خوشه‌بندی اطلاعات

چکیده

خوشه‌بندی، فرایند سازماندهی عناصر به گروه‌هایی است که اجزای آن به هم شبیه هستند. یك خوشه، مجموعهٔ عناصری است که با هم مشابهت دارند و با اجزای دیگر خوشه‌ها ناهمگون می‌باشند. هدف خوشه‌بندی، دستیابی سریع و مطمئن به اطلاعات همبسته، و شناسایی ارتباط منطقی میان آنهاست. بنابراین، الگوریتم‌های خوشه‌بندی می‌تواند دربسیاری از حوزه‌های موضوعی به‌کار گرفته شود. از آنجا که نتایج خوشه‌بندی می‌تواند با تعداد اصطلاح‌های مورد استفاده تغییر یابد، روش‌های تجربی متعددی برای تشخیص تعداد تقریبی اصطلاح‌هایی که می‌توان انتظار داشت تا توزیع متناسب داده‌ها را در میان خوشه‌ها فراهم سازد و حدود بالا و پایین الگوریتم خوشه‌بندی را تعیین کند، ارائه شده است. یکی از کاربردهای خوشه‌بندی، سازماندهی اصطلاحنامه‌هاست. در این مقاله، با بهره‌گیری از روش مطالعهٔ کتابخانه‌ای، ضمن بررسی مفهوم خوشه‌بندی اطلاعات، روش‌های مؤثر خوشه‌بندی اطلاعات و کاربرد آن در ساختار تزاروس (اصطلاحنامه) بیان شده است. از جمله روش‌های عمدهٔ خوشه‌بندی چهار شیوهٔ خوشه‌بندی دسته‌ای، تک‌پیوندی، ستاره‌ای، و رشته‌ای است. نتیجهٔ این مطالعه نشان می‌دهد که بهره‌گیری از الگوریتم‌های متناسب و روش‌های سودمند خوشه‌بندی اطلاعات می‌تواند نقشی مهم در سازماندهی بهینهٔ اصطلاحنامه داشته باشد.

کلیدواژه‌ها

خوشه‌بندی اطلاعات، تزاروس (اصطلاحنامه)، نظام‌های بازیابی اطلاعات، خوشه‌بندی مدارك

خوشـه‌بندی اطلاعـات

نرگس محمدعلیپور[1] | فریبرز درودی[2]

دریافت: ۱۳۸۶/۵/۳۰ پذیرش: ۱۳۸۶/۹/۱

مقدمه

خوشه‌بندی اطلاعات و منابع اطلاعاتی یکی از راهکارهای مؤثر در سازماندهی اطلاعات به‌شمار می‌آید. با انجام فرایندهای خوشه‌بندی اطلاعات، حیطهٔ گسترده‌ای از داده‌های پراکنده در گروه‌های مدون و سازمان‌یافته قرار می‌گیرند. گروه‌های متعدد ایجاد شده، با برخورداری از ویژگی‌های مشترک درون هر گروه، دارای ارتباط ارگانیک و ساختاری با یکدیگر هستند. یکی از کاربردهای مؤثر و سودمند خوشه‌بندی در ایجاد تزاروس یا اصطلاحنامه[3] می‌باشد. برای بهره‌گیری از اصطلاحات تخصصی نیاز به تدوین و سازماندهی روشمند واژگان است. در این رابطه انتخاب اصطلاح‌های اصلی، به‌همراه مترادف‌ها و هم‌معناهای آن، یکی از کاربردهای خاص خوشه‌بندی می‌باشد. در نظام‌های بازیابی اطلاعات نیز خوشه‌بندی نقشی مؤثر ایفا می‌کند. برای دسترسی به منابع اطلاعاتی، شناسایی ابعاد و جنبه‌های خاص موضوعی مفاهیم، که با اصطلاح‌های تخصصی مشخص می‌شوند، ضروری است. با تدوین راهبردهای اثربخشی که در خوشه‌بندی اطلاعات وجود دارد، می‌توان براساس نیاز اطلاعاتی کاربر، به نتایج مناسب‌تری در بازیابی اطلاعات دست یافت. شیوه‌های مؤثر خوشه‌بندی، با تدوین رده‌های اصلی و زیررده‌های وابسته شکل می‌گیرد. انواع الگوریتم‌ها و روش‌های تخصصی در تدوین روابط میان خوشه‌ها

۱. عضو هیئت علمی دانشگاه آزاد اسلامی واحد رودهن
۲. دکتری کتابداری و اطلاع‌رسانی سازمان اسناد و کتابخانهٔ ملی ایران
f-doroudi@nlai.ir

3. Thesaurus

و تعامل ساختاری، به کارآمد ساختن فرایندهای خوشه‌بندی یاری می‌رساند. خوشه‌بندی اطلاعات، درواقع، تدوین رده‌های ویژه برای گروه‌بندی مدون و منظم اطلاعات است که به بازیابی بهینهٔ آن منجر می‌شود.

هدف پژوهش

هدف اصلی این پژوهش، تبیین دقیق و مناسب مفهوم خوشه‌بندی اطلاعات است و انواع شیوه‌ها و روش‌های مناسب خوشه‌بندی مورد بررسی قرار گرفته است. همچنین، مراحل انجام فرایند خوشه‌بندی اطلاعات تشریح شده، و روابط خوشه‌بندی بیان می‌شود. بررسی انواع الگوریتم‌های خوشه‌بندی مورد توجه قرار گرفته و کاربرد خوشه‌بندی در ساختار تزاروس و اصطلاحنامه مطرح شده و در این رابطه نکات مهم تشریح می‌شود.

روش پژوهش

روش انجام این پژوهش مطالعهٔ کتابخانه‌ای و جست‌وجو در منابع اطلاعاتی، به‌منظور به‌دست آوردن اطلاعات جدید در زمینهٔ خوشه‌بندی اطلاعات، فرایندها، روش‌ها، مراحل و انواع شیوه‌های مؤثر خوشه‌بندی است.

پیشینهٔ پژوهش

جین، مورتی و فلین[4] (۱۹۹۹)، در اثر تحقیقی خود باعنوان «مروری بر خوشه‌بندی داده‌ها»، که در سال ۱۹۹۹ به انجام رسیده، بیان می‌کنند که خوشه‌بندی، طبقه‌بندی غیرنظارتی الگوها، شامل مشاهدات، اقلام داده‌ای یا بُردارهای خصیصه‌ای، به گروه‌ها و خوشه‌های منظم است. آنان شرح می‌دهند که مشکل خوشه‌بندی در بسیاری از زمینه‌ها، توسط محققان علوم مختلف مورد توجه قرار گرفته است. این امر، نشانگر فایده و درخواست وسیع آن به‌عنوان یکی از مراحل تحلیل داده‌کاوی است. خوشه‌بندی، فرایند تخصصی دشواری است و اختلاف در فرضیه‌ها و زمینه‌های موضوعی باعث شده تا انتقال مفاهیم مفید به کندی صورت پذیرد. آنان همچنین، مروری کلی بر الگوهای خوشه‌بندی از منظر الگوشناسی آماری، با هدف فراهم‌آوری مراجع و توصیه‌های مفید درخصوص مفاهیم بنیادین خوشه‌بندی ارائه می‌دهند. علاوه بر آن، در این اثر یک طبقه‌بندی از فنون خوشه‌بندی و پیشرفت‌های اخیر در زمینهٔ برش مقطعی موضوع‌های مختلف ارائه شده است. در ادامه، برخی از کاربردهای حائز اهمیت الگوریتم‌های خوشه‌بندی از قبیل قطعه‌بندی تصاویر، بازشناسی موضوعی و بازیابی اطلاعات توضیح داده می‌شود.

در پژوهش انجام یافته توسط برو و رید[5] (۲۰۰۵)، دربارهٔ خوشه‌بندی سلسله‌مراتبی

4. . Jain, Murty & Flynn
5. Breaux & Reed

اطلاعات، با استفاده از زبان‌های هستی‌شناسی، در سال ۲۰۰۵، مشخص شد که ابزارهای تحلیل و تجسم‌سازی اطلاعات از منابع ناهمگن متعدد، با تکیه بر بهبود و پیشرفت روش‌های آماری، بهره گرفته است. تحقیقات انجام شده درخصوص زبان‌های هستی‌شناسی، راه‌های نویدبخشی را برای غلبه بر محدودیت‌های موجود ارائه می‌دهد. در این تحقیق، معلوم شد که در زبان‌های هستی‌شناسی، استفاده از خصیصه‌های معنایی، که به شکل رمز درآمده‌اند، می‌تواند از طریق روش‌هایی مانند جست‌وجوی کلیدواژه و خوشه‌بندی برای تحلیل و تصویرسازی مدارک در سطوح ادراکی بالاتر به‌کار گرفته شود. آنان مشخص کردند که یافته‌های حاصل از نظام خوشه‌بندی سلسله‌مراتبی اصلاح‌شده برای شاخص‌بندی زبان‌های هستی‌شناسی و اجرای آن در آزمون‌های موضوعی مجموعهٔ مدرک‌هایی که کمتر از ۲۰۰ واژه را پوشش می‌دهند، به‌خوبی می‌تواند به‌کار گرفته شود.

جین و لا[6] (۲۰۰۵)، در مقالهٔ خود باعنوان «خوشه‌بندی داده‌ها: معمای غیرقابل حل کاربر»، در سال ۲۰۰۵، به تحلیل وضعیت خوشه‌ها در ارتباط با ابزارهای کاوش خودکار در برنامهٔ گروه‌سازی مجموعهٔ الگوها می‌پردازند. آنان تشریح می‌کنند که با وجودآنکه بیش از ٤٠ سال از پژوهش‌های مربوط به این حوزه می‌گذرد، هنوز چالش‌های زیادی در فرایند خوشه‌بندی داده، چه از لحاظ نظری و چه از ابعاد عملی وجود دارد. همچنین، آنان تلاش کرده‌اند تا چند پیشرفت اخیر در ارتباط با فعالیت خوشه‌بندی داده‌ها را توضیح داده و ویژگی‌های آن را بیان کنند. خوشه‌بندی دسته‌جمعی، گزینش خصیصه‌ها و خوشه‌بندی دارای محدودیت، از زمره مواردی است که در این اثر دربارهٔ آنها بحث شده است.

سونگ[7] (۲۰۰۵)، نیز در پژوهشی که در سال ۲۰۰۵ با عنوان «الگوریتم خوشه‌بندی اطلاعات بااهمیت» به انجام رسانده است، بر موضوع خوشه‌بندی اطلاعات بااهمیت، برمبنای بهینه‌سازی کمینه‌سازی و بیشینه‌سازی اطلاعات متقابل، تمرکز داشته است. از مزایای اصلی این روش خوشه‌بندی اطلاعات آن است که یک روش غیرپارامتری است، و الگوریتم سادهٔ خوشه‌بندی داده‌ها را براساس فاصلهٔ اقلیدسی مربع خطاها تشکیل می‌دهد.

آنتونی و دِژاردَن[8] (۲۰۰٦)، در مقاله‌ای که دربارهٔ مشکلات مطرح در خوشه‌بندی رابطه‌ای داده‌ها، در سال ۲۰۰٦، ارائه داده‌اند، بیان می‌کنند که وظیفهٔ خوشه‌بندی داده‌ها، شناسایی الگوها در مجموعه‌ای از داده‌هاست. بیشتر الگوریتم‌ها، داده‌های غیررابطه‌ای را به‌عنوان ورودی درنظر گرفته و گاهی نیز قادر به یافتن الگوهای معنی‌دار نیستند. بسیاری از مجموعهٔ داده‌ها، می‌توانند علاوه بر داشتن نشانه‌های موضوعی مستقل، شامل اطلاعات رابطه‌ای نیز باشند. آنان همچنین اظهار می‌دارند: در جایی که الگوریتم‌های غیررابطه‌ای شکست می‌خورند، خوشه‌بندی رابطه‌ای داده‌ها می‌تواند به یافتن الگوهای معنی‌دار کمک کند.

آنتونی و دِژاردَن (۲۰۰۷)، در اثر دیگر خود، دربارهٔ خوشه‌بندی داده‌ها با مدل رابطه‌ای

6. Jain & Law
7. Song
8. Anthony & desJardins

فشار- کشش، توضیح می‌دهند که خوشه‌بندی رابطه‌ای داده‌ها، نوعی یادگیری رابطه‌ای است که داده‌ها را با استفاده از ساختار رابطه‌ای مجموعه داده‌ها به‌صورت خوشه‌های متعدد درآورده تا فرایند خوشه‌بندی را هدایت کند. در این میان، روش‌های متعددی برای خوشه‌بندی رابطه‌ای پیشنهاد شده است. فرضیهٔ متداول در این پژوهش آن است که رابطه‌ها گرایش پیوندی دارند. یعنی فرض می‌شود لبه‌ها بیشتر در درون خوشه‌ها قرار می‌گیرند تا بین خوشه‌ها.

همچنین، در ارتباط با کاربرد خوشه‌بندی اطلاعات در بخش تجاری، پرلیش و روزت[9] (۲۰۰۷)، در تحقیقی که دربارهٔ شناسایی بسته‌های انتخابی محصول با استفاده از خوشه‌بندی اطلاعات متقابل به انجام رسانده‌اند، روش بدیعی را به‌منظور کاهش تعداد انتخاب‌های پیکربندی محصولات پیچیده، از طریق شناسایی مجموعه‌های معنی‌داری مشخص ساخته‌اند. آنان روش‌های سنجش مختلف آماری و نظریهٔ اطلاعات، تا ثبت میزان همبستگی میان گزینه‌های هر جفت مؤلفهٔ محصول را بررسی کرده‌اند. در این تحقیق، از خوشه‌بندی سلسله‌مراتبی برای شناسایی مجموعه‌های معنی‌داری از مؤلفه‌ها که می‌توانند با یکدیگر ترکیب شده تا تعداد مشخصات منحصربه‌فرد محصول را کاهش داده و استانداردسازی تولید را افزایش دهد، استفاده کرده‌اند. کانون توجه و تحلیل پدیدآورندگان بر بررسی تأثیر اختلاف در سنجش شباهت‌ها، در ارتباط با توانایی خوشه‌بندی برای یافتن خوشه‌های معنی‌دار است.

مفهوم خوشه‌بندی

جین، مورتی و فلین (۱۹۹۹)، درخصوص کاربرد خوشه‌بندی بیان می‌کنند که خوشه‌بندی برای انواع الگوی‌های تحلیل اکتشافی[10]، گروه‌بندی[11]، تصمیم‌گیری، و موقعیت‌های فراگیری ماشینی[12]، شامل: داده‌کاوی[13]، بازیابی مدارک[14]، بخش‌بندی تصویر[15]، و طرح رده‌بندی سودمند است. نویل[16] و همکاران (۲۰۰۳)، خوشه‌بندی را یک فعالیت توصیفی می‌دانند که شناسایی گروه‌بندی طبیعی داده‌ها را مورد کاوش قرار می‌دهد. جیونیز[17] (۲۰۰٤)، خوشه‌بندی را یک مرحلهٔ مهم از فرایند پردازش تحلیل داده[18]، همراه با کاربردهای آن در حوزه‌های متعدد، معرفی می‌کند و اظهار می‌دارد که براساس تعریفی ساده و اولیه، خوشه‌بندی به‌عنوان مقولهٔ بخش‌بندی داده‌ها درون گروه‌ها یا خوشه‌ها تعریف شده، که این داده‌های موضوعی در همان گروه مشابه مرتبط هستند؛ در صورتی‌که، این عناصر، در گروه‌های مختلف دارای مشابهت نیستند. کراسکف[19] و دیگران (۲۰۰۵)، اظهار می‌دارند که مقصود از خوشه‌بندی، جداسازی عناصر درون دسته‌هایی است که صرفاً در بُردار مشخصه[20] - مجموعه‌ای از

9. Perlich & Rosset
10. Exploratory pattern-analysis
11. Grouping
12. Machine-learning situations
13. Data mining
14. Document retrieval
15. Image segmentation
16. Neville
17. Gionis
18. Data analysis
19. Kraskov
20. Characteristic vector

اجزا و ویژگی‌ها- به‌کار می‌رود. روسل[21] (۲۰۰۶)، هدف خوشه‌بندی را بخش‌بندی یک مجموعهٔ ساختارنیافته از عناصر، درون خوشه‌ها یا گروه‌های مشخص معرفی می‌کند و شرح می‌دهد که شخص اغلب می‌خواهد اجزای خُرد را به‌عنوان عوامل مشترک در همان خوشه‌ای قرار دهد که دارای صفات یکسان هستند؛ و عناصر غیرمشترک را، تا حد ممکن، در خوشه‌ای جای دهد که به آن تعلق دارد. وی همچنین بیان می‌کند که خوشه‌بندی در بسیاری از حوزه‌های موضوعی به‌کار گرفته شده و انواع زیادی از الگوریتم‌های خوشه‌بندی برای مقاصد و موقعیت‌های متفاوت وجوددارد. کوالسکی[22] (۱۹۹۷)، با طرح کاربرد خوشه‌بندی در کتابخانه‌ها و مراکز اطلاع‌رسانی، هدف اصلی فرایند خوشه‌بندی را یاری‌رسانی به کاربر درتشخیص محل دقیق اطلاعات می‌داند، که این عمل در نهایت موجب تهیهٔ طرح‌های نمایه‌سازی برای سازماندهی بهتر مدارک در کتابخانه‌ها، و استانداردهای مرتبط با استفاده از نمایه‌های الکترونیکی[23] شده است. در مجموع، با بیان نظر متخصصان دربارهٔ مفهوم خوشه‌بندی می‌توان اظهار کرد که خوشه‌بندی عبارت است از: مرتب‌کردن واژه‌ها یا مدارک شبیه به‌هم در یک رده باعنوان کلی. از لحاظ کاربردی، خوشه‌بندی سبب بهینه‌سازی فعالیت جست‌وجوی اطلاعات شده و زمان جست وجوی کاربر را کاهش می‌دهد. در یک کاوش نظام‌مند و بر مبنای استفاده از پرس‌وجوهای مبتنی بر راهبردهای کاوش، خوشه‌بندی سبب ایجاد ارتباط میان خوشه‌های مختلف شده و در مجموع به نتایج سودمندی منجر می‌شود. از سوی دیگر، خوشه‌بندی سبب شده تا گروهی از موضوعات مشابه در زیر یک رده، باعنوان کلی سازماندهی شوند. این فعالیت، در دستیابی به اطلاعات مرتبط با موضوع خواسته شده تأثیر بسزایی دارد و باعث دسترسی مطلوب به اطلاعات هم‌موضوع می‌شود.

انواع خوشه‌بندی بر اساس روش کاری

فرایند خوشه‌بندی اطلاعات، روشی است که هرچه بهتر انجام شود، بازیابی اطلاعات دقیق‌تر انجام خواهد پذیرفت. توجه به کاربرد اثربخش خوشه‌بندی ما را به این نکته رهنمون می‌سازد که کلید خوشه‌بندی موفق در رعایت دو عامل مهم است: نخست انتخاب یک مقیاس مناسب برای تعیین تشابه‌ها؛ و دوم به‌کارگیری الگوریتم کارآمد، برای جانمایی مکان اجزا در رده‌های مشابه و مناسب. برای انجام چنین فعالیتی می‌توان از روش‌های دستی و یا ماشینی خوشه‌بندی بهره‌گرفت. هریک از روش‌های نامبرده دارای ویژگی‌های خاص خود است. در ادامه به توضیح این دو نوع خوشه‌بندی می‌پردازیم.

21. Rosell
22. Kowalski
23. Electronic indexes

خوشه‌بندی‌دستی

براساس بیان کوالسکی(۱۹۹۷، ص ۱۲۹– ۱۳۰)، خوشه‌بندی دستی به‌منظور تعیین موضوع و دامنهٔ موضوعی و در واقع حد و حدود فعالیت کاری انجام می‌پذیرد. این فرایند را به دو منظور انجام می‌دهیم: ۱) تقلیل اشتباهات مربوط به واژه‌های هم‌نوشت و یا هم‌املا، ۲)کمک به افزایش تمرکز در ایجاد خوشه‌ها. از واژه‌نامه‌ها به‌عنوان نقطهٔ شروع کار برای تهیهٔ اصطلاحنامه استفاده شده که منجر به تهیه و تولید کشف‌اللغات می‌شود. کشف‌اللغات[۲۴] عبارت است از فهرست الفبایی واژه‌ها، همراه با تعداد تکرار آن در متن و تعیین محل دقیق واژه. به‌عبارت دیگر، نمایه‌ای از واژه‌های اصلی هر اثر که جای آنها را در متن نشان داده و معمولاً واژه‌های قبل و بعد از مطلب را ذکر می‌کند و گاهی اوقات تعریف واژه‌ها را نیز ارائه می‌دهد. برای ساخت اصطلاحنامه، به‌صورت دستی، کافی است واژه‌های مناسب انتخاب شود، البته ابزارهای دیگری نیز در انتخاب واژه‌ها مفید خواهند بود از قبیل: لغات کلیدی خارج از متن[۲۵]، لغات کلیدی داخل متن[۲۶] لغات کلیدی و متن[۲۷]، و لغات کلیدی خارج از عنوان[۲۸].

خوشه‌بندی‌ماشینی

خوشه‌بندی خودکار (ماشینی) واژه‌ها نوع دیگری از این فرایند است که انجام می‌پذیرد. برای ایجاد یک اصطلاحنامهٔ آماری، روش‌های زیادی برای ماشینی کردن واژه‌ها وجود دارد. تمامی آنها به‌عنوان اساس کار، از این مفهوم استفاده می‌کنند که هر چه تکرار دو واژه در بخش‌های مشابه بیشتر باشد احتمال اینکه مفاهیم مشابهی داشته باشند بیشتر است. در ساده‌ترین مورد، داده‌ها یکبار در ایجاد رده‌ها به‌کار گرفته می‌شود. وقتی تعداد رده‌های ایجاد شده بسیار زیاد باشد، ممکن است رده‌های ابتدایی برای ایجاد رده‌های خلاصه‌تر یک سلسله‌مراتب ایجاد کنند، که به‌عنوان نقطهٔ شروع استفاده می‌شوند. از ویژگی‌های این نوع خوشه‌بندی آن است که حاوی رده‌هایی هستند که کاربرد واژه‌ها را منعکس می‌کنند. همچنین، رده‌ها به‌طور طبیعی اسم ندارند؛ بلکه به‌صورت گروه‌هایی با شرایط آماری مشابه هستند. به‌علاوه، تکنیک بهینه برای ایجاد رده‌ها نیاز به محاسبات زیادی دارد (کوالسکی، ۱۹۹۷، ص ۱۲۹– ۱۳۱). خوشه‌بندی ماشینی روشی است که با تعیین معیارهای سنجش توسط رایانه به انجام می‌رسد و کلیهٔ فرایندهای خوشه‌بندی با تعیین رده‌های خاص در قالب هر خوشه، به‌نحوی مؤثر در سازماندهی و دسته‌بندی داده‌ها مؤثر واقع می‌شود. تنظیم متعادل هر خوشه با بهره‌گیری از الگوریتم خاص تعریف شده برای برنامه اعمال می‌شود. این الگوریتم روند تعیین اجزای هر رده را به‌خوبی مشخص ساخته و از راهبردهای رده‌شناسی[۲۹] به‌نحو مطلوبی بهره می‌گیرد.

24. Concordance
25. Key Word out of context (Kwoc)
26. Key Word in context (Kwoc)
27. Key Word and context (Kwoc)
28. Key Word out of Title (Kwoc)
29. Taxonomy

الگوریتم‌های خوشه‌بندی

برای انجام بهینهٔ خوشه‌بندی، باید از الگوریتم‌های مناسب این کار استفاده کرد. در این حوزه و فعالیت تخصصی، الگوریتم‌ها به ما کمک می‌کنند تا با بهره‌گیری از راه‌حل‌های فرمولی و مدون به خوشه‌بندی مطلوب اطلاعات مبادرت ورزیم. این الگوریتم‌ها در رده‌های مختلف و با کاربردهای گوناگون معرفی شده‌اند. شناخت آنها به متخصصان کمک می‌کند تا به شیوه‌ای مناسب به انجام فرایند خوشه‌بندی بپردازند.

ماتوچی[30] (۲۰۰۳)، الگوریتم‌های خوشه‌بندی را در چهار ردهٔ کلی معرفی می‌کند که‌شامل: خوشه‌بندی انحصاری[31]، خوشه‌بندی همپوشانی[32]، خوشه‌بندی سلسله‌مراتبی[33]، و خوشه‌بندی احتمالی[34] می‌شود. این تقسیم‌بندی کلی است و برمبنای انواع کاربردهای خوشه‌بندی ارائه شده است. برخی از این خوشه‌بندی‌ها در فعالیت‌های این حوزه مورد توجه قرارگرفته، و در مواردی بسط داده شده است. چنانچه لیوسکی[35] (۲۰۰۲)، با توضیح پیرامون الگوریتم خوشه‌بندی متراکم سلسله‌مراتبی[36] اظهار می‌کند که این الگوریتم، سلسله‌ای از خوشه‌ها را ایجاد می‌کند که با ساخت ارتباط درختی، در جایی‌که هر گره به‌عنوان خوشه‌ای از مفاهیم مطرح می‌شود، خوشه‌ها را مطابق با زیرمجموعه – فرزند– خود قرار می‌دهد که از یک بخش کامل از آن خوشه اخذ شده‌اند. کانانگو و همکاران (۲۰۰۲)، به شرح الگوریتم خاصی در خوشه‌بندی می‌پردازند که به نام الگوریتم خوشه‌بندی کی‌مینز[37] شناخته شده است. این روش، یکی دیگر از الگوریتم‌های سودمند برای فرایند خوشه‌بندی است. یک بحث اکتشافی عام درخصوص این خوشه‌بندی الگوریتم لیود[38] می‌باشد، که آن را الگوریتم پالایش[39] نیز می‌نامند. کاربرد این الگوریتم ساده است و به راهبرد درخت کی‌دی[40]، به‌عنوان تنها ساختار دادهٔ اصلی، نیاز دارد. الگوریتم کی‌مینز به‌شکل اکتشافی برای حل مشکلات مطرح در این روش مورد استفاده قرار می‌گیرد و بر مبنای طرح تکرارشونده نمونه[41] پایه‌ریزی شده است، که جهت دستیابی به یک راه‌حل حداقلی به‌صورت موضعی عمل می‌کند. کوهن[42] (۲۰۰۶) نیز، با طرح موضوع خوشه‌بندی معنایی[43]، بیان می‌کند که این شیوه یک روش غیرتعاملی و کنترل نشده است، که در ارتباط نزدیک با کارکردهای نمایه‌سازی معنایی پنهان[44] قرار دارد. دربارهٔ این روش نمایه‌سازی، نظام نرم‌افزاری درون خوشه‌ها گروه‌بندی می‌شود. خوشهٔ معنایی، یک رسته از موضوع‌هایی است که در همان مجموعهٔ واژگان به‌کار برده می‌شود. براین‌اساس، هر خوشه، مفاهیم متفاوتی را مشخص می‌سازد که در نظام قابل بازیابی است. در نهایت، به‌طور ذاتی، مفاهیم بدون مشخصه، بابهره‌گیری از برچسب‌های ویژه، از واژگان کدهای منبع مشخصه دریافت می‌کنند. الگوریتم خودکار، هر خوشه را با واژگانی که بیشترین ارتباط را دارند، برچسب‌گذاری

30. Matteucci
31. Exclusive clustering
32. Overlapping clustering
33. Hierarchical clustering
34. Probabilistic clustering
35. Leuski
36. Hierarchical agglomerative clustering algorithm
37. K-means clustering algorithms
38. Lloyd's algorithm
39. Filtering algorithm
40. K-D Tree
41. Simple iterative scheme
42. Kuhn
43. Semantic clustering
44. (Latent Semantic Indexing (LSI

می‌کند. این روش، برای انسان قابل درک بوده و به‌منظور فهم مفاهیم اصلی در نظام نرم‌افزاری تهیه می‌شود. روسل (۲۰۰۶)، مطرح می‌سازد که برای به‌کارگیری الگوریتم خوشه‌بندی[45] دو عامل مورد نیاز است: ارائهٔ یک عنصر و یا موضوع ویژه، و نیز معیار اندازه‌گیری شباهت یا اختلاف میان آن عنصر یا موضوع خاص. الگوریتم خوشه‌بندی، بخش‌بندی یک مجموعه از عناصر ویژه، در ارتباط با برخی از معیارهای پایه در چنین شرایطی را تکمیل می‌کند. وکالی[46] (۲۰۰۷)، دربارهٔ منشأ و معیار کلی الگوریتم بیان می‌کند که الگوریتم‌های خوشه‌بندی از حوزه‌های مختلف تخصصی چون آمار، شناسایی و تشخیص الگو، و نیز یادگیری ماشینی نشأت گرفته‌اند. هر طرح خوشه‌بندی بهینه اصولاً باید دو معیار مهم را رعایت کرده باشد: نخست، فشرده‌سازی یعنی[47] داده‌های درون هر خوشه باید تا حد ممکن به یکدیگر نزدیک باشند. روش معمول سنجش انسجام، واریانس یا مجذور انحراف معیار است که باید در حداقل باشد. دوم، جداسازی[48] است، به‌این‌معنا که خوشه‌ها باید به‌طور قابل ملاحظه‌ای از هم جدا باشند. مفهوم فاصلهٔ خوشه‌ای معمولاً با تعیین اندازهٔ تفکیک به‌کار می‌رود، که باید در مرز حداکثر باشد.

برای خوشه‌بندی فایل‌های داده‌های دیجیتالی، در ارتباط با داده‌های چندبُعدی، الگوریتم‌های خاص خوشه‌بندی وجود دارد که می‌توان با بهره‌گیری از آنها به تبیین وضعیت داده‌ها پرداخت. داده‌های چندبُعدی، یکی از انواع داده‌های مبتنی بر بُعد هستند که در فرایندهای مختلف رایانه‌ای کاربرد دارند. از جمله کاربری‌های آنها می‌توان به راهبردهای مصورسازی[49] اشاره‌کرد، که به نحوی با رویکردهای دیداری در نظام‌های بازیابی اطلاعات ارتباط دارد. برای تعیین گروه‌بندی نوع خاص داده‌های ابعاد بالا[50]، می‌توان از الگوریتم خوشه‌بندی داده[51] بهره گرفت. در مجموع، براساس انواع الگوریتم‌های پیشنهادی که توسط متخصصان این حوزه ارائه شده، می‌توان نتیجه گرفت که ساختار گوناگون الگوریتم‌های خوشه‌بندی براساس نوع داده، ارتباط میان آنها، و ویژگی‌های خاص موجود با معیارهای تعیین شده به‌صورت کاربردی در فرایندهای خوشه‌بندی به‌کار گرفته می‌شود. این الگوریتم‌ها فرایند کاری خوشه‌بندی را فرمول‌بندی کرده، و با تدوین مشخص آن به یاری روش سازماندهی خوشه‌ها می‌شتابند. الگوریتم‌های خوشه‌بندی با ارائهٔ الگوهای پیشنهادی در مسیر بهینه‌سازی کارکرد خوشه‌بندی حرکت می‌کنند.

45. Clustering algorithm
46. Vakali
47. Compactness
48. Separation
49. Visualization
50. Data high-dimensional
51. Data clustering algorithm

مراحل خوشه‌بندی

در فعالیت‌های حرفه‌ای کتابداری و اطلاع‌رسانی، برای نخستین‌بار خوشه‌بندی در کاربرد خاص واژه‌ها در تزاروس یا اصطلاحنامه به‌کار گرفته شد. یکی از کاربردهای اولیهٔ خوشه‌بندی،

رده‌بندی مدارک با موضوعات مشابه بوده است. در اصطلاحنامه‌های تخصصی برای ایجاد ارتباط منطقی میان مفاهیم مشترک و اصطلاح‌های مرتبط از راهبردهای خوشه‌بندی به شیوهٔ اثربخش استفاده شده است. در این میان، شناخت مراحل مختلف خوشه‌بندی به ما کمک می‌کند تا بتوانیم به‌شیوهٔ مناسبی از آن بهره‌گیری کنیم.

کوالسکی (۱۹۹۷، ص ۱۲۶– ۱۲۷)، برای خوشه‌بندی چهار مرحله ذکر می‌کند. مرحلهٔ نخست تعریف دامنهٔ موضوعی خوشه‌بندی است. در این مرحله، حدود کار که دارای یک دامنهٔ کاربردی است، مشخص می‌شود. مانند فعالیتی که برای تهیهٔ یک اصطلاحنامهٔ پزشکی یا علوم کتابداری و اطلاع‌رسانی به‌انجام می‌رسد. مرحلهٔ دوم، تعیین خصوصیات و مشخصات واژه‌ها و مدارکی است که باید خوشه‌بندی شوند. به‌تعبیر دیگر، موضوع‌ها شامل چه بخش‌هایی هستند؛ نظیر تعیین خصوصیات مفرد و جمع اصطلاح‌ها. مرحلهٔ سوم، تعیین میزان رابطهٔ بین ویژگی‌های واژه یا مدرک، با رده‌ای است که در آن قرار می‌گیرد؛ مانند تعیین این نکته که در فرهنگ مترادف‌ها چه کلماتی با یکدیگر هم‌معنا هستند و میزان ربط معنایی آنها به چه میزان است. مرحلهٔ نهایی نیز تهیهٔ یک الگوریتم– راه‌حل فرمولی– برای تعیین رده‌های هر مقوله است و شامل مدارک و واژه‌هایی می‌گردد که به آن اختصاص داده می‌شود.

هریک از مراحل نامبرده در حیطهٔ خاص خود به‌شکل کاربردی به مرحلهٔ اجرا درمی‌آید. تعریف دامنهٔ موضوعی یکی از فرایندهای مهم کاری به‌شمار می‌آید. در این مرحله، ساختار اصلی رده‌ها و عناصر اطلاعاتی که باید درون دسته‌های معین قرار گیرند، مشخص می‌شود. حیطهٔ کلی رده‌ها، ارتباط میان رده‌های ایجاد شده و سازه‌های مهمی که در ساخت آن نقش دارد، مورد بررسی و تحلیل قرار می‌گیرد. ویژگی‌های مربوط به واژه‌ها که فرایند خوشه‌بندی، با توجه به آن، به مرحلهٔ اجرا درمی‌آید، از موارد دیگر است. شناسایی و تحلیل واژه‌ها و روابط میان آنها در فعالیت خوشه‌بندی نقشی مؤثر ایفا می‌کند. ارتباط میان واژه‌ها و اصطلاح‌ها از جنبه‌های متعدد مورد بحث واقع می‌شود. روابط اعم و اخص، کل و جزء، رابطهٔ هم‌ارز، ارتباط معنایی، و تعیین مترادف‌ها و متضادها از مواردی است که باید به‌دقت بررسی شده و در ساختار رده‌های مربوط به فرایند خوشه‌بندی لحاظ شود. یکی از موارد حساس و تعیین‌کننده، انتخاب الگوریتم خاص راهبرد خوشه‌بندی است. الگوریتم‌های خوشه‌بندی در واقع، مرحلهٔ اجرایی کردن فعالیت خوشه‌بندی به‌شمار می‌آیند. ساختار الگوریتم، با توجه به شیوهٔ کاری در ارتباط با گروه‌بندی کلی مفاهیم، تعیین دامنهٔ موضوعی واژگان، مشخص ساختن رده‌ها، ایجاد ارتباط میان آنها، تحلیل اصطلاح‌های انتخاب شده، ساختار رسته‌بندی[52] گروه‌های مختلف اجزای داده‌ها، و در

52. Categorization

مجموع روش اجرایی و کاربرد فرایندهای خوشه‌بندی اطلاعات طراحی می‌شود. در ادامه، به توضیح بیشتری دربارهٔ الگوریتم‌های خوشه‌بندی می‌پردازیم.

روابط خوشه‌بندی

آی تچیسون و گیل‌کریست[54] (۱۹۷۲)، سه نوع رابطه را مشخص کرده‌اند:

رابطهٔ هم‌ارز[55] که متداول‌ترین رابطه بوده و مترادف‌های واژه را ارائه می‌دهد. در این نوع رابطه، واژه‌هایی موردنظر است که همپوشانی معناداری بین آنها وجود دارد، اما از لحاظ واژگانی متفاوت هستند. مانند درد که دارای مترادف‌هایی چون الم، بیماری، تألم، رنج، سوز، کسالت، و مرض است؛ یا واژه‌های عکس و چاپ که گاهی می‌تواند مترادف هم تعریف شود (با توجه به اینکه کلمه چاپ شامل لیتوگرافی نیز می‌شود).

رابطهٔ سلسله‌مراتبی[56] در این رابطه، یک واژه به‌عنوان ردهٔ اصلی انتخاب شده و مدخل‌ها، زیرمجموعه‌ها یا نمونه‌های خاصی از واژهٔ کلی هستند. نظیر رایانه که به‌عنوان رده اصلی انتخاب، و ریزپردازنده‌ها، پنتیوم، و مانند آن به‌عنوان زیرمجموعه در زیرردهٔ کلی قرار می‌گیرد.

رابطهٔ غیرسلسله‌مراتبی[57] که انواع دیگر روابط بین واژه‌ها، غیر از دو مورد قبلی، از قبیل موضوع و ویژگی‌های مربوط را دربرمی‌گیرد، به‌عنوان مثال: کارمند←عنوان شغل.

در سال ۱۹۸۵، ونگ[58] و دیگران (۱۹۸۵)، طرح جدیدتری از ارتباط واژه‌ها را ارائه کردند که شامل پنج رابطهٔ جزء- کل[59]، همنشینی- ترتیبی[60]، نمونه‌ای[61]، رده‌شناسی و مترادف‌ها[62]، و رده‌شناسی و متضادها[63] بود. در این میان، ارتباطات همنشینی و نمونه‌ای نیازمند توضیح بیشتری هستند. رابطهٔ میان واژه‌ها به‌صورت همنشینی، ترتیبی یک مقیاس آماری است که از رابطهٔ لغاتی که در کنار هم می‌آیند (در جمله، عبارت و یا پارگراف) تشکیل می‌شود. مانند سلام، حالت چطور است؛ و یا خسته‌ام، حالم خوب نیست. رابطهٔ میان واژه‌ها به‌صورت نمونه‌ای عبارت است از واژه‌هایی که معنای مشابهی دارند و در یک رده قرار می‌گیرند. نظیر فرمول و معادله. رابطهٔ بین واژگان از نظر معنای واژه‌ها، روابط دیگری را نیز دربرمی‌گیرد. به‌طور نمونه، لغات متقابلی چون خُرد و جزء که بیان کنیم «کُره» جزء احجام هندسی است. یا کلان و کل، نظیر آنکه بگوییم «پی» که بخشی از ساختمان است. یا مثلاً لغت‌هایی که مفاهیم دیگری را برای ما تداعی می‌کند. مثل: عطر که با واژه‌های بوی خوش، رایحه، و شمیم ارتباطی نزدیک دارد.

برقراری ارتباط‌های متعدد میان واژه‌ها برای دستیابی به اطلاعات مرتبط از اهمیت بالایی برخوردار است. ایجاد رابطه در رده‌های تزاروس یا اصطلاحنامه، در کاربرد خاص

54. Aitchison & Gilchrist
55. Equivalence relationship
56. Hierarchical relationship
57. Non-hierarchical relationship
58. Wang
59. Parts-Wholes
60. Collocation
61. Paradigmatic
62. Taxonomy and synonymy
63. Taxonomy and antonymy

نظام‌های اطلاع‌رسانی، در واقع برای بالا بردن میزان بازیابی سودمند و مرتبط اطلاعات با نیازهای اطلاعاتی مطرح شده صورت می‌پذیرد. از این‌رو، کاربردی ساختن روابط واژه‌ها در واقع به‌منظور دستیابی به اطلاعات مؤثر است. ارتباط واژه‌ها از این حیث که سطوح متعدد برداشت کاربر از مفاهیم را فراهم می‌سازد قابل توجه است. فنون برقراری رابطه میان مجموعهٔ عناصر و واژگان، بر مبنای نوع کاربردی که از آن انتظار می‌رود، متفاوت است، که در روش‌های خوشه‌بندی بیان می‌شود.

روش‌های خوشه‌بندی

برای انجام مؤثر فرایند خوشه‌بندی داده‌ها، می‌توانیم از روش‌های مختلفی بهره گیریم. این روش‌ها، در واقع، برای انجام بهتر تنظیم رده‌های موضوعی در چارچوب گروه‌های همسان و مشترک به‌کار گرفته می‌شوند. یکی از این شیوه‌ها، روش ارتباط کامل بین واژه‌هاست. در این شیوه، شباهت موجود میان هر جفت واژه به‌عنوان مبنایی برای مشخص کردن رده‌ها محاسبه می‌شود. ساده‌ترین راه برای فهم این شیوه، در نظر گرفتن یک مدل بُرداری است. مدل بُرداری توسط یک ماتریس نشان داده می‌شود که در آن سطرها، موارد مستقل، و ستون‌ها، واژه‌های خاصی هستند. ارقام نیز نشانگر میزان تکرار واژه در هر مدرک هستند. در واقع، اعداد در هر ردیف، نشانگر میزان تکرار و انباشتگی لغت در هر مدرک هستند. شکل ۱ نمونه‌ای از یک بُردار را نشان می‌دهد. در این شکل میان واژه‌ها و عناصر تعیین شده رابطهٔ مستقیم وجود دارد. با بهره‌گیری از فرمول تعیین شباهت می‌توان به ایجاد ماتریس‌های محاسباتی واژه به واژه و روابط واژه مبادرت ورزید.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Item 1	0	4	0	0	0	2	1	3
Item 2	3	1	4	3	1	2	0	1
Item 3	3	0	0	0	3	0	3	0
Item 4	0	1	0	3	0	0	2	0
Item 5	2	2	2	3	1	4	0	2

شکل ۱

نمونه‌ای از یك بُردار
(کوالسکی، ۱۹۹۷، ص ۱۳۲)

برای درک بهتر این روش، از فرمول سادهٔ زیر استفاده می‌کنیم:

$$SIM^{64}(Term_i, Term_j) = \sum (Term_{k,i})(Term_{k,j})$$

در این فرمول، SIM نشان‌دهندهٔ میزان مشابهت عناصر در فرایند خوشه‌بندی است. i, j مقادیر کمی از واژه‌های موجود در یک گروه خوشه‌بندی شده از اطلاعات است. k نیز مجموعهٔ تمامی مقوله‌هاست. و نیز $\sum$ جمع کل است.

در نتیجه، فرمول، دو ستون از واژهٔ تحلیل شده را گرفته و پس از ضرب آنها مقادیر مجموع را در هر سطر به‌دست می‌آورد. نتایج را می‌توان در یک ماتریس m×m به نام ماتریس واژه– واژه قرار داد (شکل ۲) که m تعداد واژه‌ها در ماتریس اصلی است. ارزش‌های ارائه شده در این ماتریس نشانگر همبستگی هر واژه با خود آن واژه و واژه‌های دیگر است.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Term 1		7	16	15	14	14	9	7
Term 2	7		8	12	3	18	6	17
Term 3	16	8		18	6	16	0	8
Term 4	15	12	18		6	18	6	9
Term 5	14	3	6	6		6	9	3
Term 6	14	18	16	18	6		2	16
Term 7	9	6	0	6	9	2		3
Term 8	7	17	8	9	3	16	3	

شکل ۲

ماتریس واژه به واژه
(کوالسکی، ۱۹۹۷، ص ۱۳۳)

64. SIM= Similarity

مرحلهٔ بعد، انتخاب آستانه‌ای برای تعیین درجهٔ شباهت دو واژه، به‌منظور قرار دادن آنها در یک رده است. در این مثال، آستانه عدد ۱۰ درنظر گرفته شده است. بنابراین، در صورتی که مقدار تشابه بین آنها ۱۰ یا بیشتر باشد، دو واژه در کنار هم قرار می‌گیرند. با این اقدام، یک ماتریس دوتایی جدید به‌نام ماتریس رابطه واژه‌ها (شکل۳) ایجاد می‌شود و واژه‌هایی را تعریف می‌کند که مشابه هم هستند. مرحلهٔ نهایی در ایجاد رده‌ها، مشخص کردن این معناست که در چه صورت دو واژه در خوشه‌ای یکسان قرار می‌گیرند. برای این کار روش‌های بسیار متفاوتی وجود دارد که موجب می‌شود خوشه‌بندی به شیوه‌های زیر به‌وجود آید. متداول‌ترین آنها عبارت اند از: خوشه‌بندی دسته‌ای[65]، خوشه‌بندی تک پیوندی(یک پیوندی)[66]، خوشه‌بندی ستاره‌ای[67]، و خوشه‌بندی رشته‌ای[68] (کوالسکی، ۱۹۹۷، ص ۱۳۲). در ادامه به توضیح هریک از انواع نامبرده خواهیم پرداخت.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Term 1		0	1	1	1	1	0	0
Term 2	0		0	1	0	1	0	1
Term 3	1	0		1	0	1	0	0
Term 4	1	1	1		0	1	0	0
Term 5	1	0	0	0		0	0	0
Term 6	1	1	1	1	0		0	1
Term 7	0	0	0	0	0	0		0
Term 8	0	1	0	0	0	1	0	

شکل ۳

ماتریس روابط واژه
(کوالسکی، ۱۹۹۷، ص ۱۳۳)

خوشه‌بندی دسته‌ای: در این روش تمامی واژه‌هایی که مشابه تشخیص داده شده‌اند، با توجه به مقدار ورودشان، در یک رده قرار می‌گیرند. باید توجه کرد که واژهٔ ۱ و ٦ در بیش از یک رده قرار گرفته‌اند. یکی از مهم‌ترین خصوصیات این شیوه آن است که واژه‌ها

65. Cliques clustering
66. Single link clustering
67. Star clustering
68. String clustering

را می‌توان در چندین رده یافت. با به‌کارگیری الگوریتم برای شکل ۳ رده‌های زیر ایجاد می‌شوند که در پنج رده می‌توان آنها را نمایش داد:

Class 1 (Term 1, Term 3, Term 4, Term 6)

Class 2 (Term 1, Term5)

Class 3 (Term 2, Term4, Term 6)

Class 4 (Term 2, Term 6, Term 8)

Class 5 (Term 7)

خوشه‌بندی تک‌پیوندی(یک‌پیوندی): در این شیوه، هر واژه‌ای که شبیه واژهٔ دیگری باشد در یک رده قرار می‌گیرد و یک واژه نمی‌تواند در دو رده قرار داشته باشد. با توجه به الگوریتم به‌کار رفته برای تهیهٔ خوشه‌ها با استفاده از شیوهٔ تک‌پیوندی ماتریس رابطه واژه در شکل ۳ به ایجاد رده‌های زیر منجر می‌شود که در قالب دو ردهٔ اصلی قابل نمایش است:

Class 1(Term 1, Term 3, Term 4, Term 5, Term 6, Term 2)

Class 2 (Term 7)

خوشه‌بندی ستاره‌ای: در این روش، یک واژه را انتخاب کرده و سپس تمام واژه‌های مربوط به آن واژه را در همان رده قرار می‌دهیم. واژه‌هایی که هنوز در رده‌ای قرار نگرفته‌اند، به‌عنوان هسته‌های جدیدی انتخاب می‌شوند تا اینکه تمام واژه‌ها به رده‌ای اختصاص یابند. در این شیوه ،یک واژه می‌تواند در چندین رده قرار گیرد. رده‌های زیر حاصل خوشه‌بندی ستاره‌ای است که در سه ردهٔ اصلی ظاهر شده‌اند:

Class 1 (Term 1, Term 3, Term 4, Term 6)

Class2 (Term 2, Term 4, Term 8)

Class 3 (Term 7)

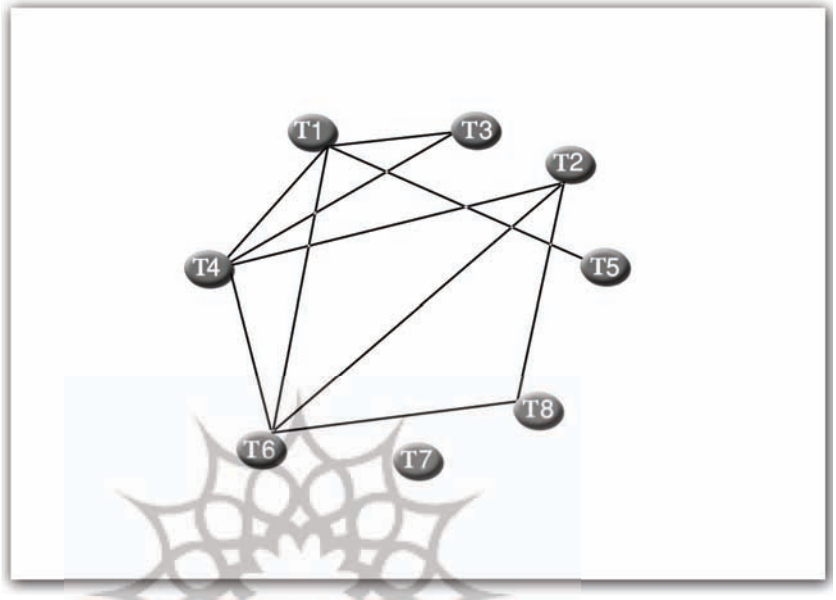
خوشه‌بندی رشته‌ای: در روش خوشه‌بندی رشته‌ای کار با یک واژه شروع می‌شود و واژه‌های مشابه، به‌صورت رشته‌ای به‌آن متصل می‌شوند. این فرایند ادامه می‌یابد تا اینکه هیچ واژهٔ دیگری در این رده قرار نگیرد. در این روش، واژه‌های جدید به‌عنوان یک گروه جدید استفاده شده، و بر این اساس فرایند ادامه می‌یابد. با تکیه بر این رویکرد، سه ردهٔ اصلی ایجاد شده که به‌صورت زیر به نمایش درمی‌آید:

Class 1 (Term 1, Term 3, Term 4, Term 2, Term 8, Term 6)

Class 2 (Term 5)

Class 3 (Term 7)

براین اساس، میان اصطلاح‌های تعریف شده ارتباطی چندوجهی ایجاد می‌شود که می‌توان این روابط را به‌صورت یک نمودار شبکه‌ای با ترسیم خطوط ارتباطی به شکل زیر نشان داد:



شکل ۴

نمودار شبکه‌ای مشابهت‌های اصطلاح (کوالسکی، ۱۹۹۷، ص ۱۳۵)

به‌عنوان نتیجه‌گیری از مبحث روش‌های خوشه‌بندی، در صورتی که بخواهیم یک مقایسه میان دو نوع از انواع مهم و پراستفادهٔ خوشه‌بندی، یعنی خوشه‌بندی دسته‌ای و خوشه‌بندی تک‌پیوندی انجام دهیم باید به نکات زیر اشاره کنیم:

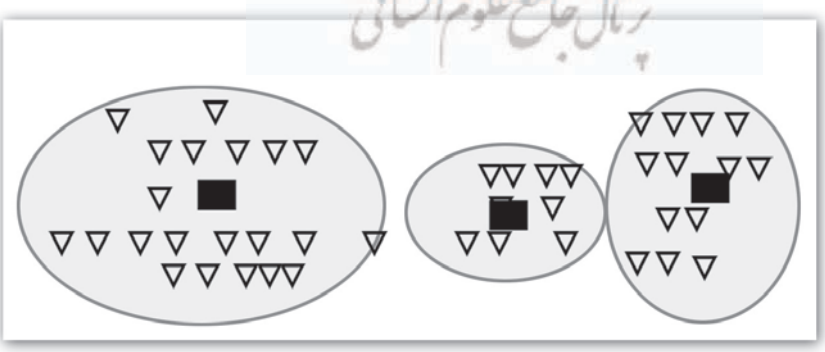
خوشه‌بندی دسته‌ای، رده‌هایی را ایجاد می‌کند که قوی‌ترین روابط بین تمام واژه‌های آن وجود دارد؛ این رده احتمالاً مفهوم خاصی را تشریح می‌کند؛ بیشترین رده‌ها را نسبت به سایر الگوریتم‌ها ایجاد می‌کند؛ تعداد واژه‌های یک رده را کاهش می‌دهد؛ در این الگوریتم مانعیت بیشتری وجود داشته و جامعیت آن کاهش می‌یابد (کوالسکی، ۱۹۹۷، ص ۱۳۳–۱۳۶). ولی ویژگی‌های خوشه‌بندی تک‌پیوندی را می‌توان به شرح ذیل برشمرد: واژه‌ها را بین رده‌ها تقسیم‌بندی می‌کند (جزءبندی)؛ این روش کمترین تعداد رده را ایجاد می‌کند؛ ناپایدارترین ارتباط بین لغات را درنظر می‌گیرد؛ این احتمال نیز وجود دارد که دو واژه‌ای که مشابهت آنها نزدیک به صفر است در یک رده قرار گیرد (سالتون،۱۹۷۲)؛ رده‌ها به‌جای رساندن یک مفهوم، تنوعی از مفاهیم را دربرمی‌گیرند؛ جامعیت را به حداکثر می‌رساند ولی مانعیت را کاهش می‌دهد.

خوشه‌بندی با استفاده از خوشه‌های موجود[۶۹]

یک روش دیگر برای ایجاد خوشه‌ها، استفاده از مجموعهٔ خوشه‌های موجود در شروع کار است. این شیوه، تعداد محاسبات مورد نیاز برای تعیین مشابهت‌ها در تهیهٔ خوشه‌ها را کاهش می‌دهد. روش نامبرده دارای ویژگی‌هایی به این قرار است: برای به‌حداقل رساندن محاسبات، مراکز ثقل برای هر خوشه تعیین می‌شود. در فیزیک به مرکز جرم یک مجموعه از اجسام، مرکز ثقل گفته می‌شود. در زمینهٔ بُردارها، این مفهوم برابر میانگین تمام بُردارها در یک خوشه است؛ اولین تخصیص واژه به خوشه‌ها، مراکز ثقلی را ایجاد می‌کند که ارتباطی به خوشه‌بندی نهایی ندارد؛ تشابه بین کل واژه‌های موجود و مراکز ثقل خوشه‌ها قابل محاسبه است؛ واژه به خوشه‌هایی راه می‌یابد که بیشترین تشابه را داشته باشند؛ این روند آنقدر تکرار می‌شود تا اینکه به یک وضعیت پایدار برسد (کوالسکی، ۱۹۹۷، ص ۱۳۷). نمایش نمودارهای واژه‌ها و مراکز ثقل اولیه نشان می‌دهد که چگونه رده پس از تعیین مکان اولیه جابه‌جا می‌شود. در واقع، خوشه‌ها، سازماندهی و ترکیب نهایی رده‌های اصلی را درون ساختار منظم گروه‌بندی برعهده دارند.

شکل ۵

مراکز ثقل اولیه برای خوشه‌ها (کوالسکی، ۱۹۹۷، ۱۳۷)



شکل ۶

مراکز ثقل پس از تخصیص مجدد واژه‌ها (کوالسکی، ۱۹۹۷، ص ۱۳۷)

69. Clustering using existing clusters

مربع توپر سیاه نشانگر مراکز ثقل رده‌هاست که در این روش با تخصیص مجدد واژه‌ها به ایجاد یک وضعیت متعادل منجر شده است. می‌توان به‌عنوان نتیجه رده‌های سه‌گانهٔ زیر را مشخص ساخت:

Class 1= (Term 1 and Term 2)

Class 2= (Term 3 and Term 4)

Class 3= (Term 5 and Term 6)

به این ترتیب مراکز ثقل زیر برای هر رده به‌شرح زیر مشخص می‌شود:

Class 1= (0+4)/2, (3+1)/2, (3+0)/2, (0+1)/2, (2+2)/2

= 4/2,4/2,3/2,1/2,4/2

Class 2= 0/2,7/2,0/2,3/2,5/2

Class 3= 2/2,3/2,3/2,0/2,5/2

علاوه بر آن، جدول مخصوص انتقال ردهٔ تکرار شونده در قالب زیر قابل ارائه است. در این جدول، موقعیت هر واژه در ارتباط با هریک از رده‌ها مشخص شده و محاسبهٔ مربوط به آن در جدول درج می‌شود:

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Class 1	29/2	29/2	24/2	27/2	17/2	32/2	15/2	24/2
Class 2	31/2	20/2	38/2	45/2	12/2	34/2	6/2	17/2
Class 3	28/2	21/2	22/2	24/2	17/2	30/2	11/2	19/2
Assign	Class2	Class1	Class2	Class2	Class3	Class2	Class1	Class1

شکل ۷

انتقال‌های رده تکرار شونده (کوالسکی، ۱۹۹۷، ص ۱۳۸)

اگرچه این شیوه نیاز به محاسبات کمتری نسبت به روش رابطه کامل واژه دارد، دارای برخی محدودیت‌های ذاتی نیز می‌باشد. یکی از مشکلات این روش، آن است که ابتدا تعداد رده‌ها تعریف می‌شود و نمی‌توان آنها را توسعه داد. این امکان وجود دارد که در انتهای فرایند، تعداد رده‌ها کمتر شود. تمام واژه‌ها باید به یک رده اختصاص یابند. بنابراین،

حتی اگر مشابهت آنها نسبت به سایر واژه‌های قرار گرفته بسیار ضعیف باشد، واژه‌ها را مجبور به قرارگرفتن در یک رده می‌کند. وضعیت جدید گروه‌بندی منظم و متعادل برای تخصیص واژه‌ها در قالب ردههای تعیین شده، به شکل جدول زیر قابل نمایش است.

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
Class 1	23/3	45/3	16/3	27/3	15/3	36/3	23/3	34/3
Class 2	67/4	45/4	70/4	78/4	33/4	72/4	17/4	40/4
Class 3	12/1	3/1	6/1	6/1	11/1	6/1	9/1	3/1
Assign	Class2	Class1	Class2	Class2	Class3	Class2	Class3	Class1

شکل ۸

انتقال رده‌ها و مراکز ثقل جدید (کوالسکی، ۱۹۹۷، ص ۱۳۹)

کاربرد خوشه‌بندی اطلاعات در ساختار اصطلاحنامه

در فرایند خوشه‌بندی رده‌های موضوعی در ساختار اصطلاحنامه تعیین وضعیت رده‌ها از اهمیت زیادی برخوردار است. بر این اساس باید هر رده کاملاً تعریف شده و از معنای مناسبی برخوردار باشد، زیرا گاهی اوقات واژهٔ انتخاب شده باعث سوءتعبیر می‌شود. بنابراین، بهتر است از اعداد برای شناسایی رده‌ها استفاده کرد؛ اندازهٔ رده‌ها باید از نظر بزرگی تقریباً یکسان باشد. یکی از مصارف ابتدایی رده‌ها، بسط دادن جست‌وجوها یا نتیجهٔ موارد بازیابی شده است. داخل هر رده، مدارک و واژه‌ها نباید بر ردهٔ اصلی برتری داشته باشد. برای مثال، کامپیوتر یا ریزپردازنده و نظایر آن؛ یک واژه یا مدرک می‌تواند به چند رده تعلق داشته باشد و یا به‌عنوان موضوعی اصلی انتخاب شود. این فرایند با تأکید بر دو رویکرد به انجام می‌رسد: نخست آنکه به‌صورت اخص باشد، یعنی واژه‌ای که بر موضوعی خاص دلالت دارد و مفهوم آن به مراتب محدودتر از واژه‌ای است که معنای عام و وسیعی را پوشش می‌دهد. مانند کبوتر یا پرندگان؛ دوم آنکه موضوع، از نظر معنایی، قابلیت قرار گرفتن در رده‌های مختلف را داشته باشد. به‌تعبیر دیگر، واژهٔ کلی، برای نشان دادن رابطهٔ معنایی بین توصیفگرهای زبان در اصطلاحنامه غنی باشد (کوالسکی، ۱۹۹۷، ص ۱۲۷).

در خوشه‌بندی اطلاعات اصطلاحنامه، اینکه یک موضوع در چندین رده قرار بگیرد یا

تنها در یک رده، بستگی به تصمیمی دارد که در آغاز گرفته می‌شود. با توجه به اینکه زبان معمولاً ایهام دارد، بهتر است که موضوع در چندین رده قرار گیرد. این عمل انعطاف‌پذیری بیشتری را در پی خواهد داشت. در ایجاد یک اصطلاحنامه مسائل مهم دیگری نیز وجود دارد که جزء خوشه‌بندی مدارک بیان نمی‌شود (آی تچیسون و گیل کریست، ۱۹۷۲) از قبیل: رویکرد همارایی واژه[۷۰]، رابطهٔ میان واژه‌ها[۷۱]، تفکیک لغات با املای یکسان[۷۲]، و محدودیت‌های واژگان[۷۳]. در روش همارایی واژه، عبارات در قالب اصطلاح‌های خاص و جداگانه رده‌بندی شده که به‌صورت روش‌های پس‌همارایی و پیش‌همارایی ظاهر می‌شود. در مقولهٔ رابطه میان واژه‌ها در تهیهٔ یک اصطلاحنامه که در آن انسان نیز می‌تواند دخالت داشته باشد – در مقابل اصطلاحنامهٔ خودکار – ایجاد رابطه‌های متنوعی بین واژه‌ها امکان‌پذیر است. در بارهٔ تفکیک لغات با املای یکسان (هم‌نوشت) باید بیان کرد که منظور، مرزبندی و مشخص ساختن واژه‌ای است که املای یکسان با معانی کاملاً مختلف دارد. مثلاً: شیر که می‌تواند شامل شیر خوراکی، شیر آب، یا شیر جنگل باشد. برای تدوین راه‌حل این مشکل پیشنهاد شده که وقتی می‌خواهیم از این کلمات استفاده کنیم باید همزمان با آن کلمهٔ دیگری آورده شود که نتیجهٔ صحیح جست‌وجو را ارائه دهد. ولی در موضوع محدودیت‌های واژگان با دو مقولهٔ عادی‌سازی و اخص‌شدن مواجه هستیم (آی تچیسون و گیل کریست، ۱۹۷۲).

خوشه‌بندی مدارک

فرآیند خوشه‌بندی مدارک بسیار شبیه خوشه‌بندی واژه‌ها برای ایجاد یک اصطلاحنامه است که به دو صورت دستی و خودکار انجام می‌شود. در روش خوشه‌بندی دستی مدارک جزء ذاتی هر کتابخانه یا نظام بایگانی است. در این شیوه، فردی مدرک را مطالعه می‌کند و سپس تصمیم می‌گیرد که مدرک به چه گروه یا گروه‌هایی تعلق دارد. در سیستم دستی معمولاً هر مدرک به یک گروه تعلق می‌گیرد. با به‌کارگیری نمایه‌سازی، یک مدرک در یک گروه اولیه ذخیره شده و می‌تواند در گروه‌های دیگر به‌صورت نمایه‌ای تعریف شود. با ظهور نظام‌های الکترونیکی، نگهداری مدارک خوشه‌بندی به‌شیوهٔ خودکار امکان‌پذیر شده است. روش‌هایی که برای خوشه‌بندی واژه‌ها بیان شد، برای خوشه‌بندی مدارک نیز به کار می‌رود. تشابه بین مدارک براساس دو معیار تعیین می‌شود: نخست واژه‌های مشترک، سپس رابطه‌های مشترک. فرمول آن نیز چنین است:

$$SIM\ (Item_i, Item_j) = \sum\ (Item_i k)(Term_j k)$$

بر همین اساس، ماتریسی ایجاد می‌شود که ستون‌ها و ردیف‌هایش مبیّن مدارک است:

70. Word coordination approach
71. Words relationship
72. Homograph resolution
73. Vocabulary constraints

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1		11	3	6	22
Item 2	11		12	10	36
Item 3	3	12		6	9
Item 4	6	10	6		11
Item 5	22	36	9	11	

شکل ۹

ماتریس مدرک به مدرک (واژه‌های مشترک) (کوالسکی، ۱۹۹۷، ص ۱۴۱)

	Item 1	Item 2	Item 3	Item 4	Item 5
Item 1		1	0	0	1
Item 2	1		1	1	1
Item 3	0	1		0	0
Item 4	0	1	0		1
Item 5	1	1	0	1	

شکل ۱۰

رابطه‌های مشترک (کوالسکی، ۱۹۹۷، ص ۱۴۱)

در این شیوه، با استفاده از الگوریتم خوشه‌بندی دسته‌ای، رده‌های زیر به‌دست می‌آید:

Class 1= Item 1, Item 2, Item 5

Class 2= Item 2, Item 3

Class 3= Item 2, Item 4

در صورتی که از الگوریتم خوشه‌بندی تک‌پیوندی استفاده کنیم، ردهٔ زیر حاصل می‌شود. در این روش، کلیهٔ مدارک در یک رده قرار گرفته‌اند(مدارک ۳ و ٤ به علت تشابه آنها با مدرک ۲ به این رده اضافه شده‌اند):

Class 1= Item 1, Item 2, Item 5, Item 3, Item 4

چنانچه قصد بهره‌گیری از الگوریتم خوشه‌بندی ستاره‌ای داشته باشیم، رده‌های زیر

به‌عنوان نتیجهٔ فرایند خوشه‌بندی مدارک ظاهر می‌شود:

Class 1 - Item 1, Item 2, Item 5

Class 2 - Item 2, Item 3, Item 4, Item 5

و، در نهایت، با استفاده از الگوریتم خوشه‌بندی رشته‌ای رده‌های زیر حاصل می‌شود:

Class 1- Item 1, Item 2, Item 3

Class 2- Item 4, Item 5

باید بر مطالب پیشین افزود که در محدودهٔ واژه‌ها، آن بخش که دارای املاهای مشابه هستند، سبب بروز ایهام و نیز خطا می‌شود. در محدودهٔ مدارک نیز، عنوان‌های چندگانه برای یک مدرک می‌تواند باعث مشکلات مشابهی شود. این مشکل به‌صورت خاص، هنگام مرزبندی میان مدارک، بیشتر مشخص می‌شود. بدون پیش‌همارایی معنای مفاهیم، مدرکی که در مورد «سیاست» در «آمریکا» و «اقتصاد» در «مکزیک» است می‌توانند در یک رده خوشه‌بندی شوند که بر روی سیاست در مکزیک تمرکز دارد.

ایجاد سلسله‌مراتب در فرایند خوشه‌بندی

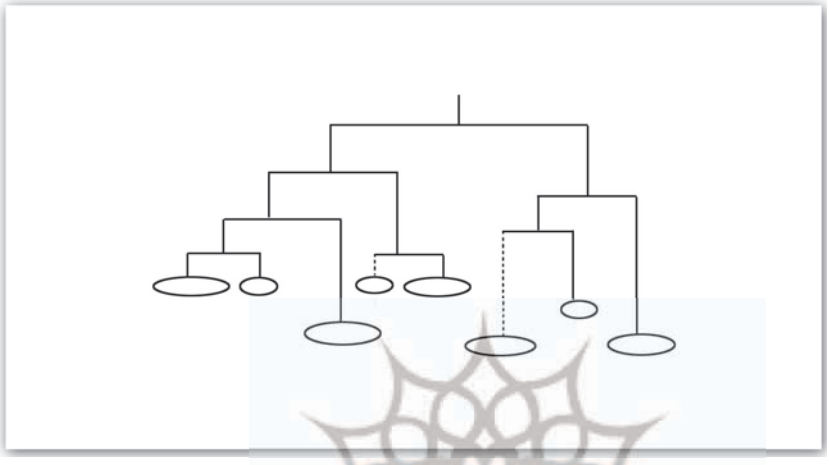
مجموعه‌ای از رده‌ها، که برای نمایش رابطه‌های نوعی فراتر، فروتر، و هم‌سطح بین رده‌های موازی و همتراز منظم شده‌اند، به‌عنوان سلسله‌مراتب خوشه‌ها[۷۴] شناخته می‌شوند. در این سلسله‌مراتب، تقدم و تأخر موضوع‌ها در طرح رده‌بندی به‌نحوی است که هریک از عناصر مذکور در توالی موضوعی، وابسته به عنصر منحصربه‌فرد ماقبل خود باشد. یعنی درست همانگونه که در خدمت دانش سنتی مرسوم بوده است.

بر این اساس، کاهش سرریز اطلاعات از طریق اجرای جست‌وجوهای بالا به پایین، تعیین مراکز ثقل خوشه‌ها در سلسله‌مراتب، و حذف شاخه‌های زائدی که مرتبط نیستند، انجام می‌شود. باید بیان کرد که نمایش تصویری از فضای اطلاعاتی مدارک کار بسیار دشواری است. استفاده از نمودار درختی همراه با نشانه‌های تصویری، تعیین اندازهٔ خوشه‌ها، و توانایی ایجاد ارتباط بین آنها، بهترین روش برای ایجاد سلسله‌مراتب است. این روش سودمند به توسعهٔ بازیابی مدارک مرتبط منجر می‌شود که دارای اثربخشی زیادی در فرایند بازیابی اطلاعات است.

استفاده از نمودار درختی فواید خاص خود را دارد. ویژگی‌های مهم این شیوه شامل سه مورد زیر می‌شود: ۱) این مدل به استفاده‌کننده اجازه می‌دهد مسیر جست وجو در پایگاه اطلاعاتی را به آسانی بازیابی کند؛ ۲) استفاده‌کننده، تمامی مدارک یا واژه‌های موردنظر در خوشه‌ها را ملاحظه می‌کند؛ ۳) استفاده‌کننده می‌تواند اخص بودن مدرک یا

74 . Hierarchy of cluster

واژه‌ها را از طریق رفتن به خوشه‌های پایین‌تر افزایش دهد، و یا با رفتن به خوشه‌های بالایی جست‌وجو را به شیوهٔ عمومی‌تری دنبال کند. ساختار سلسله‌مراتبی، در واقع، یکی از روش‌های مناسب برای نمایش روابط میان رده‌های تعیین‌شده است. معمولاً در ساختار برنامه‌های رایانه‌ای نیز در بازنمایی اطلاعات مرتبط از روش سلسه‌مراتبی بهره می‌گیرند.



شکل ۱۱
نمودار روابط سلسه‌مراتبی (کوالسکی، ۱۹۹۷، ص ۱۴۳)

خوشه‌بندی در محیط وب؛ رویکردی نوین

منابع وب، با توجه به پارامترهای خاص خود، و نیز ویژگی‌هایی از قبیل معروف بودن، ساختار متمایز و یا محتوای آن دسته‌بندی می‌شوند. خوشه‌بندی در وب می‌تواند یکی از انواع زیر باشد:

خوشه‌بندی کاربر وب[۷۵]: یعنی ایجاد گروه‌هایی از کاربران که الگوهای مرور[۷۶] مشابهی را انجام می‌دهند. چنین مهارتی خصوصاً برای اشاره به آمار کاربران به‌منظور انجام فعالیت‌های متنوعی از قبیل تقسیم‌بندی بازار در کاربردهای تجارت الکترونیک مفید است. به‌همین ترتیب، این نوع خوشه‌بندی، به فهم بهتر رفتار هدایتی کاربران کمک کرده و درخواست خدمت کاربران وب را از طریق کاهش طول مسیرهای خط سیر هدایت وب بهبود می‌بخشد.

خوشه‌بندی اسناد وب[۷۷]: که عبارت است از گروه‌بندی اسناد با محتوای مرتبط. این اطلاعات در بسیاری از کاربردها مفید و سودمند است. برای مثال، کاربرد آن در موتورهای جست‌وجوی وب، به‌منظور بهبود فرایند بازیابی اطلاعات (نظیر خوشه‌بندی پرس‌وجوهای وب) است. علاوه بر آن، خوشه‌بندی اسناد وب موجب افزایش دسترسی به اطلاعات وب شده و ارسال محتوا را بهبود می‌بخشد (واکالی، ۲۰۰۷). تقسیم‌بندی خوشه‌بندی در وب به دو گروه کاربران و اسناد، یکی از طبقه‌بندی‌هایی است که در پژوهش‌های مربوط به حوزهٔ

75. Web user clustering
76. Browsing patterns
77. Web document clustering

خوشه‌بندی وب کاربرد دارد. باید بیان کرد که فرایند خوشه‌بندی در وب، براساس داده‌های ارائه شده، مبتنی بر خوشه‌بندی اسناد وب است. در این میان، بیشترین توجه بر ساختار مدارک و اطلاعات موجود در وب بوده است. ولی اگر خوشه‌بندی اسناد و اطلاعات را به‌عنوان پایهٔ اصلی این فرایند در مطالعات وب‌مدار در نظر بگیریم، با دو رویکرد عمده در خوشه‌بندی منابع وب روبه‌رو خواهیم بود.

می‌توان روش‌های خوشه‌بندی را به دو رویکرد خوشه‌بندی مبتنی بر متن[۷۸] و خوشه‌بندی مبتنی بر پیوند[۷۹] تقسیم کرد. رویکرد خوشه‌بندی مبتنی بر متن از محتوی متن اسناد برای برآورد شباهت میان اسناد استفاده می‌کند. خوشه‌بندی مبتنی بر متن اسناد وب، معمولاً توسط مدل‌های فضای بُرداری[۸۰] در فضای بُرداری ابعاد بالا[۸۱] نمایش داده می‌شود، که در آن عبارات با مؤلفه‌های بُردار، همبستگی دارند. هنگامی‌که اسناد وب، بُرداری شدند، روش‌های خوشه‌بندی بُرداری، خوشه‌های اسناد وب را فراهم می‌سازد. ولی در روش خوشه‌بندی مبتنی بر پیوند، وب به‌صورت یک نمودار مستقیم تلقی می‌شود که در آن گروه‌ها، نمایشگر اسناد وب، همراه با آدرس ونشانی مکان‌یاب یکسان منابع[۸۲] هستند و لبه‌های میان گروه‌ها، نمایشگر فراپیوندهای بین اسناد وب می‌باشند. فنون مبتنی بر پیوند از وضعیت مکان‌شناسی[۸۳] وب‌سایت برای خوشه‌بندی اسناد وب استفاده می‌کنند (واکالی، ۲۰۰۷). در شیوهٔ خوشه‌بندی مبتنی بر پیوند، تعیین ویژگی‌های خاص پیوندهای برقرار شده از اهمیت بالایی برخوردار است. نشانی اینترنتی نقشی مؤثر ایفا می‌کند و بر اساس آن می‌توان بهتر و بیشتر از امکان خوشه‌بندی بهره جست. خوشه‌بندی در محیط وب، بیشتر با اطلاعات دیجیتالی در فضای مجازی ارتباط می‌یابد، و بُردارها نقشی بسزا در تثبیت خوشه‌ها برعهده دارند.

نتیجه‌گیری

خوشه‌بندی داده‌ها یکی از راهکارهای سودمندی است که می‌توان با کمک آن به ایجاد توازن در گروه‌بندی اطلاعات اقدام کرد. خوشه‌بندی داده، گروهی از داده‌های همسان را که از ویژگی‌های مشابه برخوردار هستند، در یک رده، باعنوان واحد، سازماندهی می‌کند. بر این اساس، خوشه‌بندی، مرتب‌کردن واژه‌ها یا مدارک شبیه به‌هم در یک رده، زیر یک عنوان کلی است. خوشه‌بندی یک مرحلهٔ مهم از فرایند پردازش تحلیل داده است که هدف آن تقسیم‌بندی منطقی یک مجموعه ساختارنیافته از اجزا، درون خوشه‌ها یا گروه‌های مشخص است.

در واقع، هدف اصلی فرایند خوشه‌بندی، کمک به استفاده‌کنندگان برای تعیین محل دقیق اطلاعات است.

این فعالیت با شناسایی گروه‌بندی طبیعی داده‌ها شکل می‌گیرد. از کارکردهای مهم خوشه‌بندی داده می‌توان به ساخت و ایجاد تزاروس و اصطلاحنامه اشاره کرد. در

78. Text-Based clustering approach
79. Link-Based clustering approach
80. Vector space model
81. High-dimension vector space
82. Uniform Resource Locator (URL)
83. Topology

مراحل کاری خوشه‌بندی داده، تعیین حدوحدود کار اولین مرحلهٔ آن است. سپس، تعیین خصوصیات و مشخصات واژه‌ها و مدارک مطرح می‌شود. از مراحل دیگر، باید به تعیین میزان رابطهٔ بین ویژگی‌های واژه اشاره کرد. آخرین مرحله نیز ایجاد الگوریتم است. انواع الگوریتم‌های خوشه‌بندی، با توجه به نوع اطلاعات و ساختار و روابط مورد نظر مورد استفاده قرار می‌گیرد. روش‌های عمدهٔ خوشه‌بندی نیز چهار شیوهٔ خوشه‌بندی دسته‌ای، تک‌پیوندی، ستاره‌ای، و رشته‌ای است. روش دیگری که باید علاوه بر روش‌های پیشین نام برد، روش خوشه‌بندی با استفاده از خوشه‌های موجود است. این شیوه تعداد محاسبات مورد نیاز برای تعیین مشابهت‌ها در تهیهٔ خوشه‌ها را کاهش می‌دهد. خوشه‌بندی مدارک نیز شبیه خوشه‌بندی واژه‌ها برای ساخت اصطلاحنامه به‌کار می‌رود، که به دو صورت دستی و خودکار انجام می‌شود. خوشه‌بندی وب‌مدار یکی از انواع جدید خوشه‌بندی است که با فضای بُرداری و متون و پیوندهای موجود در محیط وب ارتباط دارد. تمامی فرایندهای خوشه‌بندی، در نهایت تضمین‌کنندهٔ دستیابی سریع و مطمئن به اطلاعات همبسته، و شناسایی ارتباط منطقی میان آنهاست.

منابع

Aitchison, J.; Gilchrist, A. (1972). ***Thesaurus construction: A practical manual***, London: ASLIB.

Anthony, Adam; desJardins, Marie(2006). "Open Problems in relational data clustering". ICML Workshop on Open Problems in Statistical Relational Learning, Pittsburgh, PA. Retrieved Nov.30, 2007, from: http:// maple. cs.umbc.edu/papers/aanthMDJRelClustFinal.pdf

Anthony, Adam; DesJardins, Marie(2007). "Data clustering with a relational push- pull model". Retrieved Nov.30, 2007, from: http://maple.cs.umbc.edu/papers/anthonya-ClusteringRPPM.pdf

Breaux, Travis D.; Reed, Joel W. (2005). "Hierarchical information clustering Using Ontology Languages". 38th Hawaii Int'l Conf. on System Sciences(January). Retrieved Nov.30, 2007, from: http://www.csm.ornl.gov/~jreed/publications/HierarchicalInfoClustering.pdf

Gionis, Aristides; Mannila, Heikki; Tsaparas, Panayiotis (2004). "Clustering aggregation". Helsinki Institute for Information Technology, BRU, Department of Computer Science, University of Helsinki, Finland. Retrieved feb.13, 2007, from: www.cs.helsinki. fi/u/gionis/papers/icde05.pdf

Jain, A. K.; Law, Martin H.C. (2005). "Data clustering: A user's dilemma". 1st International Conference, PReMI, Kolkata, India, December 20 - 22. Retrieved Nov. 30, 2007, from: http://www.cse.msu.edu/~lawhiu/papers/JainLaw_PReMI.pdf

Jain, A.K.; Murty, M.N; Flynn, P.J (1999). "Data clustering: A review". ***ACM Computing Sur-***

veys, 31 (3): 265 - 323. Retrieved Feb.19, 2007, from: http://www.cs.rutgers.edu/~mlittman/courses/lightai03/ jain99data.pdf

Kanungo, Tapas … [et al] (2002). "An efficient k-means clustering algorithm: Analysis and implementation". IEEE Transaction on Pattern Analysis and Machine Intelligence, 24 (7): 881 - 892.

Kowalski, Gerald (1997.) ***Information retrieval systems: Theory and implementation***. Boston: Kluwer Academic Publishers.

Kraskov, A. … [et al] (2005). "Hierarchical clustering using mutual information". ***Europhysics Letters***, 70 (2) 278 – 284.

Kuhn, Adrian Ivo (2006). "Semantic clustering making use of linguistic information to reveal concepts in source code". Institut f¨ur Informatik und angewandte Mathematik. Retrieved Feb.19, 2007, from: www.iam.unibe.ch/~scg/Archive/Diploma/Kuhn06a.pdf

Leuski, Anton (2002). "Evaluating document clustering for interactive information retrieval". Retrieved Feb.19, 2007, from: wkd.iis.sinica.edu.tw/~slchuang/lecture/2002-01-17/ir-235.pdf

Matteucci, Matteo (2003). "A tutorial on clustering algorithms". Retrieved Mar.4, 2007. from: http://www.elet.polimi.it/upload/matteucc/ Clustering/tutorial_html

Neville, Jennifer; Adler, Micah; Jensen, David (2003). "Clustering relational data using attribute and link information". Retrieved Feb.19, 2007, from: http://kdl.cs.umass.edu//papers/neville-et-al-textlink2003.pdf

Perlich, Claudia; Rosset, Saharon (2007). "Identifying bundles of product options using mutual information clustering". Retrieved Nov.30, 2007, from: http://www.siam.org/meetings/proceedings/2007/ datamining/ papers/035perlich.pdf

Rosell, Magnus (2006). "Introduction to information retrieval and text clustering". Retrieved Feb.19, 2007, from: www.nada.kth.se/~rosell/undervisning/sprakt/irintro060801.pdf

Salton, G. (1972). "Experiments in automatic thesaurus construction for information retrieval". Information processing, 71th, North Holland Publishing Co., Amsterdam, p. 115 - 123.

Song, Qing (2005). "A robust information clustering algorithm". ***Neural computation***, 17 (12): 267 – 298. Retrieved Nov.30, 2007, from: http://www.galenicom.com/pt/medline/article/16212767

Vakali, Athena; Pallis, George; Angelis, Lefteris (2007). "Clustering web information sources". Idea Group Inc. Retrieved Feb.13, 2007, from: oswinds.csd.auth.gr/papers/idea06.pdf

Wang, Y-C; Vandenthorpe, J; Evans, M.(1985). "Relationship thesauri in information retrieva". ***American Society of Information Science***, 15 - 27.