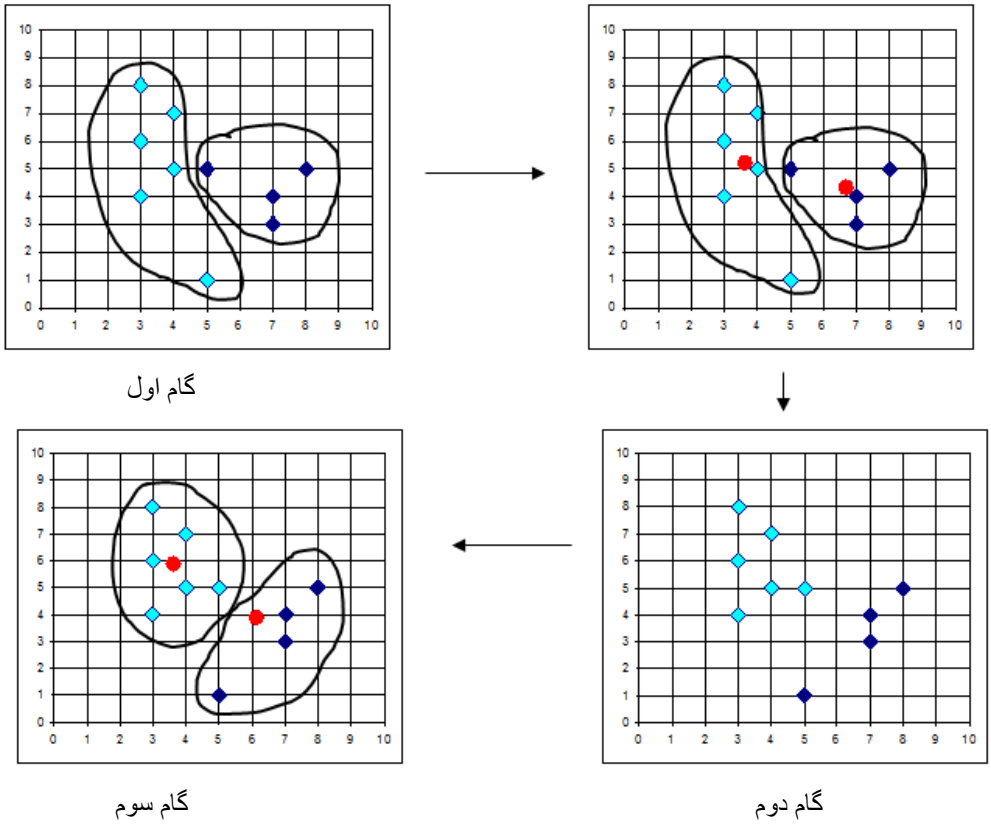
در فرمول تابعی بالا، k نشان دهنده تعداد خوشه های مورد نظر، x نشانگر شیئ، c نشانگر خوشه و mi برابر با میانگین نقاط موجود در خوشه مورد نظر است. دو علامت سیگمای تودرتو به معنای این است که مثلاً زمانی که مقدار سیگمای اول یک ($i = 1$) است برای تمام اشیائی که به خوشه C1 تعلق دارند ($x \in c_1$) مربع تفاضل هر شیئ از میانگین آن خوشه($\|x - m_1\|^2$) محاسبه شود. حال وقتی که $i = 2$ است، همین کار در خوشه C2 صورت می گیرد. این کار تا زمانی تکرار می شود که مقدار i برابر با k (تعداد خوشه ها) شود.

این الگوریتم با نقاط نماینده تصادفی یا انتخابی، کار را شروع می‌کند و در هر مرحله هر شیئ در خوشه ای قرار می گیرد که با نماینده آن کمترین فاصله را دارد. فراموش نکنیم که هر نماینده در اینجا مشخص کننده یک خوشه است. پس از آن که هر شیئ در کنار نماینده خود قرار گرفت، از میان اشیاء حاضر در خوشه، مجددا یک نماینده انتخاب می شود و جای نماینده قبلی را می‌گیرد. این عملیات زمانی متوقف می شود که با وجود تکرار عمل، دیگر تغییری در هم نشینی اشیاء موجود در خوشه حاصل نشود، یعنی دیگر اشیاء از یک خوشه به خوشه بعدی تغییر مکان ندهند. هم چنین عامل توقف می تواند نزدیک شدن به معیار همگرایی باشد (به عنوان مثال خطای مجذور مربعات کاهش یابد و یا متوقف گردد).

مهمترین مسئله در رابطه با این الگوریتم، حساسیت آن به نقاط دورافتاده[1] است که می‌تواند تاثیر بسیار زیادی روی مرکز خوشه ها داشته باشد، زیرا باعث بالا رفتن مقدار پراکندگی (مانند بالا رفتن واریانس) می شود. این روش، تا حد قابل توجهی به افراز در اولین گام و انتخاب نمایندگان اولیه وابسته است. اشکال دیگر این این

[1] -outlier

است که به دلیل انتخاب میانگین هر دسته به عنوان مرکز یا نماینده آن دسته ، این الگوریتم تنها برای اشیاء از نوع داده عددی کاربرد دارد.



شکل- الگوریتم kmeans

خلاصه این که، در روش Kmeans ابتدا تعداد دسته ها به عنوان ورودی به الگوریتم داده می‌شود در مرحله بعدی به تعداد خوشه‌ها یا دسته‌ها نمایندگانی به صورت تصادفی از مجموعه داده‌ها انتخاب می‌شوند. سپس دیگر داده‌ها در صورتی که به نماینده اول نزدیکتر باشند در دسته اول قرار می‌گیرند و در صورتی که به نماینده دوم نزدیکتر باشند در دسته دوم قرار می‌گیرند. در مرحله بعد، مراکز جدید هر دسته محاسبه می‌شود که نقاط با رنگ قرمز می‌باشند و مانند گام اول فاصله نقاط از این دو نماینده به دست می‌آیند، به هر یک نزدیکتر باشند به آن دسته تعلق دارند در نتیجه ممکن است نقاط در بین خوشه‌ها جابجا گردند. این عملیات

همان‌طور که در شکل دیده می‌شود تا مرحله سوم ادامه یافته است که دیگر جابجایی برای اشیاء در دسته‌های متفاوت دیده نمی‌شود.

روش K-medoids

روش K-medoids هیچ محدودیتی بر روی نوع داده ها و مشخصه های مربوط به داده ها ندارد. بر خلاف روش Kmeans روش K-medoids از مرکزی ترین نقطه به عنوان مرکز هر خوشه استفاده می‌کند بنابراین این روش نسبت به روش قبل به نقاط دور افتاده حساسیت کمتری دارد. علاوه بر این، این روش بر خلاف روش Kmeans به نوع داده های خاصی محدود نیست. البته این ویژگی باعث پیچیدگی بیشتر این روش شده است.

۴-۶-۳ خوشه‌بندی فازی[1]

روش‌هایی که تا کنون بیان شدند نوعی خوشه بندی گسسته بودند به این معنا که هر شیئ تنها به یک خوشه تعلق دارد. در بسیاری از موارد مثلاً در زندگی معمولی، خوشه بندی بایستی عدم قطعیت را نشان دهد. برای مثال در خوشه بندی بیان ژن، هر ژن ممکن است به چندین خوشه مرتبط باشد بنابراین، قطعیت گاهی بایستی به عدم قطعیت در خوشه بندی تبدیل شود. بر خلاف خوشه بندی با قطعیت- هر داده تنها به یک خوشه تعلق می‌گیرد- خوشه بندی فازی به داده ها اجازه می دهد تا بر اساس یک درجه عضویت به چندین خوشه تعلق داشته باشد. این درجه ها یا توابع عضویت در دامنه ۰ تا ۱ قرار دارد یعنی هر داده با احتمال بین ۰ تا ۱ به هر یک از دسته ها مربوط است به طوری که مقادیر بزرگتر نشان دهنده احتمال بیشتر برای انتساب شیء به یک خوشه می باشد . برای مثال یک داده با احتمال ۰/۲ به دسته اول و به احتمال ۰/۴ به دسته دیگر تعلق دارد در اینجا احتمال اینکه داده به خوشه دوم متعلق باشد دو برابر احتمال عضویت داده در خوشه اول است.

مهمترین الگوریتم فازی الگوریتم Fuzzy C-Means (FCM) است [8] که تعمیمی از روش سنتی Kmeans است. این روش از یک ماتریس-آرایه دوبعدی- عضویت N در K به نام U استفاده می‌کند که هر عنصر در این ماتریس با نماد u_{ij} نشان داده می شود و نماینده درجه عضویت شیء x_i به خوشه c_j است و در دامنه$[0,1]$ قرار دارد. مانند روش Kmeans هدف کمینه سازی تابع مربعات خطایی است که به صورت وزن-دار می‌آید. تابع خطا در روش فازی به صورت زیر محاسبه می شود:

[1] -fuzzy clustering

$$E(X) = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{ij}^{m} \|x_i - c_j\|^2$$

در این معادله m توان وزنی[1] است که عددی بزرگتر از یک می باشد و c_j مرکز جرم خوشه C_j است. این روش همانند روش Kmeans با نقاط اولیه تصادفی افراز یا دسته بندی را آغاز می‌کند در این حالت مقداردهی تصادفی برای ماتریس U انجام می شود و خوشه بندی داده ها به صورت یک فرآیند بهینه سازی و تکراری شبیه به روش Kmeans با بروزرسانی مقدار c_j انجام می‌شود و تا زمانی‌که به مقدار مورد نظر مسئله نزدیک گردد ادامه می‌یابد.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{2/(m-1)}}$$

در این حالت نیز اگرچه روش فازی از روش Kmeans بهتر عمل می‌کند زیرا دچار کمینه محلی[2] نمی‌شود با این وجود هنوز گرفتار همان مشکل اصلی است.

۴-۶-۴ خوشه بندی مبتنی بر چگالی[3]

یکی از مهمترین خصوصیات خوشه بندی، چگالی یا دانسیته است. چگالی یک ویژگی است که تراکم اشیاء موجود در فضا را نشان می‌دهد. یکی از روش‌های نوین برای خوشه بندی روش مبتنی بر چگالی است[10]. نکته اصلی در رابطه با این الگوریتم‌ها این است که خوشه‌ها بر اساس توابع چگالی ایجاد می‌شوند. مزیت اصلی این الگوریتم‌ها ایجاد خوشه‌هایی با شکل دلخواه می‌باشد[4].

این روشهای خوشه‌بندی بر این اصل استوارند که خوشه‌ها، ناحیه‌هایی از فضای داده با چگالی زیاد هستند که توسط نواحی با چگالی کمتر از همدیگر جدا شده‌اند. این تعریف مبنای خوشه بندی است. روش‌های مختلفی برای تحلیل خوشه ای بر اساس خوشه‌بندی میتنی بر چگالی وجود دارند[10]:

۴-۶-۵ خوشه بندی مبتنی بر شبکه

[1] - weighting exponent
[2] -local minimum
[3] -density based clustering

عموماً این نوع از الگوریتم‌ها برای داده‌کاوی فضایی پیشنهاد می‌شوند. آنها فضای جستجو را برای مجموعه داده‌ها به صورت تدریجی محدود می‌نمایند[4].

۴-۶-۶ روش های پاک سازی داده ها

۴-۶-۷ روش تعیین تعداد خوشه ها

۴-۶-۸ روش های ارزیابی نتایج خوشه بندی

خوشه‌بندی یک فرآیند بدون نظارت در داده کاوی و تشخیص الگو[1] به شمار می‌رود و بیشتر الگوریتم‌های خوشه‌بندی به پارامترهای ورودی‌شان حساس هستند از این رو ارزیابی نتایج این گونه الگوریتم‌ها بسیار اهمیت دارد. در خوشه‌بندی هیچ دسته از پیش تعیین شده ای وجود ندارد و داده‌ها در ابتدا بدون برچسب هستند بنابراین یافتن متریک یا استاندارد مناسبی که بتوان توسط آن درجه مقبولیت خوشه ها را تعیین کرد بسیار مشکل است. در این راستا روش‌ها و شاخص‌هایی برای برررسی اعتبار روش‌های خوشه‌بندی توسعه یافته‌اند[12,13] .

فرآیند ارزیابی نتایج الگوریتم‌های خوشه‌بندی، ارزیابی اعتبار خوشه‌بندی[2] نامیده می‌شوند. دو معیار اندازه گیری برای ارزیابی و انتخاب طرح خوشه بهینه پیشنهاد شده است[11] :

۱- تراکم[3]

۲- جدایی[4]

ارزیابی بر اساس تراکم

اعضای یک خوشه تا حد ممکن بایستی به هم نزدیک باشند. یکی از معیارهای رایج برای اندازه‌گیری تراکم داده‌ها واریانس می باشد.

ارزیابی بر اساس جدایی

خوشه‌ها خود بایستی به خوبی از یکدیگر جدا باشند. سه رویکرد رایج برای اندازه‌گیری فاصله مابین دو خوشه به این شرح وجود دارد

[1] -pattern recognition
[2] - cluster validity assessment
[3]- compactness
[4] - separation or isolation

- فاصله مابین نزدیکترین اعضای دو خوشه از هم
- فاصله مابین دورترین اعضای دو خوشه از هم
- فاصله مابین مراکز خوشه ها

همچنین روش‌های ارزیابی نتایج حاصل از خوشه بندی به سه دسته تقسیم می‌شوند:

- معیار های خارجی
- معیار های داخلی
- معیار های نسبی

دو معیار خارجی و داخلی بر اساس روش‌های آماری استوار هستند که به محاسبات زیاد نیاز دارند. معیارهای اعتبارسنج خارجی خوشه‌بندی را بر اساس شهود خاص کاربران انجام می‌دهند. معیارهای داخلی مبتنی بر متریک‌ها یا استانداردهایی در رابطه با مجموعه داده و طرح خوشه‌بندی اند. اشکال هر دو روش پیچیدگی محاسباتی آنها است.

اساس معیارهای نسبی مقایسه بین شماها یا طرح های مختلف خوشه‌بندی است. یک یا چند الگوریتم مختلف چندین بار با پارامترهای متفاوت بر روی یک مجموعه داده اجرا می‌گردند و بهترین شما یا طرح خوشه‌بندی از بین تمام آنها انتخاب می‌شود.

منابع

[1] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT press, 1996.

[2] Anil K. Jain and Richard C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., 1988.

[3] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern Recognition, 2nd edition. cademic Press, 2003.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. ACM Computing Surveys, 31(3):264–323, 1999.

[5] Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern Classification,Second Edition. Wiley, 2000.

[6] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3):107–145, 2001.

[7] J. B. Macqueen. Some methods for classification and analysis of multivariate observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. Berkeley, University of California Press, 1967.

[8] Bezdek J.C., Ehrlich R., and Full W. FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences, 10(2-3):191–203, 1984

[9] A. R. Web, Statistical Pattern Recognition, John Wiley & Sons, 2002.

[10] Q. He, A Review of Clustering Algorithms as Applied in IR, Graduate School of Library and Information Science University of Illinois at Urbana-Champaign, 1999.

[11] M. J. A. Berry and G. Linoff: Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley & Sons, Inc., 1996

[12] M. Halkidi, Y. Batistakis and M. Vazirgiannis: Cluster validity methods: part I, SIGMOD Rec., Vol. 31, No. 2, pp. 40-45, 2002.

[13] M. Halkidi, Y. Batistakis and M. Vazirgiannis: Cluster validity methods: part II, SIGMOD Rec., Vol. 31, No. 3, pp. 19-27, 2002.

پیوست۱: مشخصه های مدرک بر اساس استاندارد MARC

			Control Fields		001 - Control Number 003 - Control Number Identifier 005 - Date and Time of Latest Transaction 006 - Fixed-Length Data Elements - Additional Material Characteristics 007 - Physical Description Fixed Field 008 - Fixed-Length Data Elements
			Numbers and Code Fields		010 - Library of Congress Control Number (NR) 013 - Patent Control Information (R) 015 - National Bibliography Number (R) 016 - National Bibliographic Agency Control Number (R) 017 - Copyright or Legal Deposit Number (R) 018 - Copyright Article-Fee Code (NR) 020 - International Standard Book Number (R) 022 - International Standard Serial Number (R) 024 - Other Standard Identifier (R) 025 - Overseas Acquisition Number (R) 026 - Fingerprint Identifier (R) 027 - Standard Technical Report Number (R) 028 - Publisher Number (R) 030 - CODEN Designation (R) 031 - Musical Incipits Information (R) 032 - Postal Registration Number (R) 033 - Date/Time and Place of an Event (R) 034 - Coded Cartographic Mathematical Data (R) 035 - System Control Number (R) 036 - Original Study Number for Computer Data Files (NR)

					037 - Source of Acquisition (R) 038 - Record Content Licensor (NR) 040 - Cataloging Source (NR) 041 - Language Code (R) 042 - Authentication Code (NR) 043 - Geographic Area Code (NR) 044 - Country of Publishing/Producing Entity Code (NR) 045 - Time Period of Content (NR) 046 - Special Coded Dates (R) 047 - Form of Musical Composition Code (NR) 048 - Number of Musical Instruments or Voices Codes (R) 050 - Library of Congress Call Number (R) 051 - Library of Congress Copy, Issue, Offprint Statement (R) 052 - Geographic Classification (R) 055 - Classification Numbers Assigned in Canada (R) 060 - National Library of Medicine Call Number (R) 061 - National Library of Medicine Copy Statement (R) 066 - Character Sets Present (NR) 070 - National Agricultural Library Call Number (R) 071 - National Agricultural Library Copy Statement (R) 072 - Subject Category Code (R) 074 - GPO Item Number (R) 080 - Universal Decimal Classification Number (R) 082 - Dewey Decimal Classification Number (R) 083 - Additional Dewey Decimal Classification Number (R) 084 - Other Classification Number (R) 085 - Synthesized Classification Number Components (R) 086 - Government Document Classification Number (R) 088 - Report Number (R) 09X - Local Call Numbers
			Heading Fields - General		X00 - Personal Names - General Information

			Information		100 - Main Entry - Personal Name (NR) 600 - Subject Added Entry - Personal Name (R) 700 - Added Entry - Personal Name (R)) 800 - Series Added Entry - Personal Name (R) X10 - Corporate Names - General Information 110 - Main Entry - Corporate Name (NR) 610 - Subject Added Entry - Corporate Name (R) 710 - Added Entry - Corporate Name (R) 810 - Series Added Entry - Corporate Name (R) X11 - Meeting Names - General Information 111 - Main Entry - Meeting Name (NR) 611 - Subject Added Entry - Meeting (R) 711 - Added Entry - Meeting Name (R) 811 - Series Added Entry - Meeting Name (R) X30 - Uniform Titles - General Information 130 - Main Entry - Uniform Title (NR) 630 - Subject Added Entry - Uniform Title (R) 730 - Added Entry - Uniform Title (R) 830 - Series Added Entry - Uniform Title (R)
			Main Entry Fields		100 - Main Entry - Personal Name (NR) 110 - Main Entry - Corporate Name (NR) 111 - Main Entry - Meeting Name (NR) 130 - Main Entry - Uniform Title (NR)
			Title and Title-Related Fields		210 - Abbreviated Title (R) 222 - Key Title (R) 240 - Uniform Title (NR) 242 - Translation of Title by Cataloging Agency (R) 243 - Collective Uniform Title (NR) 245 - Title Statement (NR) 246 - Varying Form of Title (R) 247 - Former Title (R)

			Edition, Imprint, Etc. Fields		250 - Edition Statement (NR) 254 - Musical Presentation Statement (NR) 255 - Cartographic Mathematical Data (R) 256 - Computer File Characteristics (NR) 257 - Country of Producing Entity (R) 258 - Philatelic Issue Data (R) 260 - Publication, Distribution, etc. (Imprint) (R) 263 - Projected Publication Date (NR) 270 - Address (R)
			Physical Description, Etc. Fields		300 - Physical Description (R) 306 - Playing Time (NR) 307 - Hours, etc. (R) 310 - Current Publication Frequency (NR) 321 - Former Publication Frequency (R) 336 - Content Type (R) 337 - Media Type (R) 338 - Carrier Type (R) 340 - Physical Medium (R) 342 - Geospatial Reference Data (R) 343 - Planar Coordinate Data (R) 351 - Organization and Arrangement of Materials (R) 352 - Digital Graphic Representation (R) 355 - Security Classification Control (R 357 - Originator Dissemination Control (NR) 362 - Dates of Publication and/or Sequential Designation (R) 363 - Normalized Date and Sequential Designation (R) 365 - Trade Price (R) 366 - Trade Availability Information (R) 380 - Form of Work (R) 381 - Other Distinguishing Characteristics of Work or Expression (R)

					382 - Medium of Performance (R 383 - Numeric Designation of Musical Work (R) 384 - Key (NR)
			Series Statement Fields		490 - Series Statement **(R)**
			Note Fields		500 - General Note (R) 501 - With Note (R) 502 - Dissertation Note (R) 504 - Bibliography, etc. Note (R) 505 - Formatted Contents Note (R) 506 - Restrictions on Access Note (R) 507 - Scale Note for Graphic Material (NR) 508 - Creation/Production Credits Note (R) 510 - Citation/References Note (R) 511 - Participant or Performer Note (R) 513 - Type of Report and Period Covered Note (R) 514 - Data Quality Note (NR) 515 - Numbering Peculiarities Note (R) 516 - Type of Computer File or Data Note (R) 518 - Date/Time and Place of an Event Note (R) 520 - Summary, etc. (R) 521 - Target Audience Note (R) 522 - Geographic Coverage Note (R) 524 - Preferred Citation of Described Materials Note (R) 525 - Supplement Note (R) 526 - Study Program Information Note (R) 530 - Additional Physical Form available Note (R) 533 - Reproduction Note (R) 534 - Original Version Note (R) 535 - Location of Originals/Duplicates Note (R) 536 - Funding Information Note (R) 538 - System Details Note (R)

					540 - Terms Governing Use and Reproduction Note (R) 541 - Immediate Source of Acquisition Note (R) 542 - Information Relating to Copyright Status (R) 544 - Location of Other Archival Materials Note (R) 545 - Biographical or Historical Data (R) 546 - Language Note (R) 547 - Former Title Complexity Note (R) 550 - Issuing Body Note (R) 552 - Entity and Attribute Information Note (R) 555 - Cumulative Index/Finding Aids Note (R) 556 - Information About Documentation Note (R) 561 - Ownership and Custodial History (R) 562 - Copy and Version Identification Note (R) 563 - Binding Information (R) 565 - Case File Characteristics Note (R) 567 - Methodology Note (R) 580 - Linking Entry Complexity Note (R) 581 - Publications About Described Materials Note (R) 583 - Action Note (R) 584 - Accumulation and Frequency of Use Note (R) 585 - Exhibitions Note (R) 586 - Awards Note (R) 588 - Source of Description Note (R) 59X - Local Notes
			Subject Access Fields		600 - Subject Added Entry - Personal Name (R) 610 - Subject Added Entry - Corporate Name (R) 611 - Subject Added Entry - Meeting Name (R) 630 - Subject Added Entry - Uniform Title (R) 648 - Subject Added Entry - Chronological Term (R) 650 - Subject Added Entry - Topical Term (R) 651 - Subject Added Entry - Geographic Name (R)

					653 - Index Term - Uncontrolled (R) 654 - Subject Added Entry - Faceted Topical Terms (R) 655 - Index Term - Genre/Form (R) 656 - Index Term - Occupation (R) 657 - Index Term - Function (R) 658 - Index Term - Curriculum Objective (R) 662 - Subject Added Entry - Hierarchical Place Name (R) 69X - Local Subject Access Fields (R)
			Added Entry Fields		700 - Added Entry - Personal Name (R) 710 - Added Entry - Corporate Name (R) 711 - Added Entry - Meeting Name (R) 720 - Added Entry - Uncontrolled Name (R) 730 - Added Entry - Uniform Title (R) 740 - Added Entry - Uncontrolled Related/Analytical Title (R) 751 - Added Entry - Geographic Name (R) 752 - Added Entry - Hierarchical Place Name (R) 753 - System Details Access to Computer Files (R) 754 - Added Entry - Taxonomic Identification (R)
			Linking Entry Fields		760 - Main Series Entry (R) 762 - Subseries Entry (R) 765 - Original Language Entry (R) 767 - Translation Entry (R) 770 - Supplement/Special Issue Entry (R) 772 - Supplement Parent Entry (R) 773 - Host Item Entry (R) 774 - Constituent Unit Entry (R) 775 - Other Edition Entry (R) 776 - Additional Physical Form Entry (R) 777 - Issued With Entry (R) 780 - Preceding Entry (R)

					785 - Succeeding Entry (R) 786 - Data Source Entry (R) 787 - Other Relationship Entry (R)
			Series Added Entry Fields		800 - Series Added Entry - Personal Name (R) 810 - Series Added Entry - Corporate Name (R) 811 - Series Added Entry - Meeting Name (R) 830 - Series Added Entry - Uniform Title (R)
			Holdings, Location, Alternate Graphics, Etc. Fields		841 - Holdings Coded Data Values (NR) 842 - Textual Physical Form Designator (NR) 843 - Reproduction Note (R) 844 - Name of Unit (NR) 845 - Terms Governing Use and Reproduction (R) 850 - Holding Institution (R) 852 - Location (R) 853 - Captions and Pattern - Basic Bibliographic Unit (R) 854 - Captions and Pattern - Supplementary Material (R) 855 - Captions and Pattern - Indexes (R) 856 - Electronic Location and Access (R) 863 - Enumeration and Chronology - Basic Bibliographic Unit (R) 864 - Enumeration and Chronology - Supplementary Material (R) 865 - Enumeration and Chronology - Indexes (R) 866 - Textual Holdings - Basic Bibliographic Unit (R) 867 - Textual Holdings - Supplementary Material (R) 868 - Textual Holdings - Indexes (R) 876 - Item Information - Basic Bibliographic Unit (R) 877 - Item Information - Supplementary Material (R) 878 - Item Information - Indexes (R) 880 - Alternate Graphic Representation (R) 882 - Replacement Record Information (NR) 886 - Foreign MARC Information Field (R)

					887 - Non-MARC Information Field (R)

پیوست ۲: مشخصه های مدرک بر اساس استاندارد Dublin Core

عناصر متاداده‌ی ساده دوبلین کور			پالایشگرهای عناصر متاداده‌ی ساده دوابلین کور	
Element	***توضیح عنصر***	***طرح‌های نشانه‌گذاری***	***Refinement(s)***	***توضیح پالایشگر***
Title	نامی که به یک اثر از سوی پدیدآور یا ناشر داده شده است		Alternative	هر نوع عنوان دیگر که بتواند جانشین عنوان رسمی بشود، حتی عنوان به زبان دیگر
Creator	شخص/اشخاص یا سازمانی/ سازمان‌های مسئول در برابر محتوای اثر از لحاظ فکری. خالق اثر		–	–
Subject	موضوع اثر، همچنین کلیدواژه‌ها عبارات یا رده‌بندی که موضوع یا محتوای اثر را توصیف می کند	LCSH MeSH DDC LCC UDC	-	–
Description	توصیف محتوای اثر به صورت متن، شامل چکیده مربوط به یک اثر مکتوب یا توصیف محتوای یک اثر دیداری یا شنیداری		Table of Contents	فهرست بخش‌های فرعی تشکیل دهنده‌ی اثر
			Abstract	خلاصه‌ای از محتوای اثر
Publisher	مسئول در دسترس ساختن اثر به شکل موجود، مانند ناشر، یک گروه آموزشی در دانشگاه یا یک سازمان		-	
Contributor	شخص یا اشخاص، سازمان یا سازمان‌هایی که از لحاظ فکری در خلق اثر شرکت داشته‌اند، اما نقش درجه دوم دارند		-	
Date	تاریخی که یک اثر به شکل موجود در دسترس قرار گرفته است	DCMI Period W3C-DFT	Created	تاریخی که اثر خلق شده است. مثل تاریخ انتشار
			Valid	تاریخ اعتبار اثر(معمولا به یک محدوده زمانی اشاره دارد)
			Available	تاریخی که یک اثر در دسترس قرار گرفته است یا در دسترس قرار می گیرد(معمولا به یک محدوده زمانی اشاره دارد)
			Issued	تاریخ صدور اثر به صورت رسمی(مثلا تاریخ انتشار)

			Modified	تاریخی که در اثر تغییراتی ایجاد شده است
Type	نوع اثر، مانند home page، داستان، شعر، مقاله، گزارش فنی، واژه‌نامه و ...		-	
Format	شکل ارائه اثر، مانند text/HTML، ASCCI, PDF، کاست و ...		Extent	طول یا مدت یک اثر
		IMT	Medium	ماده یا کاریر مادی که اثر روی آن قرار دارد. مانند نوار کاست
Identifier	یک رشته یا عدد مورد استفاده برای متمایز ساختن اثر از سایر آثار. نمونه‌ای از شناسه‌گر ها در محیط شبکه آدرس اینترنتی اثر (URL) است.	URI	-	
Source	ماخذ و منشاء اثر که از آن طریق می‌توان به اصل اثر دست یافت.	URI	-	
Language	زبان یا زبان‌های یک اثر	ISO 639-2 RFC 1766	-	
Relation	ارتباط یک اثر با سایر آثار	URI	Is Version Of	بیان می‌کند که اثر حاضر نسخه، ویرایش یا برگرفته از چه اثر یا آثار دیگری است
			Has Version	بیان می‌کند که چه آثار دیگری نسخه، ویرایش یا برگرفته از اثر حاضر هستند
			Is Replaced By	بیان می کند که اثر حاضر دارای چه ضمائم، نسخه‌های مشابه یا در ارتباط با چه آثار قبلی است
			Replaces	بیان می‌کند که اثر حاضر ضمیمه، نسخه‌ی مشابه یا اثر مقدم بر چه اثر یا آثاری است
			Is Required By	بیان می‌کند که اثر حاضر مورد نیاز چه اثر یا آثار دیگری است
			Requires	بیان می‌کند که اثر حاضر به چه آثار دیگری نیاز دارد
			Is Part OF	بیان می‌کند که اثر حاضر بخشی فیزیکی یا منطقی از یک ماخذ دیگر است
			Has Part	بیان می کند که اثر حاضر از لحاظ فیزیکی یا منطقی برای کار به چه اثر یا آثار دیگری نیاز دارد
			Is Referenced	بیان می‌کند که اثر حاضر از سوی چه آثار دیگری

			By	ارجاع داده شده، یا استناد شده یا مورد اشاره قرار گرفته است
			References	بیان می‌کند که اثر حاضر به چه آثار دیگری ارجاع داده یا استناد کرده یا چه آثاری را مورد اشاره قرار داده است
			IS Format Of	بیان می‌کند که نسخه‌ی حاضر مشابه نسخه دیگری با یک شکل دیگر است
			Has Format	بیان می‌کند که اثر دیگری مشابه نسخه حاضر است
Coverage	پوشش زمانی و مکانی یک اثر	DCMI ISO 3166 DCMI Box TGN	Spatial	پوشش مکانی اثر
		DCM Period W3C-DFT	Temporal	پوشش زمانی اثر
Rights	یک پیوند مانند یک آدرس اینترنتی(در صورت وجود) که به مسائل مربوط به حق مولف و مواردی از این قبیل اشاره دارد		-	

پیوست ۳: سیاهه ای از واژه های ممنوعه در انگلیسی

able	co	happens	me	rather	thereupon	while
about	co.	hardly	mean	rd	there've	whilst
above	com	has	meantime	re	these	whither
abroad	come	hasn't	meanwhile	really	they	who
according	comes	have	merely	reasonably	they'd	who'd
accordingly	concerning	haven't	might	recent	they'll	whoever
across	consequently	having	mightn't	recently	they're	whole
actually	consider	he	mine	regarding	they've	who'll
adj	considering	he'd	minus	regardless	thing	whom
after	contain	he'll	miss	regards	things	whomever
afterwards	containing	hello	more	relatively	think	who's
again	contains	help	moreover	respectively	third	whose
against	corresponding	hence	most	right	thirty	why
ago	could	her	mostly	round	this	will
ahead	couldn't	here	mr	said	thorough	willing
ain't	course	hereafter	mrs	same	thoroughly	wish
all	c's	hereby	much	saw	those	with
allow	currently	herein	must	say	though	within
allows	dare	here's	mustn't	saying	three	without
almost	daren't	hereupon	my	says	through	wonder
alone	definitely	hers	myself	second	throughout	won't
along	described	herself	name	secondly	thru	would
alongside	despite	he's	namely	see	thus	wouldn't
already	did	hi	nd	seeing	till	yes
also	didn't	him	near	seem	to	yet
although	different	himself	nearly	seemed	together	you

always	directly	his	necessary	seeming	too	you'd
am	do	hither	need	seems	took	you'll
amid	does	hopefully	needn't	seen	toward	your
amidst	doesn't	how	needs	self	towards	you're
among	doing	howbeit	neither	selves	tried	yours
amongst	done	however	never	sensible	tries	yourself
an	don't	hundred	neverf	sent	truly	yourselves
and	down	i'd	neverless	serious	try	you've
another	downwards	ie	nevertheless	seriously	trying	zero
any	during	if	new	seven	t's	
anybody	each	ignored	next	several	twice	
anyhow	edu	i'll	nine	shall	two	
anyone	eg	i'm	ninety	shan't	un	
anything	eight	immediate	no	she	under	
anyway	eighty	in	nobody	she'd	underneath	
anyways	either	inasmuch	non	she'll	undoing	
anywhere	else	inc	none	she's	unfortunately	
apart	elsewhere	inc.	nonetheless	should	unless	
appear	end	indeed	noone	shouldn't	unlike	
appreciate	ending	indicate	no-one	since	unlikely	
appropriate	enough	indicated	nor	six	until	
are	entirely	indicates	normally	so	unto	
aren't	especially	inner	not	some	up	
around	et	inside	nothing	somebody	upon	
as	etc	insofar	notwithstanding	someday	upwards	
a's	even	instead	novel	somehow	us	
aside	ever	into	now	someone	use	
ask	evermore	inward	nowhere	something	used	
asking	every	is	obviously	sometime	useful	

associated	everybody	isn't	of	sometimes	uses
at	everyone	it	off	somewhat	using
available	everything	it'd	often	somewhere	usually
away	everywhere	it'll	oh	soon	v
awfully	ex	its	ok	sorry	value
back	exactly	it's	okay	specified	various
backward	example	itself	old	specify	versus
backwards	except	i've	on	specifying	very
be	fairly	just	once	still	via
became	far	k	one	sub	viz
because	farther	keep	ones	such	vs
become	few	keeps	one's	sup	want
becomes	fewer	kept	only	sure	wants
becoming	fifth	know	onto	take	was
been	first	known	opposite	taken	wasn't
before	five	knows	or	taking	way
beforehand	followed	last	other	tell	we
begin	following	lately	others	tends	we'd
behind	follows	later	otherwise	th	welcome
being	for	latter	ought	than	well
believe	forever	latterly	oughtn't	thank	we'll
below	former	least	our	thanks	went
beside	formerly	less	ours	thanx	were
besides	forth	lest	ourselves	that	we're
best	forward	let	out	that'll	weren't
better	found	let's	outside	thats	we've
between	four	like	over	that's	what
beyond	from	liked	overall	that've	whatever
both	further	likely	own	the	what'll

brief	furthermore	likewise	particular	their	what's
but	get	little	particularly	theirs	what've
by	gets	look	past	them	when
came	getting	looking	per	themselves	whence
can	given	looks	perhaps	then	whenever
cannot	gives	low	placed	thence	where
cant	go	lower	please	there	whereafter
can't	goes	ltd	plus	thereafter	whereas
caption	going	made	possible	thereby	whereby
cause	gone	mainly	presumably	there'd	wherein
causes	got	make	probably	therefore	where's
certain	gotten	makes	provided	therein	whereupon
certainly	greetings	many	provides	there'll	wherever
changes	had	may	que	there're	whether
clearly	hadn't	maybe	quite	theres	which
c'mon	half	mayn't	qv	there's	whichever

پیوست ۴: سیاهه ای از واژه های ممنوعه در فارسی

و	شده‌است	به‌عنوان	لذا	کم	درخصوص
در	هر	اول	آنچه	می‌توانند	ازلحاظ
به	هستند	درباره	می‌گردد	مشخص	به‌وسیله
که	دارند	بسیار	بوده‌است	هم	بیش‌از
از	می‌باشد	درمورد	بلکه	بدین	کل
این	بنابراین	باشد	روی	به‌ویژه	هیچ
را	باید	چه	بالا	پایین	برروی
است	براساس	شود	حتی	چگونه	خارج/ خارج‌از
با	آنان	اگر	شده	فقط	بعداز
برای	همچنین	کلی	زیرا	البته	ازآنجاکه

آن	بیشتر	می‌شوند	پس‌از	بالاتر	بوده
خود	از	همین	اینکه	چهار	مثلا
نیز	میان	چون	ولی	سوم	پس
آنها	نسبت‌به	جهت	بدون	چند	درواقع
بر	یعنی	زیر	مستقیم	شدند	درست
یا	ما	زیاد	بودند	آشکار	نبود
بین	می‌تواند	دیگری	همان	زمانی	بدین‌ترتیب/ به‌این‌ترتیب
یک	می‌توان	گردید	همه	علاوه‌بر	عالی
می‌شود	سه	اما	تمام	بعضی/ بعضی‌از	کامل
دو	نیست	بسیاری/ بسیاری از	نه	کاملا	
بود	به‌صورت	دوم	یکدیگر	همانطورکه	
	یک	کمتر		فوق	

تا			بهتر	
دارد	ازنظر	تنها	به‌ترتیب	آیا
دیگر	برخی/ برخی‌از	وی	شده‌اند	بطوریکه
شد	چنین	هریک/ هریک‌از	درنتیجه	می‌باشند

پیوست ۵: جدول کدهای اسکی

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com